Accurate identification of genes associated with brain disorders by integrating heterogeneous genomic data into a Bayesian framework

Dan He,[a,b,e] Ling Li,[a,b,e] Huasong Zhang,[a,b] Feiyi Liu,[a,b] Shaoying Li,[a,b] Xuehao Xiu,[a,b] Cong Fan,[a,b] Mengling Qi,[a,b] Meng Meng,[c] Junping Ye,[a,b] Matthew Mort,[d] Peter D. Stenson,[d] David N. Cooper,[d] and Huiying Zhao[a,b,*]

[a]Department of Medical Research Center, Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University, Guangzhou, 510006, China
[b]Guangdong Provincial Key Laboratory of Malignant Tumor Epigenetics and Gene Regulation, Guangzhou, 510006, China
[c]School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, 510006, China
[d]Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff, CF14 4XN, UK

eBioMedicine 2024;107: 105286

Published Online 20 August 2024

https://doi.org/10.1016/j.ebiom.2024.105286

Summary

Background Genome-wide association studies (GWAS) have revealed many brain disorder-associated SNPs residing in the noncoding genome, rendering it a challenge to decipher the underlying pathogenic mechanisms.

Methods Here, we present an unsupervised Bayesian framework to identify disease-associated genes by integrating risk SNPs with long-range chromatin interactions (iGOAT), including SNP-SNP interactions extracted from ∼500,000 patients and controls from the UK Biobank, and enhancer–promoter interactions derived from multiple brain cell types at different developmental stages.

Findings The application of iGOAT to three psychiatric disorders and three neurodegenerative/neurological diseases predicted sets of high-risk (HRGs) and low-risk (LRGs) genes for each disorder. The HRGs were enriched in drug targets, and exhibited higher expression during prenatal brain developmental stages than postnatal stages, indicating their potential to affect brain development at an early stage. The HRGs associated with Alzheimer's disease were found to share genetic architecture with schizophrenia, bipolar disorder and major depressive disorder according to gene co-expression module analysis and rare variants analysis. Comparisons of this method to the eQTL-based method, the TWAS-based method, and the gene-level GWAS method indicated that the genes identified by our method are more enriched in known brain disorder-related genes, and exhibited higher precision. Finally, the method predicted 205 risk genes not previously reported to be associated with any brain disorder, of which one top-risk gene, *MLH1*, was experimentally validated as being schizophrenia-associated.

Interpretation iGOAT can successfully leverage epigenomic data, phenotype–genotype associations, and protein–protein interactions to advance our understanding of brain disorders, thereby facilitating the development of new therapeutic approaches.

Funding The work was funded by the National Key Research and Development Program of China (2024YFF1204902), the Natural Science Foundation of China (82371482), Guangzhou Science and Technology Research Plan (2023A03J0659) and Natural Science Foundation of Guangdong (2024A1515011363).

Abbreviations: SCZ, Schizophrenia; BP, Bipolar disorder; MDD, Major depressive disorder; AD, Alzheimer's disease; PD, Parkinson's disease; MG, Migraine; EPI, Enhancer-promoter interactions; HRG, High-risk gene; LRG, Low-risk gene; Astro, Astrocytes; Endo, Endothelial cells; Micro, Microglia; Oligo, Oligodendrocytes; OPC, Oligodendrocyte progenitor cells; In, Inhibitory neurons; Ex, Excitatory neurons; psyPRGs, Genes predicted as associated with two or more psychiatric disorders; dePRGs, Genes predicted as associated with two or more neurological diseases; neuronDRG, HRGs predicted by iGOAT using EPI but were not predicted as HRGs by iGOAT without using EPI; noHiCDRG, HRGs predicted by iGOAT without using EPI but were not predicted as HRGs by iGOAT using EPI data; cLRG, The LRGs predicted by both methods; siGOAT, Another version of iGOAT by excluding SNP-SNP interactions; nnDRG, HRGs predicted by iGOAT with SNP-SNP interactions but not predicted by siGOAT; snDRG, HRGs predicted by siGOAT but not predicted by iGOAT; TiGOAT, Another version of iGOAT by using sampled strategy as in rGAT-omics; iGOAT-only, Genes predicted to be HRGs associated with SCZ by iGOAT but were not predicted as risk genes associated with SCZ by coloc or TWAS; eQTL-only, Genes predicted to be associated with SCZ by coloc or TWAS but were not predicted as HRGs by iGOAT; iGOAT-onlyH, Genes predicted to be HRGs by iGOAT but were not predicted as risk genes by H-MAGMA; H-MAGMA-only, Genes predicted by H-MAGMA but were not predicted as HRGs by iGOAT

*Corresponding author. Department of Medical Research Center, Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University, Guangzhou, 510006, China.

E-mail address: zhaohy8@mail.sysu.edu.cn (H. Zhao).

[e]These authors contributed equally: Dan He, Ling Li.

Keywords: Genome-wide association studies; Bayesian framework; Psychiatric disorders; Neurodegenerative/neurological diseases; Risk genes; SNP-SNP interactions; Enhancer–promoter interactions

Research in context

Evidence before this study

Genome-wide association studies have revealed many SNPs associated with brain disorders. Most of these SNPs reside in non-coding regions, and it remains a major challenge to decipher the pathogenicity of these risk SNPs. A commonly used approach addresses this issue by aggregating SNP associations to their nearest genes. However, increasing evidence supports the claim that index SNPs often influence the expression of genes at some considerable distance on the chromosome. Another approach is gene-level genome-wide association studies, which considers the associations between a trait and all SNP markers within a gene rather than individual SNPs. However, this approach combines the effects of SNPs within genes without considering the biological connection between the gene and the diseases. With the rapid development of sequencing technology, multiple sources of data are available for use in predicting SNP-linked genes, which can contribute to exploring the pathogenicity of risk SNPs. Currently, an appropriate model that simultaneously integrates genotype data with gene networks and cell-type/stage-specific epigenetic data is not yet available.

Added value of this study

This study presents a Bayesian framework to predict genes impacted by lead SNPs by integrating long-range chromatin interactions including enhancer–promoter interactions (EPI) derived from Hi-C data of multiple brain cell types at different developmental stages, as well as SNP-SNP interactions determined using about 500,000 samples from the UK Biobank.

This method was applied to three psychiatric disorders and three neurological disorders and predicted sets of high-risk (HRGs) and low-risk (LRGs) genes for each disorder. Comparisons between this method and an eQTL-based method, a TWAS-based method, and a gene-level GWAS method indicate that the HRGs identified by our method are more enriched in known brain disorder-related genes and exhibited higher precision. This method predicted 205 HRGs not previously known to be brain disorder-related, of which one top-ranking gene was experimentally validated as being schizophrenia-associated.

Implications of all the available evidence

The application of this method to six brain disorders predicted a set of HRGs. The HRGs exhibited higher levels of expression during prenatal brain developmental stages than postnatal stages, possibly reflecting the occurrence of the initial neuronal damage at a relatively early stage of brain development. The gene co-expression module analysis and rare variants analysis indicated the HRGs associated with neurodegenerative disorders shared genetic architecture with the psychiatric disorders. The method proposed by this study has advanced our understanding of the brain disorders and facilitated the identification of new therapeutic approaches.

Introduction

With the rapid development of genome-wide association studies (GWAS), more and more SNPs have been found to be associated with disease.[1] These GWAS findings have provided potential causal SNPs of the disease but precisely how these SNPs impact the gene expression pathway or participate in the disease generation process remains a challenging problem. In most cases, genes in the immediate vicinity of the index SNP have been simply assumed to be causal. There is however increasing evidence to support the contention that index SNPs may often influence the expression of genes at some considerable distance on the chromosome. Another commonly used approach is gene-level genome-wide association studies, which considers the associations between a trait and all SNP markers within a gene rather than individual SNPs.[2–4] However, this approach only combines effects of SNPs within genes without considering biological connections between the genes and the diseases.

Researchers have endeavored to detect risk genes by integrating GWAS loci with epigenomic and transcriptomic data, and other genomics data.[5] By accommodating positional information, expression quantitative trait loci (eQTL), and chromatin interaction mapping,[6] a web-based platform, FUMA, has been presented, which provides gene-based functional annotation of GWAS results. Pardiñas et al. combined genomic fine mapping, brain expression, and chromosome conformation data to detect causal genes for schizophrenia (SCZ)[7] whilst Fan et al. applied a systems-level analysis integrating GWAS data with transcriptomic information to identify genes associated with Alzheimer's disease.[8] OSCA is a commonly-used method leveraging omic data for analyzing associations between genes and complex traits.[9] Another method (iRIGS) was employed to identify risk genes in SCZ by integrating GWAS data with regulation networks, distal regulatory information, and genetic variant data.[10] More recently, we have proposed a risk gene predictor, rGAT-omics,[11] which integrates gene networks, gene distance to SNPs, distal regulatory elements, and gene expression information to identify those candidate genes impacted by SCZ-associated SNPs reported in.[12] However, all these methods have neglected to include important additional sources of genomic or transcriptomic information. For example, SNP-SNP interactions serve to define combinatorial effects of variants on the etiology of genetic disorders and are important for providing long-distance interactions between genes associated with disease.[13–18] Moreover, they have tended not to consider the potential utility of epigenetic information involved in different tissues and developmental stages for inferring gene candidacy. One of the most popular computational methods (H-MAGMA[19]) utilized long-range interactions in disease-related tissues to identify disease-related genes. Application of H-MAGMA to nine brain disorders revealed interesting shared biological features. It therefore follows that information on long-range chromatin interactions, such as enhancer–promoter interactions, could be very important for the identification of genes impacted by SNPs.

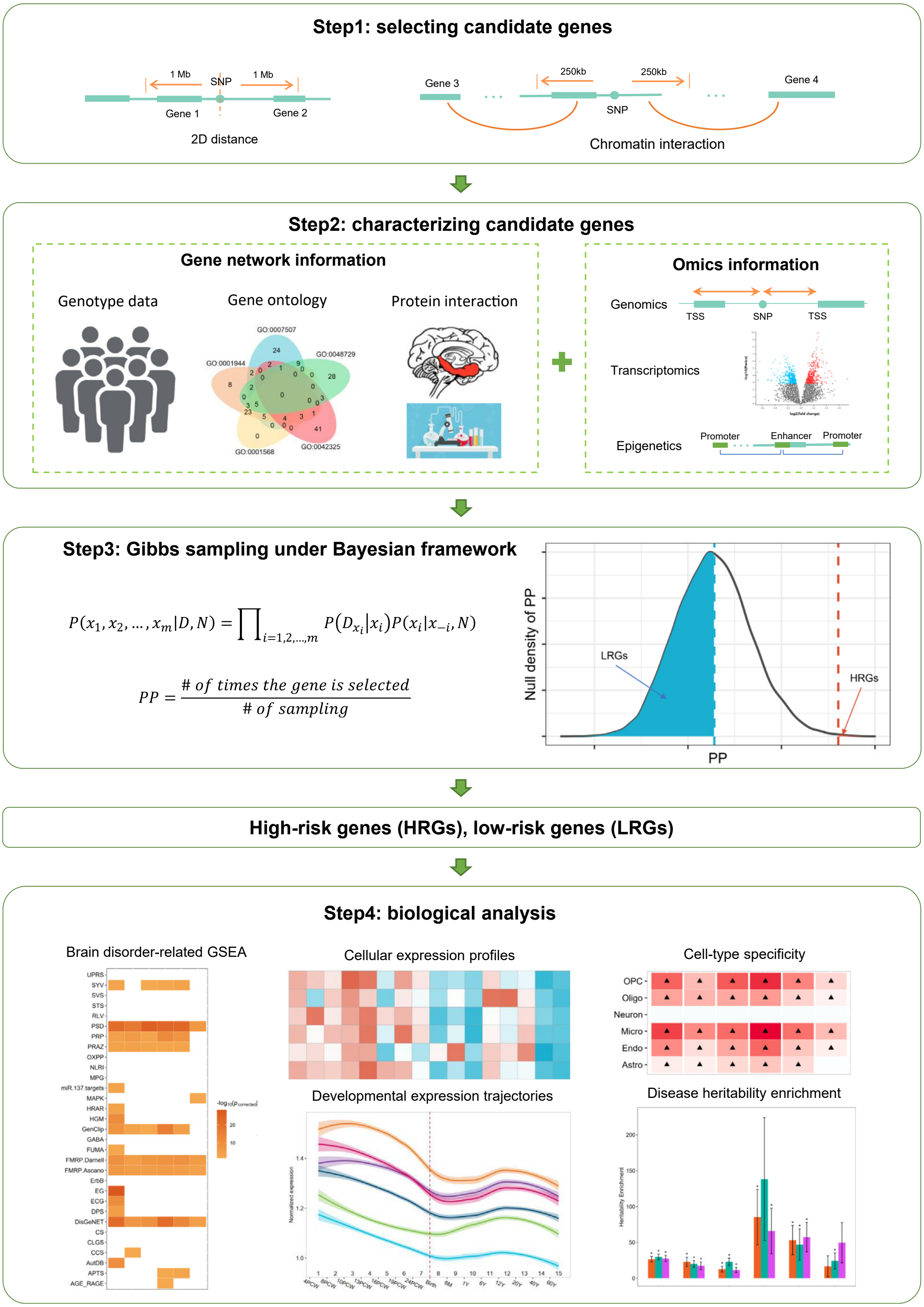
By identifying the genes potentially associated with the disorders, many studies have been performed to indicate the shared genetic mechanisms between disorders. A recent publication[20] that indicated the genetic correlations between AD and psychiatric disorders by integrating the GWAS results with human brain transcriptomes and proteomes. Another study[21] has performed proteomic sequencing of the dorsolateral prefrontal cortex in 438 older individuals, and indicated that proteins and modules associated with cerebral atherosclerosis were also associated with Alzheimer's disease, suggesting shared mechanisms between these two disorders. Additionally, a recent study has investigated the GWAS data from 1 million cases across ten neurological diseases and 10 psychiatric disorders to identify their shared genetic overlaps. In particular, they have found the shared genetic mechanisms between migraine, essential tremor, stroke and multiple sclerosis with several psychiatric disorders.[22] Thus, developing a reliable method for identifying disease-associated genes is a useful way to investigate the shared mechanisms of diseases.

In this study, we have established an unsupervised learning method **i**nte**g**rating risk SNPs with long-range chr**o**m**a**tin in**t**eractions (iGOAT) to predict high-risk genes associated with a given disease. This method employs a Bayesian framework to integrate genomic information and networks with the goal of identifying a set of most likely disease-related genes. We considered long-distance chromatin interactions, including enhancer–promoter interactions derived from multiple brain cell types at different developmental stages and weighted SNP-SNP interactions constructed by calculating the associations between genotypes and clinical phenotypes, which were estimated using data from about 500,000 samples in the UK Biobank (Fig. 1). This method is not merely an enhancement of our previous approach (rGAT-omics) by virtue of the inclusion of SNP-SNP interactions and stage-specific and cell-type-specific enhancer–promoter interactions (EPI) data. It also predicts high-risk genes (HRGs) by combining candidate genes for many SNPs into one set and introducing empirical distributions of the candidate gene set which can enlarge the sample size of the candidate gene sets while controlling false positives, thereby improving the accuracy of the predictions. This method was then applied to psychiatric disorders [schizophrenia (SCZ), bipolar disorder (BP), and major depressive disorder (MDD)] and neurodegenerative/neurological diseases [Alzheimer's disease (AD), Parkinson's disease (PD), and migraine (MG)] to identify HRGs. The predicted HRGs were consistent with our current understanding of the pathophysiology of psychiatric and neurodegenerative diseases. Importantly, iGOAT revealed many genes associated with brain disorders, which are specifically expressed in brain cells and are enriched for drug targets. Notably, experimental validation has confirmed the association of *MLH1*, one of the top ranked genes, with SCZ, a neuropsychiatric disorder that is linked to dysregulation of neural stem cell (NSC) proliferation and differentiation.[23] This predictive approach shows great potential in uncovering the underlying biological pathways, developmental windows, and cell types contributing to specific brain disorders. Thus, this type of prediction could be helpful in characterizing brain disorders and providing new therapeutic approaches.

Methods

Overview of iGOAT

An increasing number of GWAS have identified a plethora of disease-associated SNPs. However, most of these SNPs resided in noncoding genome regions, which are hard to link to disease mechanisms for further therapeutic design. Here, we developed an unsupervised Bayesian framework-based method, integrating risk SNPs with long-range chromatin interactions (iGOAT) to identify genes associated with diseases. iGOAT integrated risk SNPs with gene–gene interactions, and long-range chromatin interactions, including SNP-SNP interactions derived from genotype data and enhancer–promoter interactions derived from Hi-C data of multiple brain cell types at different developmental stages. As shown in Fig. 1, iGOAT is composed of four steps: (1) Assigning candidate genes to SNPs by leveraging SNP-to-gene distance and chromatin interaction profiles (Hi-C); (2) Scoring candidate genes by genomic data from multiple sources and constructing gene networks based on SNP-SNP


Step1: selecting candidate genes
1 Mb
SNP
1 Mb
Gene 1
Gene 2
2D distance
Gene 3
250kb
250kb
SNP
Gene 4
Chromatin interaction
Step2: characterizing candidate genes
Gene network information
Genotype data
Gene ontology
Protein interaction
GO:0007507
GO:0001944
GO:0048729
GO:0001568
GO:0042325
Omics information
Genomics
TSS
SNP
TSS
Transcriptomics
Epigenetics
Promoter
Enhancer
Promoter
Step3: Gibbs sampling under Bayesian framework
$P(x_1, x_2, \ldots, x_m | D, N) = \prod_{i=1,2,\ldots,m} P(D_{x_i} | x_i) P(x_i | x_{-i}, N)$
$PP = \frac{\#\ of\ times\ the\ gene\ is\ selected}{\#\ of\ sampling}$
Null density of PP
LRGs
HRGs
PP
High-risk genes (HRGs), low-risk genes (LRGs)
Step4: biological analysis
Brain disorder-related GSEA
UPRS
SYV
SVS
STS
RLV
PSD
PRP
PRAZ
OXPP
NLRI
MPG
miR.137.targets
MAPK
HRAR
HGM
GenClip
GABA
FUMA
FMRP.Darnell
FMRP.Ascano
ErbB
EG
ECG
DPS
DisGeNET
CS
CLGS
CCS
AutDB
APTS
AGE_RAGE
-log10(P corrected)
20
10
Cellular expression profiles
Developmental expression trajectories
Normalized expression
1.4
1.2
1.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
4PCW 8PCW 10PCW 13PCW 16PCW 19PCW 24PCW 6m 6M 1Y 6Y 12Y 20Y 40Y 60Y
Cell-type specificity
OPC
Oligo
Neuron
Micro
Endo
Astro
Disease heritability enrichment
Heritability Enrichment
200
150
100
50
0

interaction, gene ontology and protein–protein interactions; (3) Using the scores of genes as sampled probabilities, Gibbs sampling was iteratively performed to assess the association between genes and the disease under a Bayesian framework until the frequencies (posterior probabilities) of genes selected as risk genes converged. We then computed empirical p-values for candidate genes by generating null distributions of posterior probabilities. After multi-correction (Benjamini-Hochberg: BH), genes with empirical $p_{corrected}$ < 0.001 were considered high-risk genes (HRGs); genes with empirical $p_{corrected}$ > 0.5 were considered low-risk genes (LRGs). (4) The biological features of the predicted high-risk genes (HRGs) were explored and compared with those of low-risk genes (LRGs). The computer code and output results are available at (https://github.com/Dan-He/iGOAT).

Collecting SNPs and assigning candidate risk genes

The iGOAT was constructed by integrating multi-omics data from a variety of different sources. We firstly collected SNPs with genome-wide associations $p < 5 \times 10^{-8}$ from published GWAS studies. In total, we collated 108, 112, 102, 297, 181 and 75 SNPs that were significantly associated with schizophrenia (SCZ), bipolar disorder (BP), and major depressive disorder (MDD), Alzheimer's disease (AD), Parkinson's disease (PD) and migraine (MG), respectively. A detailed description on the method of collecting SNPs is shown in Supplemental Methods. The number of unique SNPs selected for this study in relation to each disorder is shown in Table S1. For each SNP, candidate risk genes were assigned to it by considering SNP-to-gene distances and chromatin interactions with genes. Briefly, a gene was considered to be a candidate for a given SNP if the gene intersected with 1 Mb of the SNP, or the SNP interacted with the transcriptional start site (TSS) of the gene according to Hi-C data.[24] We considered the SNP to be capable of interacting with the TSS if the interaction region in Hi-C data was within 250 kb of the SNP, a distance suggested by a previous study.[25] In total, there were 1756, 1534, 1311, 1958, 1883 and 806 candidate risk genes for SCZ, BP, MDD, AD, PD and MG, respectively.

Multi-omics data used in the study

This study integrated multi-omics data to predict risk genes, including distance to index SNP (DTS), the significance of differential expression (DE) in patients and controls, and the number of enhancer–promoter interactions (EPI). The DTS is the distance in base-pairs from the transcriptional start site of each gene to the position of the index SNP. The DE is the significance level (p-value) of genes expressed differentially in patients and controls by analysis of RNA-seq data from the samples listed in Table S2. The EPI data were collected from two studies.[24,26] We also obtained fetal brain Hi-C data from the paracentral cortex of three individuals at the 17-18th gestation weeks[24] and adult brain Hi-C data from the DLPFC (dorsolateral prefrontal cortex) of three individuals (aged 36, 44 and 64 years, respectively).[27] The DTS gives the distance between the candidate gene and the index SNPs, the DE provides p-values to show the expression difference of the candidate in patients and controls, and EPI provides singles to show interactions between the promoter of candidate gene and the enhancers. Detailed descriptions of all these data were included in Supplemental Material (Supplemental Methods, Tables S2 and S3). All missing data in these data sources were filled in by means of the K-Nearest Neighbors algorithm.

Establishing a weighted gene network by estimating the association between SNP-SNP interactions and a given disorder

We utilized genotype and phenotype data from about 500,000 individuals downloaded from the UK Biobank (UKB)[28] to establish SNP-SNP interactions. The definition of disease was based on the field description in the directory of UKB data (Table S4). Samples with the property "Date of disorder reported" were used to construct the SNP-SNP interaction network. From the UKB, 1089, 1476, 115,031, 1048, 2173 and 3686 individuals were recruited as SCZ, BP, MDD, AD, PD and MG patients, respectively. These samples were filtered by genotype qualification that was controlled by the standard PLINK inclusion procedure[29] (Supplemental Methods). The genotype data from cases and matched controls were used to establish the SNP-SNP interactions network.

For each disorder, we estimated associations between the SNP-SNP interactions with the disorders under a dominant–dominant model (DDM). Under the DDM, a SNP is encoded as MM = 2, Mm = 1, and mm = 0, where MM, Mm, and mm are used to denote the three genotypes of each SNP majority homozygous, heterozygous, and minority homozygous, respectively.

We used the hypergeometric test to measure associations between SNP-SNP interactions and phenotypes as

***Fig. 1:* Schematics of iGOAT**. Step 1: iGOAT assigning candidate genes to SNPs by leveraging SNP-to-gene distance and chromatin interaction profiles (Hi-C). Step 2: iGOAT constructing a Bayesian framework for scoring candidate genes through integrating genotype data, gene ontology and protein–protein interactions with multi-omics data. Step 3: Using scores of genes as sampled probabilities, Gibbs sampling was used to assess associations between gene and disease by iteratively sampling until the frequencies (posterior probabilities) of genes selected as risk genes converged. We then computed empirical p-values for candidate genes by generating an empirical distribution. After p-value correction (Benjamini-Hochberg: BH), genes with $p_{corrected}$ < 0.001 were considered high-risk genes (HRGs); genes with $p_{corrected}$ > 0.5 were considered low-risk genes (LRGs). Step 4: The biological features of HRGs were explored and compared with those of LRGs. GSEA: gene set enrichment analysis.

recommended by a previous study.[15] The effect of a pair of binary-coded SNPs, S_x and S_y, that have genotype T in a case–control cohort was calculated as Equation (1):

$$P_T\left(S_x, S_y, C\right) = 1 - \sum_{f=0}^{X} \frac{\binom{K}{f}\binom{M-K}{N-f}}{\binom{M}{N}} \tag{1}$$

where S_x and S_y are two SNPs; M is the total number of samples; N is the total number of samples in a given class C (phenotype); K is the total number of samples that have genotype T (MM or Mm in a dominant model); X is the total number of samples that have genotype T in class C; the probability of taking any N samples from M samples and the number of samples in the N samples with genotype T in class C exceeds X is calculated by $P_T(S_x, S_y, C)$. $P_T(S_x, S_y, C)$ represents the significance level of the correlation between S_x-S_y and trait C. Then we expand the trait-specific SNP-SNP interaction into a trait-specific gene network. S_x was mapped to gene g_x if S_x was located within 1 kb upstream or downstream of g_i. The weight between gene g_i and g_j in a certain class C is defined as Equation (2):

$$W_{Sij} = max\left\{-log_{10}\left(P_T\left(S_x, S_y, C\right)\right) \middle| \, S_x \xrightarrow{mapped} g_i, \, S_y \xrightarrow{mapped} g_j\right\} \tag{2}$$

Then, a weighted gene network was established, which was denoted as a SNP-SNP network (W_S).

In this study, we found the number of SNP-SNP interactions in the patients and controls is a huge number (e.g., SCZ patients have 5.63×10^9 SNP-SNP interactions in DDM). To reduce the calculation burden, only the dominant–dominant model was used here.

Constructing GO and protein–protein interaction networks

Gene networks used in this study include gene ontology (GO),[30] protein–protein interaction (PPI) from BioGRID,[31] and tissue-specific PPI networks from TissueNET.[32] We computed the weighted matrices for the GO network (W_G), BioGRID PPI network (W_P), and tissue-specific PPI network (W_T), respectively (Supplemental Methods). The weight calculation is to evaluate the difference between these candidate genes annotated with certain functions compared to all the human genes in the same situation. The same strategy was used in our previous study to weight a gene–gene interaction network.[11] The Random Walking with Restart algorithm (RWR) was used to calculate the probability of reaching x_i when starting from a set of selected genes x_{-i} in the weighted networks (SNP-SNP network W_S, GO network W_G, PPI network W_P, and tissue-specific network W_T), respectively (Supplemental Methods).

Construction of a Bayesian framework for the prediction of risk genes

A Bayesian framework was constructed to prioritize the candidate risk genes with known omics and the networks including SNP-SNP interactions, GO networks, and protein–protein interactions. Denote omics data as D and the networks as N. Then, the goal was to maximize $P(x_1, x_2, ..., x_m|D, N)$ so that we can identify most likely risk genes. $[x_1, x_2, ..., x_m]$ denotes a set of genes. $P(x_1, x_2, ..., x_m|D, N)$ denotes the probability of selecting a given set of genes as associated with the disease under the conditions of D and N. More details on the calculation of $P(x_1, x_2, ..., x_m|D, N)$ is shown in Equation (2) of the Supplementary Material. It is unrealistic to directly calculate $P(x_1, x_2, ..., x_m|D, N)$ of each combination case, since there will be a combination explosion. Thus, we decomposed the overall probability into multiple factors (Equation (3)), which was more estimable. The derivation process of Equation (3) is shown in the Supplementary Material. Then we can maximize $P(x_1, x_2, ..., x_m|D, N)$ by selecting m genes with the largest $P(D_{x_i}|x_i)P(x_i|x_{-i}, N)$.

$$P(x_1, x_2, ..., x_m|D, N) \propto \prod\nolimits_{i=1,2,...,m}\left[P\left(D_{x_i}|x_i\right)P(x_i|x_{-i}, N)\right] \tag{3}$$

where

$$P(x_1, x_2, ..., x_m|D, N) = \tfrac{P(N)}{P(D,N)} P(x_1, x_2, ..., x_m|N)$$

$\prod_{i=1}^{m} P(D_i|x_i)$ ($\frac{P(N)}{P(D,N)}$ is a constant if D and N are fixed), and $P(x_1, x_2, ..., x_m|N) \approx \prod_{i=1}^{m} P(x_i|x_{-i}, N)$ was approximated by one-dimensional conditional likelihoods as performed in a previous study.[10] $[x_1, x_2, ..., x_m]$ denotes genes to be selected, and x_{-i} means $(x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_m)$, a set of selected genes. The overview of the framework of this method is shown in Fig. 1.

For the first factor $P(D_{x_i}|x_i)$, we estimated it using gene features described by multi-omics data. Each candidate gene was annotated in terms of multi-omics features (Supplemental Methods). A cumulative distribution of each feature in a set of candidate genes was used to score the genes. In short, if a gene x_i was an element of a candidate gene set for a given disease, the feature j of gene x_i was given by x_{ij}. The percentile rank of x_{ij} in feature j was used to represent the score of the gene x_i for feature j, which was denoted as $p_{x_{ij}}$. $p_{xij} = \frac{C_{x_{ij}}}{M}$, where $C_{x_{ij}}$ is the cumulative frequency of genes with scores higher than or equal to the score of gene x_i on feature j (if the lower score means the higher risk of functional impact on the gene, $C_{x_{ij}}$ is the number of candidate risk genes with scores higher than or equal to the score of the candidate risk gene x_i on feature j), and M

is the total number of candidate risk genes for a specific disorder. The $p_{x_{ij}}$ was converted to a -log scale, and greater -log $(p_{x_{ij}})$ means higher risk of a gene to be disease-associated. The summation of the -log $(p_{x_{ij}})$ yielded the final score of gene x_i (*Equation (4)*). The equation (4) denotes the importance of gene x_i for all the feature, D_i.

Then, we defined the weighted score of D_{x_i} for x_i across each feature to be the sum of $-\,log(p_{x_{ij}})$, which made greater scores mean higher risk. The $p_{x_{ij}}$ was converted to a -log scale, and greater $-log(p_{x_{ij}})$ means higher possibility of a gene to show feature D. D_{x_i} represents the features of gene $\mathrm{x_i}$ and it can be described as vectors of feature scores $(x_{i1}, x_{i2}, \ldots, x_{in})$, and $x_{i1}, x_{i2}, \ldots, x_{in}$ are independent with each other (n is the number of features). The summation of the $-log(p_{x_{ij}})$ yielded the final score of gene x_i to denotes the importance of gene x_i for all the feature D (*Equation (4)*).

$$P\left(D_{x_i}|x_i\right) \approx -\sum_{j=1}^{n} log\left(p_{x_{ij}}\right) \tag{4}$$

where n is the total number of features for gene x_i, and D_{x_i} represents the multi-omics information for gene x_i.

The above omics scores were integrated with the second factor (gene network weights), $P(x_i|x_{-i}, N)$, in which x_{-i} denotes the set of candidate risk genes already selected ($x_i \notin x_{-i}$), and N refers to the constructed weighted gene networks (Supplemental Methods). The Random walking with Restart algorithm (RWR) was used to calculate the distance between the gene x_i and the set of previous-round selected genes x_{-i} in the network, N. $P(x_i|x_{-i}, N)$ can be calculated as the sum of the sub-vectors extracted from N, which are rows for x_i and columns for x_{-i} in N. $P(D_{x_i}|x_i)$ and $P(x_i|x_{-i}, N)$ were integrated together to form a Bayesian prior probability factor $P(x_i|x_{-i}, N)P(D_{x_i}|x_i)$, which means the probability of selecting x_i as risk gene based on x_{-i}, information in D (genomic data) and N (gene networks). $P(x_i|x_{-i}, N)P(D_{x_i}|x_i)$ was used as a sampled weight for gene x_i in Gibbs sampling. Here, Gibbs sampling was used to approximate the posterior probabilities (PPs) of each gene. After Gibbs sampling was converged, according to PPs, a set of genes that were maximized $\prod_{i=1,2,\ldots,m} P(D_{x_i}|x_i)P(x_i|x_{-i}, N)$ were selected as most likely to be associated with the disease. Here, $P(x_1, x_2, \ldots, x_m|D, N) \propto \prod_{i=1,2,\ldots,m} P(D_{x_i}|x_i)P(x_i|x_{-i}, N)$ represents the probability of a set of genes associated with disease. Gibbs sampling was initiated with 30 genes (initial x_{-i}) that were detected as being associated with index SNPs with the lowest $P_{FDR} < 0.05$ according to the cis-eQTL result obtained from the dorsolateral prenatal cortex[33] and trans-eQTL summary data of blood.[34] The sampling probability of each gene was $P(x_i|x_{-i}, N)P(D_{x_i}|x_i)$.

In Gibbs sampling, x_{-i} was initially set as a random set of m genes from candidates ($x_{-i}^{(0)} = [x_1^{(0)}, x_2^{(0)}, \ldots, x_m^{(0)}]$). Then in the next round of sampling, $x_{-i}^{(1)}$ was updated based on $x_{-i}^{(0)}$, D, and N. Using $P(x_i^{(1)}|x_{-i}^{(0)}, N)P(D_{x_i}|x_i)$ as sample weight of gene $x_i^{(1)}$, we sampled another m genes, noted as $x_{-i}^{(1)} = [x_1^{(1)}, x_2^{(1)}, \ldots, x_m^{(1)}]$. At the same time, we recorded the selected frequency (posterior probability) of each candidate gene. Set $Freq_i = \frac{\#\ of\ times\ the\ gene\ x_i\ is\ selected}{\#\ of\ sampling}$, $F = [Freq_1, Freq_1, \ldots, Freq_M]$. In each round, we updated F when x_{-i} changed. Repeat the above sampling process many times until F did not change much in the last two rounds. That is $\left\| F_{last} - F_{penultimate} \right\| < E_{Gibbs}$ ($E_{Gibbs} = 0.01$). F_{last} were the posterior probabilities (PPs) of candidate genes. Then $P(x_1, x_2, \ldots, x_m|D, N)$ was approximated by $\prod_{i=1,2,\ldots,m} P(x_i|x_{-i}^{(last)}, N)P(D_{x_i}|x_i)$.

To address concerns of false positives, like genes associated with many other genes in the network, we built a null distribution of the posterior probability (PP) for each candidate in order to identify real high-risk genes (HRGs) and low-risk genes (LRGs). The genes with $p_{corrected}$ (Empirical p-value) < 0.001 were considered as HRGs, and the genes with $p_{corrected}$ (Empirical p-value) > 0.5 were used as LRGs in this study. The number of HRGs could further reduced by considering the rank of the PP in the candidate genes. More details are given in the Supplemental Methods.

Tissue- and cell-type specificity analysis

iGOAT was evaluated by tissue specificity of the predicted HRGs. The tissue specificity analysis was performed by using several gene expression data from brain tissues. Briefly, we downloaded the median expression levels, Reads Per Kilobase of transcript, per Million mapped reads (RPKM) of genes in brain tissues from GTEx (V8, https://gtexportal.org/home/datasets), and the expression levels of genes in ten brain regions from BrainEAC (http://www.braineac.org/, cerebellar cortex [CRBL], frontal cortex [FCTX], hippocampus [HIPP], medulla [MEDU], occipital cortex [OCTX], putamen [PUTM], substantia nigra [SNIG], thalamus [THAL], temporal cortex [TCTX], and intralobular white matter [WHMT]). For the cell-type specificity analysis, we used a single-cell transcriptomic dataset merged from three sources.[35–37] The dataset contains scRNAseq results from 4249 cell samples associated with 35 cell types from fetal and adult brains. Only genes with expression values > 0 in more than 1% of samples from adult brains were included in the analysis. The details of how the tissue- and cell-type specificity of gene expression were evaluated are given in the Supplemental Methods.

When plotting heatmaps, we used median expression levels of HRGs/LRGs in a cell type to represent the expression level of the entire gene set in that cell type. To compare expression levels across different cell types or gene sets, we normalized the median expression values to obtain "scaled expression".

Gene set enrichment analysis

Brain disorder-related gene sets

iGOAT was further evaluated by enrichment of predicted genes in many brain disorder-related gene sets. We collated 41 brain disorder-related gene sets in total. SCZ-related gene sets were collected by following the procedures described in a previous study.[11] There are seven gene sets (CCS, FMRP.Ascano, FMRP.Darnell, PRAZ, PRP, PSD and SYV) that are involved in synaptic, presynaptic, or voltage-gated calcium-channel functions considered to be associated with the six brain disorders under study. We also collected genes reported to be associated with the six brain disorders from the two datasets, GenCLiP,[38] and DisGeNET.[39] In addition, we obtained 13 disease-associated pathways from KEGG.[40] These pathways were searched in KEGG by using 'Schizophrenia', 'Bipolar disorder', 'Major depressive disorder', 'Alzheimer's disease', 'Parkinson's disease', and 'Migraine' as keywords, respectively. Genes in these pathways were collected since they are associated with various brain disorders. The numbers of genes in the gene sets and abbreviations ascribed to these gene sets are listed in Table S5. The gene set enrichment analysis were performed by GOST function in gprofiler2 package in R4.2. The results with $p_{FDR} < 0.05$, 15≤"term_size"≤600 and "intersection_size" ≥5 were considered as significant.

Genes harbouring de novo variants or rare variants

iGOAT was further evaluated by enrichment of predicted genes in genes harbouring *de novo* protein-disrupting variants causing developmental disorders. Genes that harbour *de novo* protein-disrupting variants causing developmental disorders were obtained from the Deciphering Developmental Disorders Study,[41] which includes 93 developmental disorder (DD) risk genes enriched in damaging *de novo* mutations (DNMs) with genome-wide significance ($p < 5 \times 10^{-7}$). We also obtained 102 Autism spectrum disorder (ASD)-associated genes harbouring a burden of rare variants with $p_{FDR} < 0.1$ in a cohort including 11,986 cases and 23,598 controls.[42] Finally, from a previous study,[43] we extracted 54 SCZ risk genes harbouring an elevated burden of rare variants with $p_{FDR} < 0.3$ by analyzing exomes from 4264 cases and 9343 controls. These variants were employed as a comprehensive dataset of rare variants and *de novo* variants associated with brain disorders.

Inherited disease genes from HGMD

We collected all genes known to harbour heritable mutations either causing or associated with the six brain disorders plus ASD from the Human Gene Mutation Database (HGMD[44]). Here, we only considered genes with mutations in the categories "DM" or "DM?" (Table S6).

We employed the one-sided Fisher's Exact test to evaluate the enrichment of the HRGs in these gene sets compared to LRGs and treated Benjamini–Hochberg corrected $p < 0.05$ (Fisher's Exact test) as being statistically significant.

Spatiotemporal transcriptome profile

The spatiotemporal transcriptome profile was downloaded from GEO (GSE25219[45–47]) and was used to explore the spatiotemporal transcriptome of HRGs in the human brain. GSE25219 constitutes exon-level transcriptome data from 16 brain regions from 1340 samples belonging to 57 postmortem human brains and represents gene expression data from 15 developmental stages of the human brain (ranging from pre-to postnatal development). A detailed description of the methods using these data is included in Supplemental Methods.

Gene co-expression module analysis

The high-risk genes (HRGs) predicted by iGOAT were examined on their enrichment in gene co-expression modules that are associated with the brain disorders. In order to construct the gene co-expression modules, we downloaded the gene expression profiles of the six brain disorders (SCZ, BP, MDD, AD, PD, and MG) and their corresponding controls from the GEO database (Table S7). From the profiles, we removed outliers which were defined as those samples with standardized sample network connectivity Z scores < −2. In total, 2400 samples were used for the analysis after removing 124 samples as outliers. Then, we used the gene expression profiles of all controls (totally 1356 samples) to construct a gene co-expression module by WGCNA.[48] The associations between gene co-expression modules and disease were tested by a linear regression model using contact country, platform id, and tissue type as covariates [R code: lm (ME ~ diagnosis + contact_country + platform_id + tissue_type, data)], where ME is the module eigengene that is the first principal component of the gene expression matrix of the corresponding module. The linear regression model was constructed by using data from all patients and controls. Then, backward best subset selection was performed to identify the best covariates. In the linear regression model of a gene co-expression module and disease tags in all samples, if $\beta > 0$ and $p_{corrected} < 0.05$ (t-test), then the co-expression module was considered to be significantly upregulated in the disease. If $\beta < 0$ and $p_{corrected} < 0.05$ (t-test), we assumed the gene co-expression module to be significantly downregulated in the disease.

Experimentally validating the association between the *MLH1* gene and SCZ

Cell culture

Validation was performed in mouse neural stem cells (NSCs). NSCs that were derived from brain tissues of embryonic day 14.5 C57BL/6 mice were purchased from

Cyagen Biosciences (Guangzhou, People's Republic of China) and were grown in serum-free growth medium (Dulbecco's modified Eagle's medium (DMEM)/F12 1:1; Gibco) containing 1 × B27 Supplement minus vitamin A (ThermoFisher, Cat.NO: 12587010), 1 × N2 supplement (ThermoFisher, Cat.No: 17502048), 20 ng/mL basic fibroblast growth factor (STEMCELL, Cat.No: 78003), 20 ng/mL epidermal growth factor (STEMCELL, Cat.No: 78006), 2 μg/mL heparin (STEMCELL, Cat.No: 07980) and 1% penicillin/streptavidin. The multipotent neurospheres were passaged every 3–4 days to single-cell suspension for continuing growth and further experiments. Cells were cultured at 37 °C with 95% air and 5% CO_2. Mycoplasma tests were performed periodically using specific PCR primers, and no mycoplasma contamination was detected in cells used in this study.

Knockdown MLH1 in mouse NSCs

Two distinct short hairpin RNAs (shRNAs) targeting *MLH1* (MLH1-Sh#1 and MLH1-Sh#2) were constructed using pLKO.1 vector. The 21 bp targeting sequences are MLH1-Sh#1 (5′-GCTAATTCAGATCCAAGACAA-3′) and MLH1-Sh#2 (5′-CCGAAGCATTTCACAGAAGAT-3′). The control scramble shRNA sequence is Control (5′-CCTAAGGTTAAGTCGCCCTCG-3′). Lentiviruses were generated according to the manufacturer's protocol. After 72 h viral infection, cells were treated with puromycin (1 μg/mL) to select NSCs stably expressing indicated shRNAs.

Western blotting

Whole-cell protein extracts were lysed by RIPA lysis buffer (Epizyme, PC101) and centrifuged at 12,000 g for 15 min. The G250 (Beyotime) was used to quantify the protein concentrations. Proteins were separated by 10% SDS polyacrylamide gel electrophoresis, transferred to polyvinylidene difluoride membrane, blocked by 5% bovine serum albumin (BSA) for 1 h, and incubated with primary antibodies overnight at 4 °C. Primary antibodies used include rabbit anti-MLH1 (1:500, Affinity DF6057), rabbit anti-NeuN (1:1000, ABclonal, A19086) and mouse anti-GAPDH (1:10,000, Proteintech, 60004-1-IG). The membrane was then incubated for 1 h at room temperature with the appropriate secondary antibodies (1:1,000, Abclonal, AS-003, AS-014). The chemiluminescence signals were detected with enhanced chemiluminescence (ECL) and quantified by densitometry using the ImageJ software (NIH, Bethesda, USA). At least three independent experiments were carried out and representative results are shown.

NSCs proliferation and differentiation

To investigate the role of *MLH1* in mouse NSCs, we performed proliferation and differentiation assays. The cell proliferation capabilities were examined by the Cell Counting Kit-8 (CCK-8) and EdU incorporation assays. For CCK-8 assay, 1×10^4 cells were plated into 96-well plates (pre-coated with 10 μg/mL laminin). When the cells were cultured for 24 and 48 h, respectively, CCK-8 solution was added to the culture medium, and cells were incubated for 2 h at 5% CO_2, 37 °C. The absorbance at 450 nm wavelength was measured according to the manufacturer's instructions. For EdU assay, 1×10^6 NSCs were plated into 6-well plates. After these cells were cultured for 12 h, 10 μM EdU (APE × BIO, K1078) was added to label dividing cells. After incubation for 3 h, an EdU Flow Cytometry Assay Kit was used to detect EdU-labeled cells as described in the protocol. Briefly, cells were centrifuged, washed by PBS, fixed with 4% PFA for 15 min, permeabilized with 0.5% Triton X-100 in PBS for 20 min, and then incubated with 100 μL Click-iT reaction mix for 30 min. The treated cell suspension samples were analyzed by flow cytometry (646/662 nm).

For the differentiation assay, NSCs were cultured for three days in a differentiation medium that included DMEM/F12 containing 1% penicillin/streptavidin, 1 × N2 supplement (Gibco), 1 × B27 supplement (Gibco), and 2 μg/mL heparin. Immunostaining was performed to count the number of neurons. Briefly, cells were fixed with 4% PFA and permeabilized with PBS containing 0.5% Triton-X100, then blocked with 5% bovine serum albumin (BSA) for 1 h at room temperature. The cells were then incubated with rabbit anti-NeuN antibody (1:50, ABclonal, A19086) overnight at 4 °C. Finally, they were incubated with goat anti-rabbit antibody (1:100, Abbkine, A23420) at room temperature for 1 h. The numbers of NeuN + cells were quantified with Image J software and statistical analysis was performed using a two-tailed Student's t-test. All experiments were performed in three independent assays with at least three replicates per group. Data were represented as mean ± SD. Differences were considered to be statistically significant if $p < 0.05$.

Role of funders

The funders had no role in study design, data collection, analysis, interpretation or writing of the report.

Results

High-risk genes predicted by iGOAT are enriched in brain-disorder associated gene sets and account for a significantly enriched heritability

iGOAT integrated enhancer–promoter interaction (EPI) information from microglia, neurons, and oligodendrocytes, respectively. The numbers of HRGs predicted by iGOAT using the three types of EPI data were listed in Table S8. The total number of HRGs simultaneously identified using three different EPI types are 199, 195, 187, 229, 210, and 122 for SCZ, BP, MDD, AD, PD, and MG, respectively. A Venn diagram is shown in Fig. S1 to describe the number of overlapped genes. We compared the impact of using EPI data from different cell types (microglia, neurons and oligodendrocytes) by analysing the extent of overlap of predicted high-risk genes (HRGs) with brain disorder-related gene sets. As shown in Fig. S2, the extent of overlap of the HRGs with the brain disorder-related gene sets was similar when using EPI from the three different cell types. The EPI data from neurons were selected for use in iGOAT for following analysis.

iGOAT predicted 305, 283, 265, 326, 325, and 192 genes with empirical $p_{corrected} < 0.001$ as high-risk genes (HRGs) associated with SCZ, BP, MDD, AD, PD and MG, respectively (Table S8), and 1298, 1123, 931, 1470, 1406, and 554 genes ($p_{corrected} > 0.5$) as low-risk genes (LRGs) associated with these six disorders, respectively. Only about 10–49% of the intergenic SNPs were predicted to impact the nearest HRGs, whilst 32–60% of the intronic SNPs were predicted to impact HRGs residing nearest to them (Fig. 2a), highlighting the importance of using 3-D genomic evidence and SNP-SNP interactions to link noncoding index SNPs to candidate genes irrespective of whether they are proximal or more distant.

HRGs predicted by iGOAT were significantly enriched in brain disorder-associated genes from PSD, FMRP and DisGeNET datasets (Methods) compared to the LRGs (Fig. 2b and Table S9). More importantly, the predicted HRGs associated with five disorders SCZ, BP, MDD, PD and MG explained greater heritability enrichment ($Enrichment$ = 29.90, 20.07, 22.96, 47.04, and 24.26, and nominal $p = 3.66 \times 10^{-12}$, 3.06×10^{-5}, 2.79×10^{-5}, 9.02×10^{-2}, 4.27×10^{-2}, and 3.80×10^{-2}, respectively. Supplemental Methods) than the LRGs (Fig. 2c). No significant heritability enrichment was observed for HRGs predicted to be associated with AD ($Enrichment$ = 138.38, nominal $p = 9.02 \times 10^{-2}$). The enrichment score was calculated by the equation: Enrichment = $\frac{h^2_{HRG}/h^2_{All}}{SNP_{HRG}/SNP_{All}}$, where h^2_{HRG} and h^2_{All} represent the heritability explained by SNPs around HRGs and by all SNPs in 1000 Genomes Phase 3, respectively,

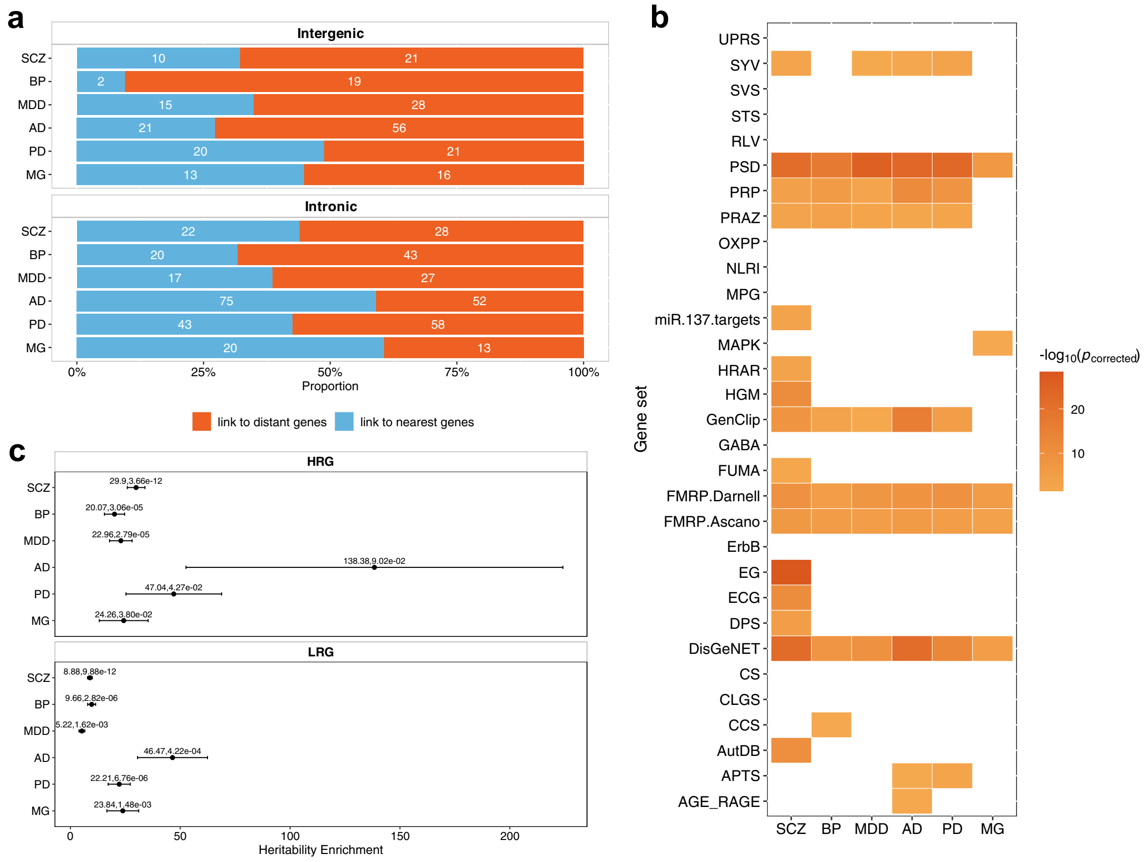


Fig. 2: **The performance of iGOAT in predicting HRGs associated with brain disorders**. (a) The proportion of intronic and intergenic SNPs mapped to the nearest and distant HRGs. The exact numbers of SNPs linked to the HRGs were shown in the middle of the bar. (b) Enrichment of the HRGs compared to LRGs in neurological disorder-related gene sets. Only significant results were shown. The intensity of shading denotes the degree of significance. (c) Disease heritability enrichment of the HRGs and LRGs. The enrichment value and Fisher's exact test p-value are given above the bar. The dots represent the enrichment values, whereas the error bars indicate standard errors.

and where SNP_{HRG} and SNP_{All} denote the number of SNPs around HRGs and the total number of SNPs in 1000 Genomes Phase 3, respectively (More details are shown in Supplementary Material). The method of analysing the disease heritability enrichment was shown in Supplemental Methods.

The sample size may influence the quality of construction of the SNP-SNP interaction network, and may further influence the performance of iGOAT. We enlarged the sample sizes of AD cases and controls to construct SNP-SNP interaction network by considering the family history of the individuals. By using this SNP-SNP interaction network, iGOAT identified 329 HRGs, most of which (87.8%) overlapped with the HRGs only considering the AD patients. The detailed results are shown in Supplementary Material (Fig. S3). The sensitivity of iGOAT to window length for defining candidate genes was tested by using five different distance parameters: 1 kb, 10 kb, 100 kb, 1 Mb and 2 Mb to generate candidate genes for SCZ. As shown in Table S10, iGOAT using different distances parameters provided similar percentages of HRGs known to be SCZ-related genes (ranging from 86.39% to 81.76%). Longer distance leads to a higher number of candidates and slightly lower coverage (Table S10). Thus, iGOAT is robust to varying choices of gene window length. We also tested the ability of iGOAT in controlling false positives by using null SNP signals. The result indicated that the HRGs identified by using the simulated SNPs are not significantly enriched with SCZ-related genes compared to the HRGs identified by using the SNPs associated with SCZ (Fig. S4). The detailed results are shown in Supplementary Material.

Incorporating epigenetics data reveals the neurodevelopmental origin of brain disorders and improves the predictive power of iGOAT

Since enhancer–promoter interactions (EPI) are highly tissue-specific,[24] we reasoned that the inclusion of EPI from different brain developmental stages in iGOAT might help to predict HRGs that reflect the neurodevelopmental origin of brain disorders. When we replaced the childhood (child Hi-C) EPI used in iGOAT with fetal Hi-C and adult Hi-C (see Methods) respectively, the expression levels of the HRGs in the whole brain developmental stages were similar to those identified using EPI from the three different developmental stages (Fig. 3a and e). When we examined the expression of the most significant (top 1% of) HRGs over the different brain developmental stages, we observed that using the EPI from different brain developmental stages influences the detected gene expression trajectories in brain development stages (Supplemental Results and Fig. S5).

We then examined the heritability enrichment scores of the HRGs predicted by using EPI derived from the fetal Hi-C, child Hi-C or adult Hi-C data, and found that the HRGs associated with the psychiatric disorders, and predicted using EPI derived from the child Hi-C data, have higher heritability enrichment scores than when the fetal Hi-C data were used. The HRGs associated with PD or MG, and predicted using the child Hi-C data, showed lower heritability enrichment scores than when the fetal Hi-C data were used (Fig. 3b). Thus, using EPI derived from child Hi-C data may serve to identify HRGs associated with psychiatric disorders that are associated with higher heritability enrichment than using EPI derived from other brain developmental stages. By contrast, using EPI derived from child Hi-C data may serve to identify HRGs for PD and MG, which show lower heritability enrichment than when using EPI derived from fetal Hi-C data.

We next explored the impact of EPI on the prediction of HRGs. We simply compared the HRGs predicted by iGOAT using EPI to the HRGs predicted by iGOAT without using EPI. Those HRGs which were predicted by iGOAT using EPI but were not predicted as HRGs by iGOAT without using EPI, were termed neuronDRGs, whereas the HRGs predicted by iGOAT without using EPI but were not predicted as HRGs by iGOAT using EPI data were termed noHiCDRGs. The LRGs predicted by both were termed cLRGs. The numbers of neuronDRGs, noHiCDRGs, and cLRGs identified are listed in Table S11. The neuronDRGs associated with all disorders have higher extents of overlap with brain disorder-related genes than the noHiCDRGs (Fig. 3c, one-sided t-test, $p = 0.24$, $t = 2.62$) and a significantly higher proportion of brain disorder-related genes than that of cLRGs. Specifically, the neuronDRGs associated with SCZ, BP, MDD and MG were enriched in more brain disorder-related gene sets compared to cLRGs than the noHiCDRGs (Fig. 3d). Moreover, the neuronDRGs exhibit greater tissue specificity in all 13 brain tissues from GTEx, and the three brain regions from BrainEAC than the cLRGs (Fig. S6a and b). By contrast, no noHiCDRGs were found to have greater tissue specificity in brain regions from the BrainEAC dataset than the cLRGs (Fig. S6b). We also observed that the neuronDRGs associated with PD and MG showed cell-type specificity in OPC, Oligo, and Micro compared to cLRGs, whereas noHiCDRGs only specifically expressed in one cell type (Oligo for PD and Endo for MG. Fig. S6a). In cell expression profile, the neuronDRGs associated with SCZ, MDD, AD, PD, and MG exhibited higher expression levels in Astro, Endo, OPC, Micro, Neuron, and Oligo than the noHiCDRGs and the cLRGs, suggesting that the neuronDRGs have potential functions in brain cells, especially neurons (Fig. S6c). Over the various developmental stages, the expression of the neuronDRGs was higher than the expression of noHiCDRGs and cLRGs (Fig. 3e). The function enrichment analysis of neuronDRGs was performed by the GOST function in gprofiler2 package in R4.2. The neuronDRGs for psychiatric disorders are enriched in

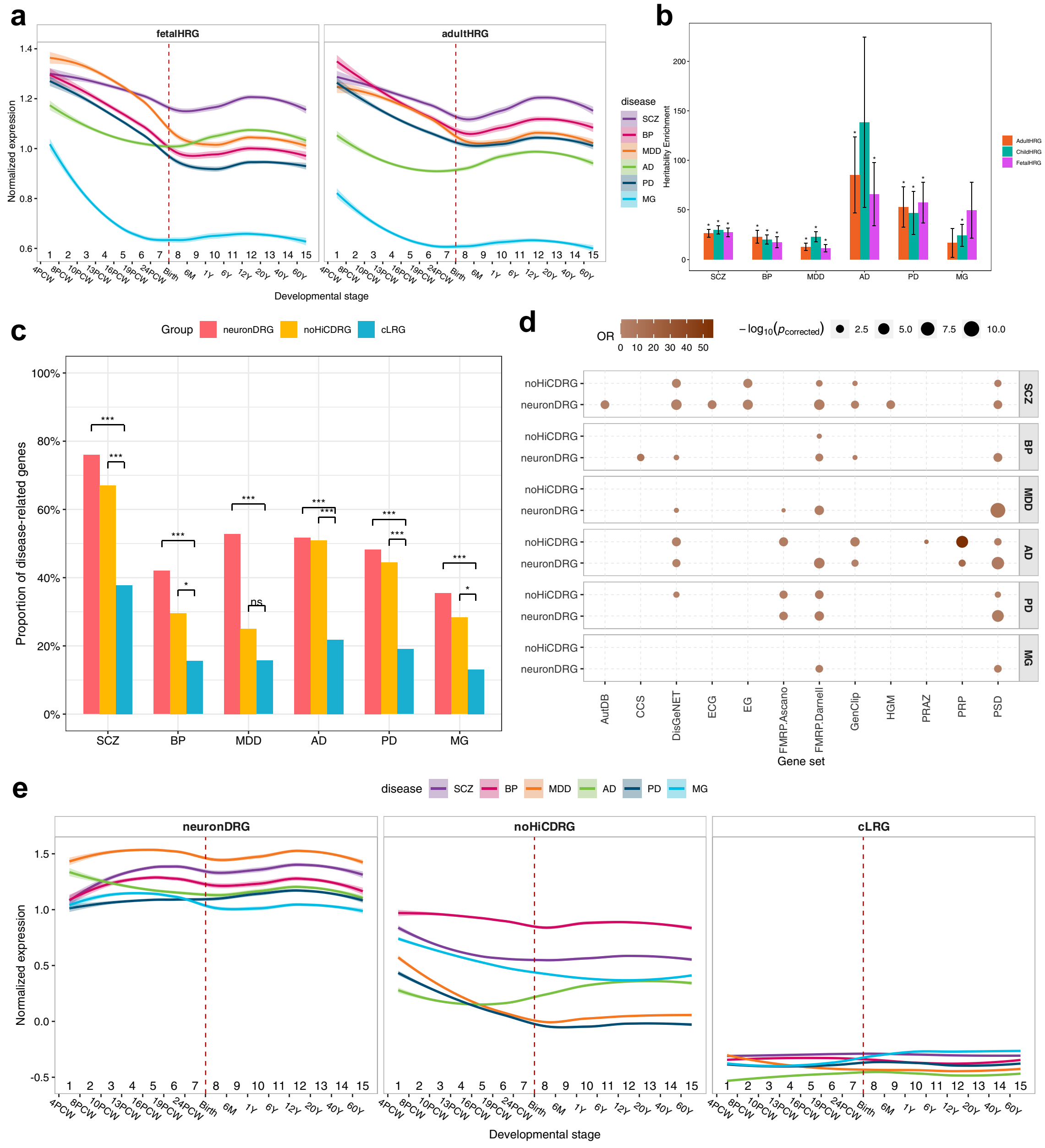


Fig. 3: **Effects of EPI data from the human fetus, child and adult on iGOAT**. (a) Developmental expression trajectory of the HRGs predicted by iGOAT using EPI data from the fetal (fetalHRG) and adult (adultHRG), respectively. (b) Heritability enrichment of HRGs predicted by iGOAT using EPI from fetus, child or adult. The error bars indicate the standard errors of the enrichment. The "*" denotes significant enrichment ($p_{Enrichment} < 0.05$). (c) Proportion of disease-related genes in neuronDRGs, noHiCDRGs, and cLRGs. The significance was obtained using one-sided Fisher's exact test comparing the number of disease-related genes in neuronDRGs/noHiCDRGs with that of cLRGs. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ns: not significant. (d) Enrichment of neuronDRGs and noHiCDRGs in brain disorder-related gene sets using the cLRGs as background genes. Only significant results are shown here. (e) Average expression levels of neuronDRGs, noHiCDRGs and cLRGs during the stages of brain development.

synaptic related functions, such as regulation of synaptic plasticity ($p_{adjust} = 1.30 \times 10^{-3}$, Fisher's exact test), modulation of chemical synaptic transmission ($p_{adjust} = 9.89 \times 10^{-4}$, Fisher's exact test), regulation of trans-synaptic signalling ($p_{adjust} = 9.95 \times 10^{-4}$, Fisher's exact test) and long-term synaptic potentiation ($p_{adjust} = 1.90 \times 10^{-3}$, Fisher's exact test) (Table S12). The neuronDRGs for neuron degeneration disorders are enriched in neural tube related functions, such as neural tube development ($p_{adjust} = 1.29 \times 10^{-2}$, Fisher's exact test), neural tube closure ($p_{adjust} = 2.86 \times 10^{-2}$, Fisher's exact test), and primary neural tube formation ($p_{adjust} = 3.22 \times 10^{-2}$, Fisher's exact test). They are also enriched in post-synapse organization ($p_{adjust} = 1.56 \times 10^{-2}$, Fisher's exact test) (Table S12). In comparison, the noHiCDRGs for psychiatric disorders and neuron degeneration disorders were not found to be enriched with any neuron function- or brain function-related GO term. Taken together, these results indicate the key role that EPI data had in allowing iGOAT to detect HRGs specifically expressed in brain tissues and brain cells, and more active in brain development.

Incorporating SNP-SNP interaction data improves the predictive power of iGOAT

The evaluation of the impact of SNP-SNP interactions on iGOAT was performed by excluding SNP-SNP interactions from iGOAT and instead predicted HGRs and LRGs by integrating the GO, BioGIRD PPI and Tissue-specific PPI networks with omics data. We termed this approach siGOAT. The HRGs predicted by iGOAT with SNP-SNP interactions but not predicted by siGOAT were termed nnDRGs. The HRGs predicted by siGOAT but not predicted by iGOAT were termed snDRGs. The common LRGs predicted by both siGOAT and iGOAT were termed cLRGs. The numbers of nnDRGs, snDRGs, and cLRGs identified are listed in Table S11. Compared to snDRGs, the nnDRGs are more likely to be brain disorder-related genes (Fig. S7a). The nnDRGs associated with SCZ, MDD, PD, and MG were enriched in more brain disorder-related gene sets than snDRGs when compared to cLRGs (Fig. S7b). In tissue specificity analysis, the nnDRGs of all disorders have greater specificity in more than ten types of brain tissue than the cLRGs whereas only the snDRGs associated with AD showed more specificity in over ten types of brain tissue than the cLRGs (Fig. S7c). The tissue specificity analysis on brain regions from BrainEAC also indicated that the nnDRGs have greater specificity in more brain regions than the snDRGs did when compared to cLRGs (Fig. S7d). Meanwhile, the nnDRGs associated with six brain disorders were found to be expressed higher in neuronal subtypes than the snDRGs (Fig. S8a), and the nnDRGs associated with BP, MDD, AD and PD manifested higher expression levels in Astro, OPC and neuronal cells than the snDRGs (Fig. S8b). By comparison, the snDRGs associated with BP, AD and MG tended to show higher expression levels in Endo, Micro, and Oligo than the nnDRGs. During brain development, the nnDRGs associated with BP and MDD were expressed more highly than the snDRGs and cLRGs (Fig. S8c). The expression of the nnDRGs increased over time with brain development whereas the expression of the snDRGs decreased with brain development. The function enrichment analysis indicated that nnDRGs are enriched in biological terms related to synapse organization, regulation of synapse, and synapse assembly and protein dephosphorylation (Table S13).

We also evaluated the impact of SNP-SNP interactions when only networks were used in iGOAT (Supplemental Results), and found higher precision in predicting brain disorder-related HRGs by the network including SNP-SNP interactions than the network not using SNP-SNP interactions (Table S14 and Fig. S9). Taken together, SNP-SNP interactions served to improve the ability of iGOAT to predict HRGs specifically or highly expressed in brain tissues, multiple brain regions and neurons.

Combining candidate genes associated with multiple SNPs improves the predictive power of iGOAT

iGOAT evaluated the associations of a group of candidate genes with the set of SNPs according to the null distribution of the features of all the candidate genes (Methods). This strategy differs from that used in previous studies[10,11] in which each SNP was assumed to affect only one gene, and the associations between SNPs and genes were assessed by considering only the genes mapping to one particular SNP. iGOAT using this method when sampling was termed Transformed iGOAT (TiGOAT). We compared iGOAT to TiGOAT in terms of their performance in predicting HRGs associated with the six brain disorders by inputting SNPs (GWAS $p < 5 \times 10^{-8}$) mapping to no more than five candidate genes by the two distinct methods. As shown in Table 1, higher proportions of HRGs associated with SCZ, MDD, AD, and MG predicted by iGOAT were included in brain disorder-related gene sets (50–100%) than those predicted by TiGOAT (40–83%). When we further compared iGOAT and TiGOAT in terms of their ability to predict HRGs by inputting SNPs (GWAS $p < 5 \times 10^{-8}$) mapping to no more than three, seven or nine candidate genes (Table S15) by the two distinct methods, we consistently observed that iGOAT predicted a higher proportion of HRGs overlapping with

Disorder	≤3 candidates		≤5 candidates		≤7 candidates		≤9 candidates	
	iGOAT[a]	TiGOAT[b]	iGOAT	TiGOAT	iGOAT	TiGOAT	iGOAT	TiGOAT
SCZ	100.00%	80.00%	100.00%	83.33%	95.00%	84.62%	93.55%	88.89%
BP	100.00%	80.00%	50.00%	54.55%	50.00%	46.67%	55.56%	52.63%
MDD	100.00%	83.33%	93.75%	71.43%	68.97%	66.67%	61.90%	70.00%
AD	100.00%	60.00%	100.00%	72.73%	87.50%	64.71%	85.71%	72.73%
PD	0%	0%	53.85%	62.50%	55.00%	61.54%	48.28%	57.14%
MG	0%	0%	100.00%	40.00%	62.50%	44.44%	60.00%	50.00%

"≤*n* candidates" refers to SNPs mapped by no more than *n* candidate genes. The proportion denotes the extent of overlap of HRGs with brain disorder-related gene sets. [a]iGOAT: HRGs associated with SNPs (≤*n* candidates) predicted by iGOAT. [b]TiGOAT: HRGs associated with SNPs (≤*n* candidates) predicted by TiGOAT.

Table 1: **Proportion of HRGs overlapping with brain disorder-related gene sets in terms of SNPs that mapped to a small number of candidates.**

known brain disorder-related genes than TiGOAT (Table 1). Thus, iGOAT, by jointly evaluating gene associations with a certain brain disorder through combining candidate genes associated with multiple SNPs, has improved the accuracy of the predictions compared to the method considering genes mapped to one individual SNP.

Tissue specificity, cell-type specificity, and expression level of the predicted HRGs at different brain developmental stages

We observed that all the HRGs identified in the six brain disorders under study had higher Tissue-specificity scores (TS scores) ($p_{corrected} < 2.93 \times 10^{-4}$, Wilcoxon test) in 13 types of brain tissue from GTEx data (Fig. 4a) than the LRGs, which was suggestive of the potential impact of HRGs on brain activity. We also evaluated the Tissue-specificity of the HRGs compared to LRGs in some tissues less related to brain disorders, like blood, artery, testis and kidney (Fig. S10). We found that the HRGs have significantly higher TS scores than the LRGs in blood ($p_{corrected} < 5.75 \times 10^{-7}$, Wilcoxon test), artery (aorta, coronary and tibial; $p_{corrected} < 1.99 \times 10^{-8}$, Wilcoxon test) and adipose (visceral and subcutaneous; $p_{corrected} < 9.95 \times 10^{-10}$, Wilcoxon test) but not in testis ($p_{corrected} > 0.89$, Wilcoxon test) and kidney ($p_{SCZ} = 0.11$, $p_{MDD} = 0.07$, Wilcoxon test). When we respectively used the TS scores of HRGs in blood, artery, adipose, testis and kidney as a negative control to compare the TS scores in the brain tissues (Wilcoxon test), we found that HRGs in the cerebellum and cerebellar hemisphere have significantly higher TS scores than some unrelated tissues. TS scores of HRGs in kidney cortex ($p_{corrected} < 0.020$, Wilcoxon test) and blood ($p_{corrected} < 0.027$, Wilcoxon test) were significantly lower than those in most brain tissues (Fig. S11).

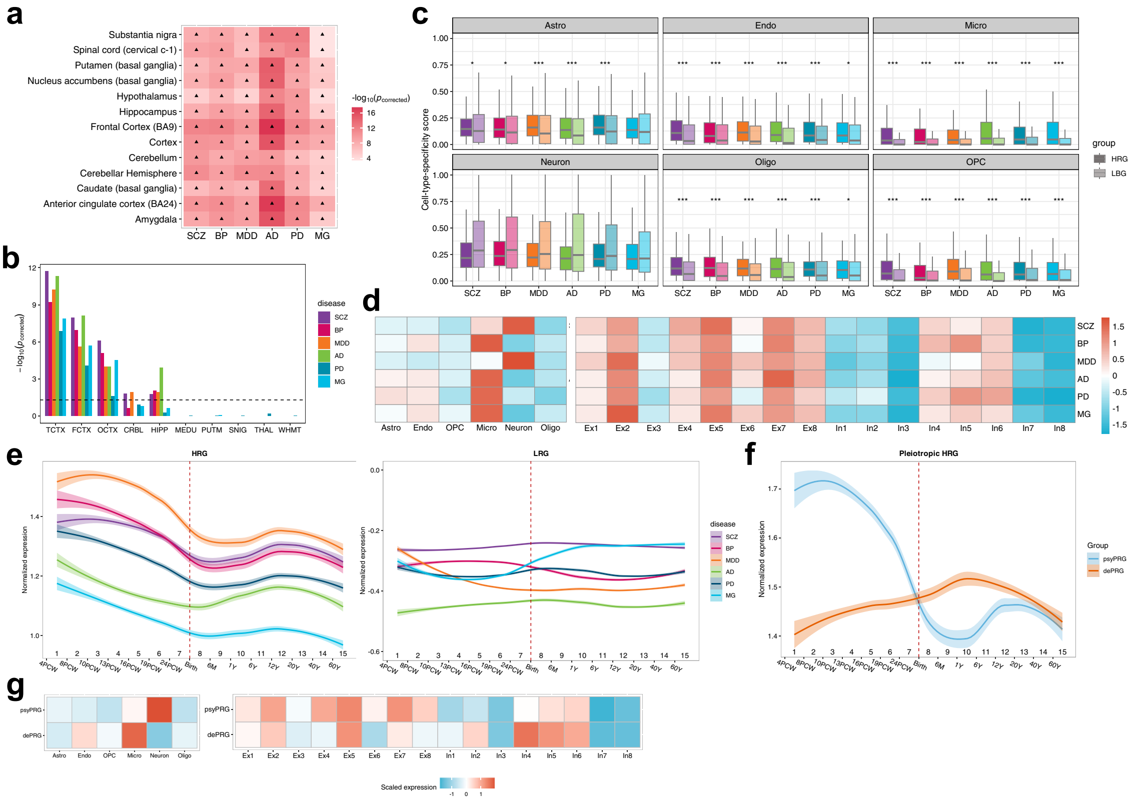


Fig. 4: **Analysis on the expression of HRGs, LRGs and pleiotropic HRGs**. (a) Tissue specificity analysis on HRGs compared to LRGs using the GTEx dataset (Methods). "▲" denotes significant result (Fisher's exact test $p_{corrected}$ < 0.05). (b) Tissue specificity analysis on HRGs compared to LRGs using expression profiles from BrainEAC (Methods). (c) Cell specificity analysis of HRGs compared to LRGs (Methods). The plot on the right shows the cell-type expression levels of HRGs and LRGs in subtypes of neurons ("*" denotes the expression of HRGs). The colour indicates the column-scaled average expression levels of genes in that cell type. Astro, astrocytes; Endo, endothelial cells; Micro, microglia; Oligo, oligodendrocytes; OPC, oligodendrocyte progenitor cells; In, inhibitory neurons; Ex, excitatory neurons. (d) Comparison of the median expression levels of HRGs across different cell types. The colour indicates the row-scaled median expression level of genes in that disorder. (e) Developmental expression trajectories of the HRGs and LRGs during the entire brain developmental stages. The dashed line represents birth. The shaded aeras represent the 95% confidence interval of expression levels. (f) Expression trajectories of the psyPRGs and dePRGs during the brain developmental stages. (g) Row-scaled average cellular expression profiles of the psyPRGs and dePRGs in brain cells and neuron subtypes.

Performing the analysis using BrainEAC data (Fig. 4b) indicated the HRGs associated with all six brain disorders exhibited significantly higher TS scores ($p_{corrected} < 0.025$, Wilcoxon test) in the temporal cortex (TCTX), frontal cortex (FCTX), and occipital cortex (OCTX) than the LRGs. Further analysis of the cell-type specificity indicated the HRGs associated with SCZ, BP, MDD, AD and PD have more cell-type specificity in OPC, Oligo, Micro, Endo and Astro than the LRGs. The HRGs associated with MG have more cell-type specificity in OPC, Oligo, Micro and Endo than the LRGs (Fig. 4c).

We then sought to determine the average expression levels of the HRGs and the LRGs in specific cell types. As shown in Fig. S14, the expression of the HRGs associated with the six brain disorders in specific cell types are all significantly higher than the LRGs. We further compared their expression levels in neuronal subtypes, and found that the HRGs were expressed much more highly than the LRGs (Fig. S12). Further, the cell specificity analysis indicated that the HRGs are more specifically expressed in neurons comparing to OPC ($p_{corrected} < 5.03 \times 10^{-4}$, Wilcoxon test) and Oligo ($p_{corrected} < 0.031$; Fig. S13) but more specifically expressed in Astro ($p_{corrected} < 0.014$, Wilcoxon test) and Micro ($p_{corrected} < 7.45 \times 10^{-10}$, Wilcoxon test) than neurons. Using cellular expression profiles of the HRGs in six brain cell types, we found that the median expression levels of the HRGs associated with AD, PD, BP and MG have the greatest median expression in microglia while the HRGs associated with SCZ and MDD have the greatest median expression in neurons (Fig. 4d). This result reflected the functional difference of the risk genes associated with neurodevelopmental and neurodegenerative disorders. Comparison of the median expression levels of the HRGs in 16 neuronal subtypes indicated HRGs associated with SCZ, BP, MDD, AD and MG were preferentially expressed in the excitatory neurons rather than in the inhibitory neurons (Fig. S16b) ($p_{corrected} < 3.45 \times 10^{-3}$Student's t− test), whereas the LRGs associated with MDD and MG were more highly expressed in the inhibitory neurons than in the excitatory neurons (Student's t-test $p_{corrected} < 3.68 \times 10^{-2}$, Fig. S16a and b).

Further analysis of the expression levels of the HRGs in brain developmental stages indicated that the HRGs associated with the six brain disorders showed remarkably similar expression patterns with decreases during stages 1–7 (prenatal stage from 4 weeks prior to birth), slight increases at stages 9–12 (infancy and childhood) and a peak at stage 12 (12Y $\leq$ age <20Y, adolescence). We noted that the HRGs associated with the three psychiatric disorders exhibited higher expression levels than the HRGs associated with AD, PD, or MG during brain development (Fig. 4e). When we investigated the expression of LRGs, we found that they exhibited much lower expression than the HRGs in the brain developmental stages whilst no obvious change in the expression of LRGs was noted during brain development (Fig. 4e).

Pleiotropic HRGs associated with brain disorders show different expression profiles during brain development

We next analysed pleiotropic HRGs (PRGs), defined as being associated with two or more psychiatric disorders, or two neurological diseases; these gene sets were termed psyPRGs (91 genes) and dePRGs (71 genes), respectively (Table S16). The expression profiles of these genes during brain development are different. As shown in Fig. 4f, the expression of psyPRGs during the prenatal stages (1–7) is much higher than during the postnatal stages (8–15), showing a decrease during stages 3–9, and a slight increase in stages 10–12. By contrast, the expression of dePRGs is higher during the postnatal stages than in the prenatal stages, showing a trend to increase during stages 1–9 and then a trend to decrease during stages 10–15 (Fig. 4f). These results evidence the elevated activity of HRGs associated with psychiatric disorders during fetal stages of brain development, whereas the HRGs associated with neurological diseases displayed elevated activity only after the infant stage. The expression profiles of the psyPRGs and dePRGs were found to expressed higher in Neurons and Microglia, respectively (Fig. 4g). Specifically, in neuronal subtypes, the psyPRGs exhibited high expression in almost all excitatory neurons whilst the dePRGs did not; the dePRGs displayed high expression in certain inhibitory neurons.

When we calculated the heritability enrichment of psyPRGs and dePRGs in psychiatric disorders and degenerative disorders, we found significant enrichment of psyPRGs in SCZ ($Enrichment = 44.44$, nominal $p = 2.62 \times 10^{-3}$), BP ($Enrichment = 32.30$, $p = 2.20 \times 10^{-3}$) and MDD ($Enrichment = 44.01$, nominal $p = 1.99 \times 10^{-3}$). While dePRGs showed no significant disease heritability enrichment in any degenerative disorder ($Enrichment_{AD} = 33.26$, $p_{AD} = 0.34$; $Enrichment_{PD} = 51.99$, $p_{PD} = 0.08$; $Enrichment_{MG} = 28.11$, $p_{MG} = 0.14$). The higher p values of enrichment for dePRGs in degenerative disorders may be due to the lower number of genes in dePRGs (71; 91 in psyPRGs). Gene function analysis on psyPRGs and dePRGs were shown in the Supplementary Material and Fig. S17. The result indicated that the psyPRGs are enriched with several development-related terms, such as stem cell differentiation ($p_{adjust} = 0.013$), cell morphogenesis involved in neuron differentiation ($p_{adjust} = 0.046$), and positive regulation of cell development ($p_{adjust} = 1.16 \times 10^{-3}$) (Fig. S17). In comparison, the dePRGs are enriched in positive regulation of cell death ($p_{adjust} = 8.28 \times 10^{-3}$), and cellular response to topologically incorrect protein ($p_{adjust} = 1.22 \times 10^{-5}$) whereas no development-related terms were enriched by dePRGs (Fig. S17).

The predicted HRGs potentially disclosing shared mechanisms between brain disorders

We next explored if the gene co-expression modules participated by the HRGs are significantly shared among the six brain disorders. We firstly constructed the gene co-expression modules using brain tissue-based RNA-seq data from controls (Method section) through WGCNA analysis. From these modules, we identified those modules significantly dysregulated in the brain disorders by using the RNA-seq data of brain tissues from both the patients and the controls through a logistic regression model (Methods section). If one module is significantly dysregulated in one brain disorder and enriched with the HRGs associated with another brain disorder, this module is considered to represent the shared mechanisms between two brain disorders.

In total, the logistic regression identified 16 gene co-expression modules significantly [$|\beta| > 0$ and $p_{corrected} < 0.05$ (t-test)] dysregulated in the brain disorders (Fig. 5a). Among them, module MEpink was found to be significantly upregulated in MDD (= 1.42×10^{-3} and $p_{corrected-MDD} = 5.04 \times 10^{-3}$, t-test) and downregulated in neurodegenerative disorders (AD and PD, $\beta_{AD} = -3.30 \times 10^{-3}$ and $p_{corrected-AD} = 1.05 \times 10^{-8}$, t-test; $\beta_{PD} = -5.12 \times 10^{-3}$ and $p_{corrected-PD} = 1.14 \times 10^{-7}$ t-test). Further analysis indicated that MEpink was significantly (Fisher's Exact test $p < 0.05$) enriched by the HRGs associated with BP, MDD, AD and PD compared to LRGs. When the heritability enrichment analysis was performed for MEpink module after excluding the HRGs associated with specific disorders, we found it had not achieved the significance for MDD (nominal $p = 0.33$), AD (nominal $p = 0.36$) and PD (nominal $p = 0.74$) (Fig. S18). Besides, module MElightgreen was found to be downregulated in BP and MDD ($\beta_{BP} = -1.97 \times 10^{-3}$ and t-test $p_{corrected-BP} = 0.017$; $\beta_{MDD} = -3.14 \times 10^{-3}$ and t-test $p_{corrected-MDD} = 2.83 \times 10^{-4}$), and upregulated in AD and PD ($\beta_{AD} = 9.24 \times 10^{-3}$ and t-test $p_{corrected-AD} = 7.10 \times 10^{-22}$; $\beta_{PD} = 5.95 \times 10^{-3}$ and t-test $p_{corrected-PD} = 3.87 \times 10^{-4}$), and was enriched by HRGs associated with AD as compared to the LRGs. However, after removing the HRGs, the module MElightgreen was not found with significant heritability enrichment for BP (nominal $p = 0.26$), MDD (nominal $p = 0.51$), AD (nominal $p = 0.15$), PD (nominal $p = 0.35$). Interestingly, we found that the HRGs associated with psychiatric disorders were enriched in the dysregulated gene-expression

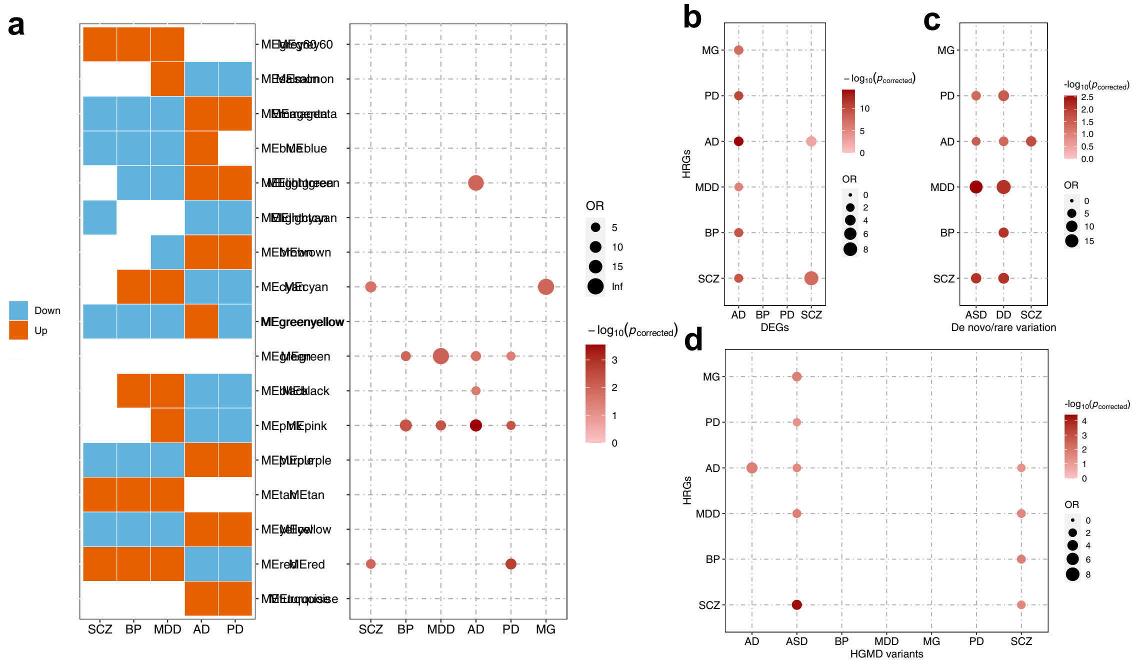


Fig. 5: **Shared genetic mechanisms of HRGs associated with the six brain disorders**. (a) Dysregulated gene modules associated with diseases are coloured in the left-hand block. The enrichment of HRGs compared to LRGs in the gene modules is depicted in the right-hand block. "ME" refers to the module eigengene. (b) Enrichment analysis of HRGs in differentially expressed genes (DEGs) in AD, BP, PD or SCZ patients and controls using LRGs as background genes. (c) Enrichment analysis of HRGs in genes harbouring *de novo*/rare variants associated with ASD, DD or SCZ using LRGs as background genes. (d) Enrichment analysis of HRGs in genes carrying disease-causing mutations in AD, ASD, BP, MDD, MG, PD or SCZ from HGMD by using LRGs as background genes. The intensity of shading denotes the degree of significance. The size of the dot indicates the magnitude of the odds ratio (OR) in the one-sided Fisher's Exact test to evaluate the enrichment of HRGs in the gene sets as compared to LRGs.

modules associated with neurodegenerative disorders (MEcyan, MEred, and MEpink) and vice versa (MEblack, MEpink, MElightgreen, MEred, and MEcyan). When we explored the biological function enriched by MEpink, we found that genes in MEpink are enriched by presynaptic endocytosis ($p_{adj} = 3.35 \times 10^{-3}$), synaptic vesicle cycle ($p_{adj} = 0.028$), regulation of GTPase activity ($p_{adj} = 0.033$) and chromatin remodeling ($p_{adj} = 0.046$) functions (Table S17). Furthermore, we found that the HRGs associated with all six brain disorders were enriched in genes that were differentially expressed (DEGs) between AD patients and controls (Fig. 5b). The HRGs associated with SCZ were enriched in DEGs of SCZ. The gene expression difference analysis identified 6 DEGs (|log2(Fold change)| > 0 and t-test $p_{corrected} < 0.05$] for BP, among which none are HRGs nor LRGs of BP. In similar vein, 79 DEGs were found for PD, among which three genes are HRGs of PD. The number of overlapping genes between HRGs and DEGs for PD were marginally higher (Fisher's exact test $p = 0.049$) than the LRGs. Thus, the enrichment of HRGs associated with psychiatric disorders in the modules dysregulated in neurodegenerative diseases may well reflect shared neurological mechanisms between these conditions.

The shared mechanisms between the six brain disorders were further explored by searching for the enrichment of HRGs identified by dint of their harbouring rare variants related to SCZ or ASD, or *de novo* mutations related to developmental disorders (DD; see Methods). The HRGs associated with SCZ, MDD, AD, and PD were found to be enriched in genes known to harbour rare variants related to ASD and DD (Fig. 5c, Table S18). The HRGs associated with AD were also enriched in genes harbouring rare variants associated with SCZ. A Venn diagram in Fig. S19a depicts the number of HRGs overlapping with those genes harbouring rare variants associated with brain disorders (SCZ, ASD, or DD).

We next explored whether HGMD pathogenic variants [disease-causing mutations (DM) or likely pathological mutations (DM?)][44] were enriched in the HRGs (see Methods). We found that the HRGs associated with five of the disorders under study here (SCZ, MDD, AD, PD, and MG) were all enriched among the genes known to carry HGMD pathogenic variants related to ASD (Fig. 5d). The HRGs identified as being associated with SCZ, BP, MDD and AD were also enriched in HGMD pathogenic variants related to SCZ. A Venn diagram in Fig. S19b depicts the number of HRGs overlapping with genes harbouring pathogenic variants in HGMD associated with seven brain disorders (SCZ, BP, MDD, ASD, AD, PD or MG). Thus, the pathogenic variants harboured by the HRGs are potentially indicative of shared genetic architecture between the brain disorders.

Comparing iGOAT with eQTL-based gene annotation tools and gene-level genome-wide association analysis

Using SCZ as an example, we compared the output of iGOAT with two eQTL-based approaches, coloc[27] and TWAS,[49] and a recently developed a gene-level genome-wide association analysis, H-MAGMA.[19] Since both TWAS and coloc were obtained from the adult human DLPFC, we compared these two methods with results of iGOAT using adult Hi-C data and H-MAGMA results derived from the adult brain (v1.08). We detected a significant overlap between the HRGs identified by iGOAT and the three other sets of genes. Indeed, the HRGs were significantly more likely to overlap with the genes predicted by coloc and TWAS than the LRGs (Fig. 6a. $p = 3.76 \times 10^{-3}$, $OR = 1.87$ [coloc]; $p = 2.68 \times 10^{-4}$, $OR = 2.60$ [TWAS], Fisher's Exact test).

Then, we compared the precisions of these methods in predicting SCZ-associated genes. H-MAGMA predicted 6664 SCZ-associated genes, including 187 genes that iGOAT identified as HRGs, which represented a significant enrichment compared to LRGs ($p = 4.32 \times 10^{-11}$, $OR = 2.35$, Fisher's exact test). We calculated the precision of the top predictions generated by these four methods (rank by p values). As shown in Fig. 6b, iGOAT exhibited a precision of 100% for its top 19 predictions, and a precision larger than 85% for its top 200 predictions. TWAS showed a comparable precision with coloc, which did not exceed 65%. H-MAGMA predicted more risk genes but with lower precisions (<45%) than other methods.

In total, 328 genes were predicted to be associated with SCZ by coloc or TWAS but were not predicted as HRGs by iGOAT, whereas 255 genes were predicted to be HRGs associated with SCZ by iGOAT but were not predicted as risk genes associated with SCZ by coloc or TWAS. These two sets of genes were termed eQTL-only and iGOAT-only, respectively. The iGOAT-only gene set was found to be significantly enriched in nine SCZ-related gene sets compared to the eQTL-only gene set (Fig. 6c). Briefly, 80.00% of the iGOAT-only genes were present in the SCZ-related gene sets whereas only 39.33% of the eQTL-only genes were present in the SCZ-related gene sets. When comparing the expression levels of iGOAT-only and eQTL-only genes in the developmental stages of brain, we observed higher expression levels of iGOAT-only than eQTL-only genes (Fig. 6d).

A recent TWAS study[50] has identified 67 SCZ-associated genes, termed nMHCGs. Among them, 51 genes were not identified as HRGs by iGOAT, which were termed nMHCG-only genes. Out of 305 HRGs identified by the iGOAT, 289 did not show significance in the TWAS study, and which were termed HRG-only genes. From the HRG-only genes, 83.93% were in the SCZ-gene database while only 56.72% nMHCG-only genes were in the SCZ-gene database. Further analysis

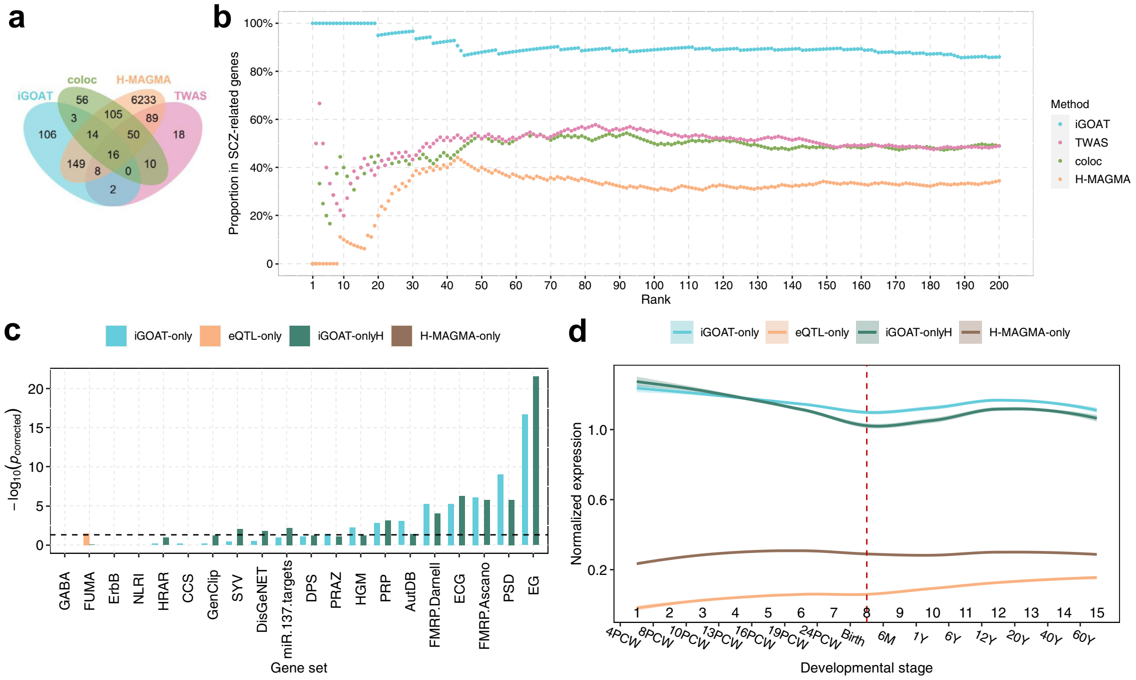


Fig. 6: **Comparing iGOAT with other risk gene annotating methods, coloc, TWAS, and H-MAGMA**. (a) The number of overlapping genes between SCZ-associated HRGs identified by iGOAT and those identified by coloc, TWAS and H-MAGMA. (b) Proportion of SCZ-related genes in the top-rank predictions. (c) Gene set enrichment analysis of iGOAT-only, eQTL-only, iGOAT-onlyH, and H-MAGMA-only. The blue bar represents the analysis on the iGOAT-only compared to eQTL-only whereas the orange bar represents the analysis on the eQTL-only compared to iGOAT-only. The green bar represents the analysis on the iGOAT-onlyH compared to H-MAGMA-only whereas the tan bar represents the analysis on the H-MAGMA-only compared to iGOAT-onlyH. The dashed line means $-log_{10}(0.05)$. (d) Normalized expression levels of the iGOAT-only and eQTL-only, iGOAT-onlyH and H-MAGMA-only in brain developmental stages. The blue, orange, green and tan solid lines represent the mean expression levels of the iGOAT-only, eQTL-only, iGOAT-onlyH and H-MAGMA-only respectively, and the shaded areas represent the 95% confidence intervals of the expression levels.

indicated that the HRG-only genes were enriched ($p_{corrected} \leq 4.93 \times 10^{-2}$, Fisher's exact test) in six SCZ-related gene sets compared to the nMHCG-only genes (Fig. S20). The nMHCG-only genes were not enriched in any SCZ-related gene sets as compared to the HRG-only genes (Fig. S20). However, these two sets of genes have not shown tissue-specificity differences whether using HRG-only as background genes ($p_{corrected} > 0.20$, Fisher's exact test) or using nMHCG-only as background genes ($p_{corrected} > 0.33$, Fisher's exact test) in the BrainEAC dataset. A similar conclusion was drawn when they were compared on the GTEx dataset ($p_{corrected} > 0.34$ using HRG-only as background genes, and $p_{corrected} > 0.99$ using nMHCG-only as background genes).

Further, we extracted risk genes ($FDR < 0.05$) detected by H-MAGMA,[19] a gene-level genome-wide association analysis incorporating chromatin interaction profiles and GWAS summary statistics to predict risk genes for brain disorders, and compared these genes with HRGs predicted by iGOAT. Compared to the HRGs predicted by iGOAT, 111 HRGs were not predicted as risk genes by H-MAGMA, whereas 6477 risk genes were predicted by H-MAGMA but were not predicted as HRGs by iGOAT. These genes were termed iGOAT-onlyH and H-MAGMA-only, respectively. Compared to H-MAGMA-only, iGOAT-onlyH were enriched in ten SCZ-related gene sets (Fig. 6c) with 81.98% SCZ-related genes while the value for H-MAGMA-only was 23.48%. Additionally, we observed that iGOAT-onlyH genes were characterized by higher expression levels during all brain developmental stages than H-MAGMA-only genes (Fig. 6d).

Taken together, iGOAT is a more powerful tool for the detection of true disease-related genes and predicted HRGs highly expressed in all brain developmental stages than coloc, TWAS, and H-MAGMA.

HRGs identified by iGOAT and experimental validation of an association between *MLH1* gene and SCZ

In total, iGOAT predicted that 1393 HRGs (empirical $p_{corrected} < 0.001$) were associated with at least one of the six brain disorders studied here. The biological functions implicated by these HRGs are listed in Supplemental Results (Fig. S21 and Table S19). We additionally found all the HRGs were significantly more likely to interact with drugs than the LRGs ($p_{corrected} \leq 7.23 \times 10^{-3}$, $OR \geq 1.76$, Table S20, Fisher's exact test), suggesting that the HRGs are likely to be potential drug targets.

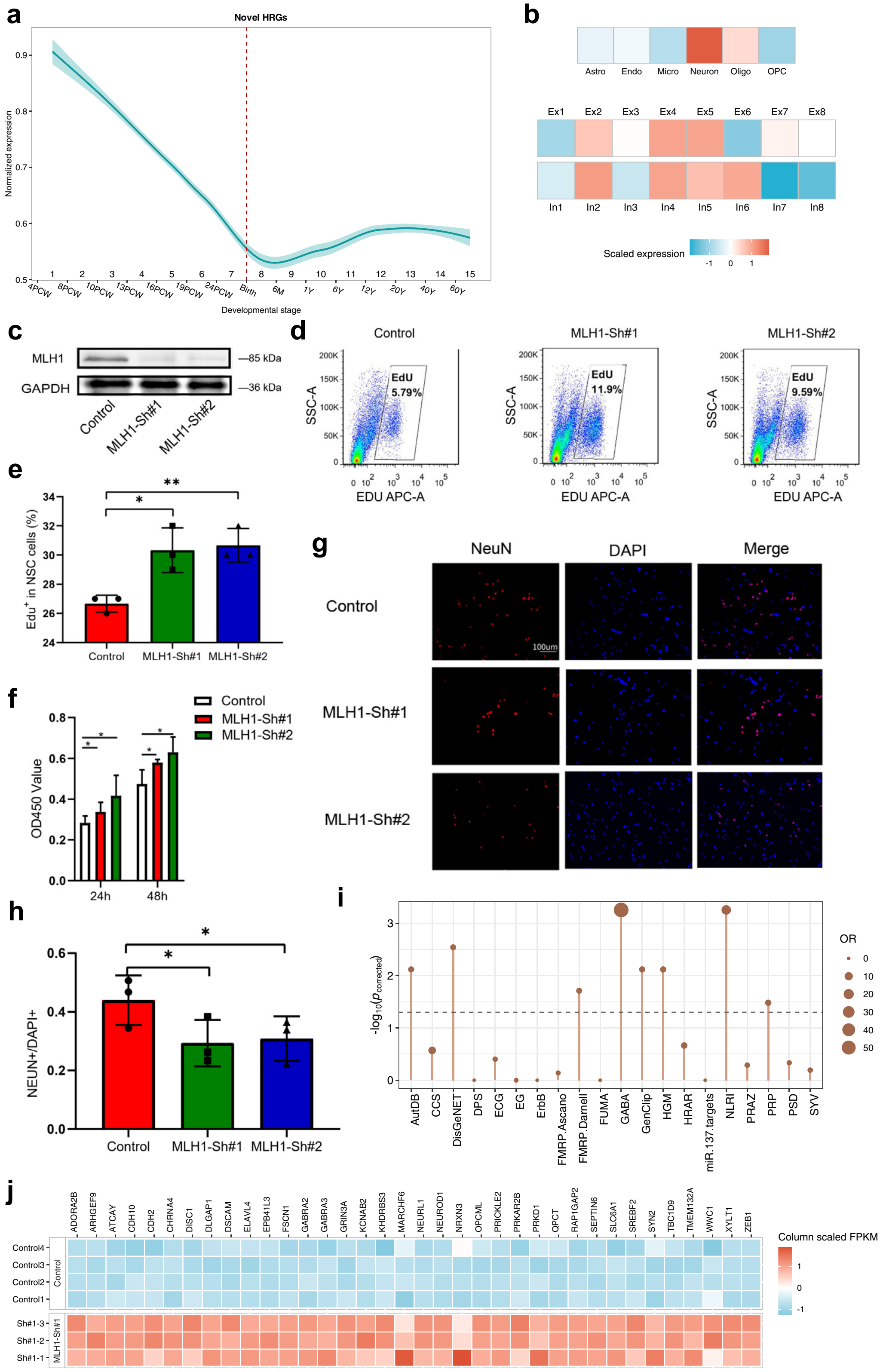
Among these HRGs, 205 genes were not included in any of the brain disorder-related gene sets (Table S21). The expression level of these HRGs in brain development was higher during the prenatal stage than the postnatal stage[51] (Fig. 7a). Moreover, cell-type profiling indicated that these HRGs exhibited higher expression levels in neurons than in other cell types (Fig. 7b). Among these genes, *MLH1* ($PP = 0.0016$, empirical $p < 10^{-4}$) was the top-ranked gene predicted to be associated with SCZ. The *MLH1* gene, located at chromosome 3p22.2, plays a crucial role in genome integrity by replacing mis-paired nucleotides during DNA replication. As a key component of the DNA mismatch repair (MMR) system, inactivated *MLH1* can affect growth regulation and apoptosis-related genes,[52] and overexpression of *MLH1* induces apoptosis and/or a mutator phenotype.[53] Therefore, maintaining precise control over the cellular levels of *MLH1* is essential for ensuring genome stability. Despite evidence suggesting that reduced MLH1 expression can result in genomic instability and tumorigenesis, little is known about its potential contribution to neurodevelopmental disorders. Huckins et al., using conditional analyses, identified independent associations between *MLH1* and SCZ.[50] However, no experiments have been performed to validate the association between *MLH1* and SCZ.

Experimental validation of the association of *MLH1* with SCZ was performed using neural stem cells (NSCs) from mouse. We validated the identity of the isolated NSCs with a Stemness marker (*NESTIN*) for NSCs (Fig. S22a). We first knocked down the expression of *MLH1* in the mouse NSCs, and then examined the knockdown efficiency of the designed shRNAs by western blotting (Fig. 7c and Fig. S22b). To assess the effect of *MLH1* on the proliferation of these cells, we performed EdU incorporation and CCK-8 assays. Both assays showed that *MLH1* knockdown significantly inhibited the proliferation of NSCs (Fig. 7d, e, and f). We further investigated the functional roles of *MLH1* in NSC differentiation and found that the proportion of neuronal nuclei (NeuN, a marker of mature neurons) positive cells, was significantly decreased in *MLH1* knockdown groups compared with controls (Fig. 7g and h). Consistently, the expression levels of NeuN protein were significantly decreased in *MLH1* knockdown groups compared with controls (Fig. S22c and d), indicating that the differentiation abilities of NSCs were impaired by *MLH1* knockdown. Overall, these experiments indicated the important role of *MLH1* in the proliferation and differentiation of NSCs, two important neurodevelopmental processes that have been frequently reported to be affected by SCZ risk genes.[54–57]

The association of *MLH1* and SCZ was further examined by RNA sequencing (RNA-seq) of murine NSCs. The NSCs were divided into two groups, a *MLH1* knockdown group and the control group. RNA-seq data analysis indicated that 66 human homologous genes (DEGs) were expressed differentially ($|\log(FC)| > 0.30$ and $FDR < 0.05$) in the two groups. We further explored the enrichment of these 66 DEGs in SCZ-related gene sets compared to background genes ($FDR > 0.1$, Fisher's exact test), and found that they were significantly enriched in eight gene sets (Fig. 7i), including one SCZ-related pathway ($p_{corrected} = 5.52 \times 10^{-4}$, Fisher's exact test) and a literature-reported set of genes (GenCLiP and DisGeNET; $p_{corrected} = 7.58 \times 10^{-3}$ and 2.86×10^{-3}, Fisher's exact test), suggesting the potential role of *MLH1* in SCZ. The expression profile of DEGs overlapping with the eight significant gene sets are shown in Fig. 7j. We observed significantly higher expression levels of DEGs in *MLH1* knockdown cells than in control cells.

Discussion

Here, we present an unsupervised method, iGOAT, that predicts high-risk genes (HRGs) associated with various neurological conditions. By integrating gene networks and multi-omics features, we identified HRGs for each disorder. These genes were consistent with results from other studies. In AD, we found AD-related HRGs were enriched in GO terms related to inflammation[58,59] (regulation of inflammatory response), immune system[60] (immune response-activating signalling pathway, immune response-regulating cell surface receptor signalling pathway, immune system development) and glial cell,[61] which have been reported to relate to AD. In PD, HRGs associated with PD include monogenic PD gene *LRRK2*, which was identified in GWAS as a risk locus for sporadic PD. Mutations in *LRRK2* make a large contribution to both sporadic and familial forms of PD. The HRGs were enriched in functions related to proteolytic stress, which underlines nigral pathology in PD.[62] In SCZ, SCZ-related HRGs were enriched in processes related to synapse and immune system, which are key factors in the development of SCZ.[63] Moreover, SCZ-related HRGs were also enriched in dopaminergic function that highly related to SCZ.[64] In BP, BP-related HRGs were enriched in functions related to mitochondrial and oxidative stress. Researchers have found evidence on mitochondrial abnormalities in BP.[65,66] There is evidence to show oxidative damage in proteins in BP[67,68] and oxidative stress plays a key role in the pathophysiology of BP.[69] Moreover, dysfunction in the endoplasmic reticulum-related stress response may be associated with BP and illness progression[70] and HRGs associated with BP were enriched in response to endoplasmic reticulum stress. Overall, HRGs were enriched


a
Novel HRGs
Normalized expression
0.9
0.8
0.7
0.6
0.5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
4PCW 8PCW 10PCW 13PCW 16PCW 19PCW 24PCW Birth 6M 1Y 6Y 12Y 20Y 40Y 60Y
Developmental stage
b
Astro Endo Micro Neuron Oligo OPC
Ex1 Ex2 Ex3 Ex4 Ex5 Ex6 Ex7 Ex8
In1 In2 In3 In4 In5 In6 In7 In8
Scaled expression
-1 0 1
c
MLH1
GAPDH
—85 kDa
—36 kDa
Control
MLH1-Sh#1
MLH1-Sh#2
d
Control
MLH1-Sh#1
MLH1-Sh#2
SSC-A
200K 150K 100K 50K 0
EdU 5.79%
EdU 11.9%
EdU 9.59%
0 10^2 10^3 10^4 10^5
EDU APC-A
e
Edu+ in NSC cells (%)
34 32 30 28 26 24
*
**
Control MLH1-Sh#1 MLH1-Sh#2
f
OD450 Value
0.8 0.6 0.4 0.2 0.0
24h 48h
*
Control
MLH1-Sh#1
MLH1-Sh#2
g
NeuN
DAPI
Merge
Control
MLH1-Sh#1
MLH1-Sh#2
100um
h
NEUN+/DAPI+
0.6 0.4 0.2 0.0
*
Control MLH1-Sh#1 MLH1-Sh#2
i
-log10(pcorrected)
3 2 1 0
OR
0 10 20 30 40 50
AutDB CCS DisGeNET DPS ECG EG ErbB FMRP.Ascano FMRP.Darnell FUMA GABA GenClip HGM HRAR miR.137.targets NLRI PRAZ PRP PSD SYV
j
ADORA2B ARHGEF9 ATCAY CDH10 CDH2 CHRNA4 DISC1 DLGAP1 DSCAM ELAVL4 EPB41L3 FSCN1 GABRA2 GABRA3 GRIN3A KCNAB2 KHDRBS3 MARCHF6 NEURL1 NEUROD1 NRXN3 OPCML PRICKLE2 PRKAR2B PRKD1 QPCT RAP1GAP2 SEPTIN6 SLC6A1 SREBF2 SYN2 TBC1D9 TMEM132A WWC1 XYLT1 ZEB1
Control4
Control3
Control2
Control1
Control
Sh#1-3
Sh#1-2
Sh#1-1
MLH1-Sh#1
Column scaled FPKM
1 0 -1

in many disorder-related gene sets and functions affecting neurons and synapses, which suggested the close connection between HRGs and these disorders.

iGOAT is an approach to construct a Bayesian framework integrating enhancer–promoter interaction (EPI) information with other genomics data including SNP-SNP interactions. We compared iGOAT with eQTL-based gene annotation tools and gene-level genome-wide association analysis, and indicated iGOAT is a more powerful tool for the detection of brain-disorder related genes (Fig. 6). iGOAT can be modified to predict genes associated with other types of heritable disease by including features pertaining to that specific disease. The general approach promises to be especially suitable for diseases where a large amount of omics data (e.g., epigenomic, transcriptomic, genomic) are available.

The impact of EPI and SNP-SNP interactions on iGOAT was evaluated by excluding sequentially these two sources of information. We found that using EPI in iGOAT increases its ability to identify genes specifically expressed in brain tissues and regions (Fig. S6a and b). In similar vein, the use of EPI data increased the probability of revealing genes that are highly expressed in multiple brain cell types. However, using EPI iGOAT predicted HRGs associated with BP or AD were found to be expressed in Astro, Endo, OPC, Micro, Neuron and Oligo at a lower or comparable level than the HRGs predicted by iGOAT without using EPI (Fig. S6c). The underlying reason may be that the EPI data used in this study were derived from the neurons of 11 different individuals aged between 5 months and 18 years, and hence do not represent all the EPI information present in brain cells, especially in different brain developmental stages. Many studies have indicated that microglia play a significant role in AD. Our study also found that the HRGs associated with AD expressed significantly higher in microglia than in other cell types (except neurons, $p \leq 4.96 \times 10^{-5}$, Wilcoxon test Fig. S23). When we only used EPI data of microglia in iGOAT, we observed the median expression level of HRGs associated with AD is significantly ($p_{corrected} \leq 1.35 \times 10^{-5}$, Wilcoxon test) higher in microglia and neurons than in other cell types (Fig. S24). This finding suggests that including the EPI data from the specific cell type having important roles in the brain disorders is helpful in identifying the HRGs highly expressed in that cell type. Moreover, as shown in Fig. S6c, including EPI of neurons in iGOAT can lead to the discovery of more genes highly expressed in neurons. For diseases that are not unrelated to neurons, we suggest that users should exclude the EPI of neurons when they run iGOAT using the data and the code in Github. More comprehensive EPI information is required to further improve the predictive ability of iGOAT in the context of identifying HRGs mapped in various cell types.

iGOAT potentially revealed the neurodevelopmental origin of the brain disorders, when it was applied to psychiatric disorders and neurological disorders. There is a long-standing debate as to whether or not adult-onset brain disorders have a neurodevelopmental origin.[71] It has become widely accepted that disturbances that occur early in brain development can contribute to the pathogenesis of SCZ later in life.[72] Applying iGOAT to the psychiatric disorders SCZ, BP and MDD, we found that the expression of HRGs associated with these disorders exhibited potential spatiotemporal and developmental convergence during brain development from the prenatal to postnatal stages (Fig. 4e). Similar conclusions were reached for AD, PD and MG, although the burden of these disorders increases with age. When we examined the expression of the top 1% most significant HRGs associated with AD, PD and MG (Fig. S5), we found that the HRGs associated with AD and MG showed gradual increases across the lifespan using child and adult Hi-C data. This finding is consistent with the result in H-MAGMA,[19] and suggests that HRGs with different levels of significance may have different expression trends during brain development. When we excluded the pleiotropic genes for the HRGs, in total 243, 226, 197, 278, 263 and 150 HRGs were only associated with SCZ, BP, MDD, AD, PD and MG, respectively. The temporal patterns of these HRGs were found to be like the HRGs not excluding the pleiotropic genes (Fig. S25). This result illustrated that the HRGs associated with AD and PD indeed exhibit a similar temporal expression pattern during brain development, which are not the same as the pleiotropic genes associated with both. However, HRGs associated with PD decreased during development. Although PD is generally associated with advanced age, it is possible that the initial neuronal damage occurs at a relatively early stage of brain development.[73] One study has even shown that epigenetic factors operating during fetal brain

Fig. 7: **Expression feature of HRGs and experimental validation of *MLH1*.** (a) The expression level of the HRGs at different stages of human brain development. (b) Row-scaled cell-type expression of the HRGs. (c) Western blot of *MLH1* in NSCs infected with *MLH1* shRNA containing lentivirus or empty control lentivirus. (d) Edu assay showing that *MLH1* knockdown slightly enhanced proliferation of NSCs. (e) Quantification data for (d). (f) CCK-8 assay showing that *MLH1* knockdown promoted significant proliferation of NSCs. n = 6 for each group. *p* values were calculated using the two-tailed Student's t-test. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. (g) *MLH1* promotes differentiation of NSCs. *MLH1* knockdown decreased the differentiation of NSCs into NeuN-positive neurons. (h) Quantification data for (g). (i) Enrichment analysis of DEGs (identified by analysing RNA-seq of murine NSCs) in SCZ-related gene sets. The point size represents the OR value in a one-sided Fisher's exact test. The dashed line was $-\log_{10}(0.05)$. (j) Expression profile of DEGs that were significantly enriched in SCZ-related gene sets. The FPKM values were log2 transformed and scaled for each gene.

development may result in PD later in life.[74] The findings of the present study provide support for the concept of a link between PD and brain development in the fetus, on the basis that the HRGs associated with PD appear to play an important role in the early stages of brain development. In this study, the genes predicted as HRGs by iGOAT using EPI but not predicted as HRGs by the method without using EPI, were termed neuronDRGs. As shown in Fig. 3e, the neuronDRGs were expressed much more highly than all the cLRGs and noHiCDRGs (HRGs predicted by method without using EPI), and they have not shown obvious expression patterns in any brain developmental stage. These results indicated that the EPI data are helpful in identifying genes highly expressed at the whole brain developmental stages.

However, this study has found the insignificant enrichment score explained by HRGs for AD. The reason may be due to the low heritability of AD. When we used LDSC to estimate the proportion of the heritability explained by all SNPs of AD, h_{AD}^2 reached 0.0045 that is much lower than the heritability of other diseases ($h_{SCZ}^2 = 0.363$, $h_{BP}^2 = 0.251$, $h_{MDD}^2 = 0.0311$, $h_{PD}^2 = 0.247$, and $h_{MG}^2 = 0.0275$). When we estimated the heritability explained by SNPs around HRGs, the $h^2{}_{AD}$ was reduced to 0. Thus, the low heritability of AD is a potential reason for the enrichment score explained by the HRGs being close to random. Besides the low heritability of AD, other factors may also influence the heritability enrichment score. When we estimated the heritability enrichment of nonHiCDRs and neuronDRGs for AD, we found that noHiCDRGs ($Enrichment = 185.003$, $p_{corrected} = 0.023$) were more enriched in terms of heritability than neuronDRGs ($Enrichment = 0.199$, $p_{corrected} = 0.971$), suggesting EPI in iGOAT influences the identification of HRGs enriched with heritability of AD. To examine if the EPI from microglia can improve the heritability enrichment of HRGs for AD, we performed iGOAT only using EPI data of microglia to predict HRGs (microDRGs). The result indicated that microDRGs ($Enrichment = 26.337$, $p_{corrected} = 0.239$) have no significant enrichment of AD heritability. The underlying reason for the low heritability enrichment of HRGs for AD is hard to ascertain based on the current analysis. We may obtain clear interpretation after we have more EPI data and more powerful GWAS of AD in the future.

Additionally, applying iGOAT to multiple brain disorders yielded clues as to the shared mechanisms between psychiatric disorders and several neurodegenerative diseases. Of special note, we found that the AD-associated gene co-expression modules (MEpink and MEcyan) are all enriched by HRGs that are associated with BP, MDD or SCZ (Fig. 5a). In total, 133, 133, 106, 146, 146 and 86 HRGs associated with SCZ, BP, MDD, AD, PD, and MG overlapped with differentially expressed genes (DEG) of AD. We performed function enrichment analysis to identify the GO terms enriched by these genes. These HRGs associated with SCZ and AD are all enriched in GO terms related with neuron projection, regulation of synapse structure or activity, DNA repair, regulation of DNA metabolic process and learning or memory (Table S22). The HRGs associated with BP and overlapping with DEGs of AD are enriched in GO terms related with DNA repair, regulation of DNA metabolic process and RNA transport, the HRGs associated with MDD and overlapped with DEGs of AD are enriched in GO terms pertaining to learning or memory and calcium ion transport, the HRGs associated with AD and MG are all enriched in GO terms related with neuron differentiation, neuron projection and regulation of synapse structure or activity. More detailed results are shown in Table S22. The HRGs associated with SCZ, BP, MDD and PD are enriched in rare variants related to developmental disorders (DD) (Fig. 5c). This finding may reflect shared genetic architecture between AD/PD and psychiatric disorders.[75–78]

The HRGs associated with SCZ are not enriched with SCZ *de novo*/rare mutations. The underlying reason is the limited number of SCZ-associated *de novo*/rare mutations. One study[43] collected 118 rare mutations that were obtained by analysing exomes from 4264 cases and 9343 controls, and had an elevated burden related to SCZ with $p_{FDR} < 0.3$. These mutations were harboured by 54 genes. A recent study has shown a higher burden of rare loss-of-function variants in individuals with severe, and extremely treatment-resistant SCZ, but did not find that any rare variants located in genes reached genome-wide significance with the gene-level burden test.[79] This previous study performed four collapsing analyses to show the association of the rare variants with SCZ; from each analysis we selected the top 10 significant rare variants and compared them to the HRGs obtained in our study. These rare variants mapped to 32 genes. Only 4 out of 32 genes were SCZ candidate genes (1,756) of our study. The most recent meta-analysis[80] of 24,248 cases, 97,322 controls and *de novo* mutations from 3402 trios implicates ten genes in which ultra-rare coding variants (URVs) are associated with SCZ and 32 genes at an FDR <5%. Out of these 32 genes, two (*GRIN2A* and *STAG1*) are indicated as HRGs by our method that has identified 305 HRGs and 1298 LRGs for SCZ. Using the LRGs as background genes, the binomial test has shown that the HRGs significantly enriched with more genes (binomial test, $p = 0.023$) resided by the rare variants associated with SCZ. There are two additional studies focusing on detecting genes resided by rare variants associated with SCZ. One study[81] has analysed the exomes of 12,332 unrelated Swedish individuals, which include 4877 affected with schizophrenia. The analysis identified 244,246 coding-sequence and splice-site ultra-rare variants (URVs) that were unique to individual Swedes. However, the single gene burden analysis has not found individual genes significantly enriched more disruptive or damage rare variants (dURVs) in cases comparing to controls exome widely. Similarly, our method only identified one (*KCNH7*) out of the genes resided by the dURVs with burden test $p < 0.01$ as HRGs of SCZ, indicating the HRGs not significantly (binomial test, $p = 0.76$) enriched with dURVs when comparing to the LRGs. Additional study[82] has presented an approach for detecting risk genes of SCZ by analysing rare variant, and have identified SCZ risk genes using extTADA. However, only 24 candidate risk genes were identified with $FDR < 0.3$ and two genes were individually significant at $FDR < 0.05$. Among these 24 genes, only *HSPA8* gene was identified as HRGs by our method. In summary, our method has defined candidate genes as those close to common variants associated with SCZ, which is hard to identify genes significantly enriched with rare variants associated with SCZ. To enhance the predictive capabilities of our methodology, it is imperative to broaden its application to forecast disease–gene associations that are influenced by rare genetic variants in proximity to genes.

ASD is highly genetically heterogeneous and may be caused by both inherited and *de novo* variants. We found that HRGs for MDD are enriched in *de novo* mutations associated with ASD. Indeed, many studies have indicated that depression or depressive symptoms are prevalent in ASD patients.[83,84] Another study through MTAG analysis indicated shared genetic loci between ASD and MDD.[85] Additionally, ASD has shown strong genetic correlations with MDD.[86] The HRGs for MDD were also found to be enriched with *de novo* mutations associated with DD. Many studies have revealed that developmental characteristics increase the risk of MDD.[87,88] Common variants in the *RBFOX1* gene and 22q11.2 region have been indicated as being associated with both MDD and DD.[89,90] Few studies have investigated the shared *de novo* variants of MDD with ASD and DD. More investigations are required to validate the roles of these *de novo* mutations in MDD.

A previous study[20] has investigated the shared genetics between eight psychiatric traits (major depressive disorder (MDD), bipolar disorder (BP), schizophrenia (SCZ), anxiety, post-traumatic stress disorder (PTSD), alcoholism, neuroticism, and insomnia), and five neurodegenerative diseases (Alzheimer's disease (AD), Lewy body dementia (LBD), frontotemporal dementia (FTD), amyotrophic lateral sclerosis (ALS), and Parkinson's disease (PD)), which has found the genetic correlations between AD and MDD, BD, PTSD, and other psychiatric diseases but not identified any genetic correlations between PD and the psychiatric diseases. This previous study did not investigate the genetic correlations between PD and ASD. Here, we found the HRGs associated with PD were enriched with *de novo* mutations associated with ASD (Fig. 5c).

In this study, we have provided a score (posterior probability, PP) and an empirical P-value to evaluate the significance of one gene associated with disease. Higher PP and lower *p*-value indicated a gene having a higher probability to be associated with the disease, and lower probability for the gene to be random. More stringent thresholds will lead to a lower false positive rate but a higher false negative rate. When we used a stringent threshold to select the top 0.5% HRGs, we observed an increasing trend of HRGs for AD, and a decreasing trend of HRGs for BP, MDD and PD during the life span (Fig. S26). This observation is similar to the result for analysing the top 1% of HRGs. Users can select genes according to the study requirement. If they need to find more HRGs for validation, a lax threshold is useful.

This methodology has many limitations. First, SNP-SNP interactions were constructed by identifying paired SNPs using $r^2 < 0.1$ as a cutoff to exclude SNPs exhibiting a high level of linkage disequilibrium. Although this cutoff is commonly used, it may oversimplify the complexities of linkage disequilibrium in the SNP-SNP interaction network. Second, this study collected over 20 brain disorder-related gene sets to measure the predictive ability of iGOAT. These genes were shown by different methods to be associated with brain disorders, and their involvement in the causation of brain disorders is not equal. They may however represent a comprehensive dataset of genes that are associated with brain disorders, and which may therefore facilitate the identification of new disease genes from the sets of HRGs predicted by iGOAT. Thirdly, the HRGs predicted by iGOAT are dependent on the power of the original GWAS. It follows that the further development of GWAS studies should dramatically improve the predictive power of iGOAT. Finally, to reduce the computational burden, we only used pairwise SNP-SNP interactions integrating with GO interactions and protein–protein interactions in constructing the Bayesian framework. The pairwise SNP-SNP interactions may miss the information on a more complex network. Nevertheless, the missing information may be compensated for by protein–protein interactions and GO interactions. Moreover, while this study validated the roles of *MHL1* gene in SCZ experimentally, more experiments on other genes will be required in future.

Altogether, we have presented here an approach, iGOAT, which integrates heterogeneous genomic data into a Bayesian framework to allow the prediction of disease-associated genes. The application of iGOAT to both psychiatric disorders and neurological diseases revealed the importance of EPI and SNP-SNP interactions in the predictions. iGOAT can potentially facilitate the development of neurologically relevant hypotheses from GWAS study results. For example, iGOAT can help us to generate the hypothesis that many SNPs associated with the disorders play a regulatory role through remote regulatory elements by identifying the HRGs rather than the nearest gene to the index SNPs. Additionally, iGOAT identified many HRGs that are not reported to be associated with brain disorders in widely used known gene sets. These HGRs can help us to generate the hypothesis that the disorders are involved in a new gene network. iGOAT is also able to be applied to non-brain disorders by taking SNPs associated with a non-brain disorder as input, and utilizing multi-omics data from tissues related to a specific disorder to construct the Bayesian framework.

Contributors

HY.Z designed the study. D.H performed the model construction and analyses, with assistance from X.X, C.F, M.Q, M.M, and J.Y. L.L performed the experiments, with assistance from HS.Z, F.L. D.N.C, M.M. and P.S collected the data and assisted the analysis. S.L, D.H and L.L wrote the manuscript. HY.Z supervised the study. All authors discussed and accessed the results, and verified the underlying data and contributed to the final version of the paper.

Data sharing statement

Details of data resources used in this study were listed in Table S23. The source code and used multi-omics data that support the findings of this study are available in GitHub (https://github.com/Dan-He/iGOAT).

Supplemental information is available on the website.

Declaration of interests

The authors declare no competing interests.

Acknowledgements

The authors thank many resources for making their data available. Data for this study were also obtained from the UK-Biobank (application number #65805 and # 51732). We appreciate Professor Jian Yang from Westlake University in China providing constructive suggestions on this work.

The following grants were awarded to HY.Z. The National Key Research and Development Program of China (2024YFF1204902) and the Natural Science Foundation of China (82371482) was used to support data collection. The Guangzhou Science and Technology Research Plan (2023A03J0659) and Natural Science Foundation of Guangdong (2024A1515011363) was used to support data analysis and study design."

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ebiom.2024.105286.

References

1 Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 2019;47(D1):D1005–D1012.

2 Zhao H, Yang Y, Lu Y, et al. Quantitative mapping of genetic similarity in human heritable diseases by shared mutations. *Hum Mutat.* 2018;39(2):292–301.

3 Zhao H, Nyholt DR. Gene-based analyses reveal novel genetic overlap and allelic heterogeneity across five major psychiatric disorders. *Hum Genet.* 2017;136(2):263–274.

4 Liu JZ, McRae AF, Nyholt DR, et al. A versatile gene-based test for genome-wide association studies. *Am J Hum Genet.* 2010;87(1):139–145.

5 Zhu Z, Zhang F, Hu H, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet.* 2016;48(5):481–487.

6 Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun.* 2017;8(1):1826.

7 Pardinas AF, Holmans P, Pocklington AJ, et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat Genet.* 2018;50(3):381–389.

8 Fan C, Chen K, Zhou J, et al. Systematic analysis to identify transcriptome-wide dysregulation of Alzheimer's disease in genes and isoforms. *Hum Genet.* 2021;140(4):609–623.

9 Zhang F, Chen W, Zhu Z, et al. OSCA: a tool for omic-data-based complex trait analysis. *Genome Biol.* 2019;20(1):107.

10 Wang Q, Chen R, Cheng F, et al. A Bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia GWAS data. *Nat Neurosci.* 2019;22(5):691–699.

11 He D, Fan C, Qi M, Yang Y, Cooper DN, Zhao H. Prioritization of schizophrenia risk genes from GWAS results by integrating multi-omics data. *Transl Psychiatry.* 2021;11(1):175.

12 Schizophrenia Working Group of the Psychiatric Genomics C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature.* 2014;511(7510):421–427.

13 Prabhu S, Pe'er I. Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease. *Genome Res.* 2012;22(11):2230–2240.

14 Jorgenson E, Witte JS. A gene-centric approach to genome-wide association studies. *Nat Rev Genet.* 2006;7(11):885–891.

15 Fang G, Wang W, Paunic V, et al. Discovering genetic interactions bridging pathways in genome-wide association studies. *Nat Commun.* 2019;10(1):4274.

16 Holzinger ER, Verma SS, Moore CB, et al. Discovery and replication of SNP-SNP interactions for quantitative lipid traits in over 60, 000 individuals. *BioData Min.* 2017;10:25.

17 Li P, Guo M, Wang C, Liu X, Zou Q. An overview of SNP interactions in genome-wide association studies. *Brief Funct Genomics.* 2015;14(2):143–155.

18 Lee KY, Leung KS, Ma SL, et al. Genome-wide search for SNP interactions in GWAS data: algorithm, feasibility, replication using schizophrenia datasets. *Front Genet.* 2020;11:1003.

19 Sey NYA, Hu B, Mah W, et al. A computational tool (H-MAGMA) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. *Nat Neurosci.* 2020;23(4):583–593.

20 Wingo TS, Liu Y, Gerasimov ES, et al. Shared mechanisms across the major psychiatric and neurodegenerative diseases. *Nat Commun.* 2022;13(1):4314.

21 Wingo AP, Fan W, Duong DM, et al. Shared proteomic effects of cerebral atherosclerosis and Alzheimer's disease on the human brain. *Nat Neurosci.* 2020;23(6):696–700.

22 Smeland OB, Kutrolli G, Bahrami S, et al. Genome-wide analyses reveal widespread genetic overlap between neurological and psychiatric disorders and a convergence of biological associations related to the brain. *medRxiv.* 2023;21:23292993.

23 Li Y, Jiao J. Deficiency of TRPM2 leads to embryonic neurogenesis defects in hyperthermia. *Sci Adv.* 2020;6(1):eaay6350.

24 Won H, de la Torre-Ubieta L, Stein JL, et al. Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature.* 2016;538(7626):523–527.

25 Waszak SM, Delaneau O, Gschwind AR, et al. Population variation and genetic control of modular chromatin architecture in humans. *Cell.* 2015;162(5):1039–1050.

26 Nott A, Holtman IR, Coufal NG, et al. Brain cell type-specific enhancer-promoter interactome maps and disease-risk association. *Science.* 2019;366(6469):1134–1139.

27 Wang D, Liu S, Warrell J, et al. Comprehensive functional genomic resource and integrative model for the human brain. *Science.* 2018;362(6420).

28 Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature.* 2018;562(7726):203–209.

29 Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81(3):559–575.

30 Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000;25(1):25–29.

31 Oughtred R, Rust J, Chang C, et al. The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.* 2021;30(1):187–200.

32 Basha O, Barshir R, Sharon M, et al. The TissueNet v.2 database: a quantitative view of protein-protein interactions across human tissues. *Nucleic Acids Res.* 2017;45(D1):D427–D431.

33 Ng B, White CC, Klein HU, et al. An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nat Neurosci*. 2017;20(10):1418–1426.
34 Võsa U, Claringbould A, Westra H-J, et al. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv*. 2018:447367.
35 Darmanis S, Sloan SA, Zhang Y, et al. A survey of human brain transcriptome diversity at the single cell level. *Proc Natl Acad Sci U S A*. 2015;112(23):7285–7290.
36 Lake BB, Ai R, Kaeser GE, et al. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science*. 2016;352(6293):1586–1590.
37 Lake BB, Chen S, Sos BC, et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol*. 2018;36(1):70–80.
38 Wang JH, Zhao LF, Wang HF, et al. GenCLiP 3: mining human genes' functions and regulatory networks from PubMed based on co-occurrences and natural language processing. *Bioinformatics*. 2019; btz807.
39 Pinero J, Bravo A, Queralt-Rosinach N, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res*. 2017;45(D1):D833–D839.
40 Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*. 2017;45(D1):D353–D361.
41 Deciphering Developmental Disorders S. Prevalence and architecture of de novo mutations in developmental disorders. *Nature*. 2017;542(7642):433–438.
42 Satterstrom FK, Kosmicki JA, Wang J, et al. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell*. 2020;180(3):568–584.e23.
43 Singh T, Kurki MI, Curtis D, et al. Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. *Nat Neurosci*. 2016;19(4):571–577.
44 Stenson PD, Ball EV, Mort M, et al. Human gene mutation database (HGMD): 2003 update. *Hum Mutat*. 2003;21(6):577–581.
45 Kang HJ, Kawasawa YI, Cheng F, et al. Spatio-temporal transcriptome of the human brain. *Nature*. 2011;478(7370):483–489.
46 Sedmak G, Jovanov-Milosevic N, Puskarjov M, et al. Developmental expression patterns of KCC2 and functionally associated molecules in the human brain. *Cereb Cortex*. 2016;26(12):4574–4589.
47 Willsey AJ, Sanders SJ, Li M, et al. Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell*. 2013;155(5):997–1007.
48 Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinf*. 2008;9:559.
49 Gandal MJ, Zhang P, Hadjimichael E, et al. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science*. 2018;362(6420).
50 Huckins LM, Dobbyn A, Ruderfer DM, et al. Gene expression imputation across multiple brain regions provides insights into schizophrenia risk. *Nat Genet*. 2019;51(4):659–674.
51 Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*. 2019;47(D1):D607–D613.
52 Shen S, Chen X, Li H, Sun L, Yuan Y. MLH1 promoter methylation and prediction/prognosis of gastric cancer: a systematic review and meta and bioinformatic analysis. *J Cancer*. 2018;9(11):1932–1942.
53 Mao G, Lee S, Ortega J, Gu L, Li GM. Modulation of microRNA processing by mismatch repair protein MutLalpha. *Cell Res*. 2012;22(6):973–985.
54 Tomita K, Kubo K, Ishii K, Nakajima K. Disrupted-in-Schizophrenia-1 (Disc1) is necessary for migration of the pyramidal neurons during mouse hippocampal development. *Hum Mol Genet*. 2011;20(14):2834–2845.
55 Mao Y, Ge X, Frank CL, et al. Disrupted in schizophrenia 1 regulates neuronal progenitor proliferation via modulation of GSK3beta/beta-catenin signaling. *Cell*. 2009;136(6):1017–1031.
56 Ishizuka K, Kamiya A, Oh EC, et al. DISC1-dependent switch from progenitor proliferation to migration in the developing cortex. *Nature*. 2011;473(7345):92–96.
57 Duan X, Chang JH, Ge S, et al. Disrupted-In-Schizophrenia 1 regulates integration of newly generated neurons in the adult brain. *Cell*. 2007;130(6):1146–1158.
58 Rubio-Perez JM, Morillas-Ruiz JM. A review: inflammatory process in Alzheimer's disease, role of cytokines. *Sci World J*. 2012;2012: 756357.
59 Kinney JW, Bemiller SM, Murtishaw AS, Leisgang AM, Salazar AM, Lamb BT. Inflammation as a central mechanism in Alzheimer's disease. *Alzheimers Dement (N Y)*. 2018;4:575–590.
60 Wu KM, Zhang YR, Huang YY, Dong Q, Tan L, Yu JT. The role of the immune system in Alzheimer's disease. *Ageing Res Rev*. 2021;70:101409.
61 Lopategui Cabezas I, Herrera Batista A, Penton Rol G. The role of glial cells in Alzheimer disease: potential therapeutic implications. *Neurologia*. 2014;29(5):305–309.
62 McNaught KS, Olanow CW. Proteolytic stress: a unifying concept for the etiopathogenesis of Parkinson's disease. *Ann Neurol*. 2003;53(Suppl 3):S73–S84.
63 Pickard B. Progress in defining the biological causes of schizophrenia. *Expert Rev Mol Med*. 2011;13:e25.
64 Kesby JP, Eyles DW, McGrath JJ, Scott JG. Dopamine, psychosis and schizophrenia: the widening gap between basic and clinical neuroscience. *Transl Psychiatry*. 2018;8(1):30.
65 Scaini G, Rezin GT, Carvalho AF, Streck EL, Berk M, Quevedo J. Mitochondrial dysfunction in bipolar disorder: evidence, pathophysiology and translational implications. *Neurosci Biobehav Rev*. 2016;68:694–713.
66 Dager SR, Friedman SD, Parow A, et al. Brain metabolic alterations in medication-free patients with BipolarDisorder. *Arch Gen Psychiatr*. 2004;61(5):450–458.
67 Magalhaes PV, Jansen K, Pinheiro RT, et al. Peripheral oxidative damage in early-stage mood disorders: a nested population-based case-control study. *Int J Neuropsychopharmacol*. 2012;15(8): 1043–1050.
68 Andreazza AC, Kapczinski F, Kauer-Sant'Anna M, et al. 3-Nitrotyrosine and glutathione antioxidant system in patients in the early and late stages of bipolar disorder. *J Psychiatry Neurosci*. 2009;34(4):263–271.
69 Pfaffenseller B, Fries GR, Wollenhaupt-Aguiar B, et al. Neurotrophins, inflammation and oxidative stress as illness activity biomarkers in bipolar disorder. *Expert Rev Neurother*. 2013;13(7):827–842.
70 Pfaffenseller B, Wollenhaupt-Aguiar B, Fries GR, et al. Impaired endoplasmic reticulum stress response in bipolar disorder: cellular evidence of illness progression. *Int J Neuropsychopharmacol*. 2014;17(9):1453–1463.
71 Levitt P, Veenstra-VanderWeele J. Neurodevelopment and the origins of brain disorders. *Neuropsychopharmacology*. 2015;40(1):1–3.
72 Weinberger DR. The neurodevelopmental origins of schizophrenia in the penumbra of genomic medicine. *World Psychiatr*. 2017;16(3):225–226.
73 Barlow BK, Richfield EK, Cory-Slechta DA, Thiruchelvam M. A fetal risk factor for Parkinson's disease. *Dev Neurosci*. 2004;26(1):11–23.
74 Faa G, Marcialis MA, Ravarino A, Piras M, Pintus MC, Fanos V. Fetal programming of the human brain: is there a link with insurgence of neurodegenerative disorders in adulthood? *Curr Med Chem*. 2014;21(33):3854–3876.
75 Samii A, Nutt JG, Ransom BR. Parkinson's disease. *Lancet*. 2004;363(9423):1783–1793.
76 Postuma RB, Berg D. Prodromal Parkinson's disease: the decade past, the decade to come. *Mov Disord*. 2019;34(5):665–675.
77 Lanctot KL, Amatniek J, Ancoli-Israel S, et al. Neuropsychiatric signs and symptoms of Alzheimer's disease: new treatment paradigms. *Alzheimers Dement (N Y)*. 2017;3(3):440–449.
78 Huang J, Zuber V, Matthews PM, Elliott P, Tzoulaki J, Dehghan A. Sleep, major depressive disorder, and Alzheimer disease: a Mendelian randomization study. *Neurology*. 2020;95(14):e1963–e1970.
79 Zoghbi AW, Dhindsa RS, Goldberg TE, et al. High-impact rare genetic variants in severe schizophrenia. *Proc Natl Acad Sci U S A*. 2021;118(51).
80 Singh T, Poterba T, Curtis D, et al. Rare coding variants in ten genes confer substantial risk for schizophrenia. *Nature*. 2022;604(7906):509–516.
81 Genovese G, Fromer M, Stahl EA, et al. Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat Neurosci*. 2016;19(11):1433–1441.
82 Nguyen HT, Bryois J, Kim A, et al. Integrated Bayesian analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders. *Genome Med*. 2017;9(1):114.

83 Torjesen I. Depression and SSRI use in pregnancy associated with traits of autism in children. *BMJ*. 2014;349:g4835.

84 Hudson CC, Hall L, Harkness KL. Prevalence of depressive disorders in individuals with autism spectrum disorder: a meta-analysis. *J Abnorm Child Psychol*. 2019;47(1):165–175.

85 Grove J, Ripke S, Als TD, et al. Identification of common genetic risk variants for autism spectrum disorder. *Nat Genet*. 2019;51(3):431–444.

86 Thapar A, Rutter M. Genetic advances in autism. *J Autism Dev Disord*. 2021;51(12):4321–4332.

87 Kapornai K, Gentzler AL, Tepper P, et al. Early developmental characteristics and features of major depressive disorder among child psychiatric patients in Hungary. *J Affect Disord*. 2007;100(1-3):91–101.

88 Chorbadjian TN, Deavenport-Saman A, Higgins C, et al. Maternal depressive symptoms and developmental delay at age 2: a diverse population-based longitudinal study. *Matern Child Health J*. 2020;24(10):1267–1277.

89 Blackburn PR, Schultz MJ, Lahner CA, et al. Expanding the clinical and phenotypic heterogeneity associated with biallelic variants in ACO2. *Ann Clin Transl Neurol*. 2020;7(6):1013–1028.

90 Olsen L, Sparso T, Weinsheimer SM, et al. Prevalence of rearrangements in the 22q11.2 region and population-based risk of neuropsychiatric and developmental disorders in a Danish population: a case-cohort study. *Lancet Psychiatr*. 2018;5(7): 573–580.