MCGA: a multi-strategy conditional gene-based association framework integrating with isoform-level expression profiles reveals new susceptible and druggable candidate genes of schizophrenia

Xiangyi Li[1,3,4,#], Lin Jiang[2,#], Chao Xue[1,3,4], Mulin Jun Li[5], Miaoxin Li[1,3,4,*]

[1]Zhongshan School of Medicine, The Fifth Affiliated Hospital, Sun Yat-sen University, Guangzhou, China; [2]Research Center of Medical Sciences, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China; [3]Key Laboratory of Tropical Disease Control (Sun Yat-sen University), Ministry of Education; [4]Center for Precision Medicine, Sun Yat-sen University, Guangzhou, China; [5]The Province and Ministry Co-sponsored Collaborative Innovation Center for Medical Epigenetics, Tianjin Medical University, Tianjin 300070, China.

* To whom correspondence should be addressed:

Miaoxin Li. Tel: +86-2087335080; Email: limiaoxin@mail.sysu.edu.cn; Add.: Medical Science and Technology Building, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510080, China.

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Abstract

Linkage disequilibrium and disease-associated variants in non-coding regions make it difficult to distinguish truly associated genes from redundantly associated genes for complex diseases. In this study, we proposed a new conditional gene-based framework called MCGA that leveraged an improved effective chi-squared statistic to control the type I error rates and remove the redundant associations. MCGA initially integrated two conventional strategies to map genetic variants to genes, i.e., mapping a variant to its physically nearby gene and mapping a variant to a gene if the variant is a gene-level expression quantitative trait locus (eQTL) of the gene. We further performed a simulation study and demonstrated that the isoform-level eQTL was more powerful

than the gene-level eQTL in the association analysis. Then the third strategy, i.e., mapping a variant to a gene if the variant is an isoform-level eQTL of the gene, was also integrated with MCGA. We applied MCGA to predict the potential susceptibility genes of schizophrenia and found that the potential susceptibility genes identified by MCGA were enriched with many neuronal or synaptic signaling-related terms in the Gene Ontology knowledgebase and antipsychotics-gene interaction terms in the drug-gene interaction database (DGIdb). More importantly, nine susceptibility genes were the target genes of multiple approved antipsychotics in DrugBank. Comparing the susceptibility genes identified by the above three strategies implied that strategy based on isoform-level eQTL could be an important supplement for the other two strategies and help predict more candidate susceptibility isoforms and genes for complex diseases in a multi-tissue context.

Introduction

Genome-wide association studies (GWASs) have been used to identify novel genotype-phenotype associations for more than a decade, and thousands of single-nucleotide polymorphisms (SNPs) have been revealed for their associations with hundreds if not thousands of complex human diseases[1,2]. Nevertheless, conventional GWAS analyses have limited power to produce a complete set of susceptibility variants of complex diseases[3]. Because most susceptibility SNPs only have small effects on a complex phenotype, conventional SNP-based association tests are generally underpowered to reveal susceptibility variants after multiple testing corrections. Moreover, the susceptibility variants scattering randomly throughout the genome are often in strong linkage

disequilibrium (LD) with numerous neutral SNPs, and makes the discrimination of truly causal variants from GWAS hits quite difficult[3]. Finally, more than 90% of the disease-associated variants are in non-coding regions of the genome, and many of them are far from the nearest known gene, and it remains a challenge to link genes and a complex phenotype through the non-coding variants [4,5]. Accordingly, corresponding methodological strategies have been proposed to make up, at least in part, for the issues mentioned above.

First, gene-based approaches can reduce the multiple testing burden by considering the association between a phenotype and all variants within a gene[6]. Assigning a variant to a gene according to the physical position of the variant from gene boundary is one of the most popular strategies for gene-based approaches. For example, MAGMA (Multi-marker Analysis of GenoMic Annotation), one of the most popular gene-based approaches, uses a multiple regression approach to incorporate LD between markers and detect multi-markers effects to perform gene-based analysis[7]. VEGAS, a versatile gene-based test for GWAS, incorporates information from a full set of markers (or a defined subset) within a gene and accounts for LD between markers by simulations from the multivariate normal distribution[8]. GATES, a rapid gene-based association test that uses an extended Simes procedure to assess the statistical significance of gene-level associations [9].

Second, evaluating associations at one gene conditioning on other genes can isolate true susceptibility genes from non-susceptibility genes[10]. Yang et al. proposed an approximate conditional and joint association analysis based on linear regression analysis for estimating the individual causal variant with GWAS summary statistics[11]. Our previously proposed conditional

gene-based association approach based on effective chi-squared statistics (ECS) could remove redundantly associated genes based on the GWAS p-values of variants. The comparison of ECS with MAGMA and VEGAS suggested that ECS might be more powerful to predict biologically sensible susceptibility genes [10].

Third, the observation that variants in the non-coding regions were enriched in the transcriptional regulatory regions implied that these variants might affect the disease risk by altering the genetic regulation of target genes[2]. Integration of expression quantitative trait loci (eQTL) studies and GWAS has been used to investigate the genetic regulatory effects on complex diseases. As many complex diseases manifested themselves in certain tissues, using the eQTLs of potentially phenotype-associated tissues might help identify the true susceptibility genes in tissue context[12]. Based on MAGMA, a method called eMAGMA, which integrated genetic and transcriptomic information (e.g., eQTLs) in a tissue-specific analysis to identify risk genes, was proposed to identify novel genes underlying major depression disorder[13]. S-PrediXcan was developed for imputing the genetically regulated gene expression component based on GWAS summary statistics and transcriptome prediction models built from the eQTL/sQTL dataset of the Genotype Tissue Expression (GTEx) project[14]. Researchers have applied S-PrediXcan to study genetic mechanisms of multiple complex traits [15-17]. In contrast to the considerable research focusing on integrating gene-level eQTLs with GWAS summary statistics, little attention has been paid to integrating isoform-level expression QTLs (isoform-level eQTLs) with GWAS summary statistics. Michael J. Ganda et al. estimated the candidate risk genes of three psychiatric disorders based on GWAS summary statistics and isoform-level expression profiles. They emphasized the importance

of isoform-level gene regulatory mechanisms in defining cell type and disease specificity[18], and similar analyses and conclusions were generated for Alzheimer's disease[19].

Though much achievement has been attained, identifying independently phenotype-associated genes with high reliability remains challenging, especially for complex diseases. In the present study, we aimed to build a more powerful conditional gene-based framework based on a new ECS and investigate whether QTLs in phenotype-associated tissues, especially the isoform-level eQTLs, can predict more susceptibility genes or not. The main assumption of our study is that isoform-level eQTLs may reflect the more real variant-expression regulatory relationship than gene-level eQTLs, and using isoform-level eQTLs can help predict novel susceptibility genes and isoforms that cannot be found by the conventional gene-based approaches and gene-level eQTLs strategy. The formation procedure of the assumption is this: gene-level eQTLs are predicted based on gene-level expression profiles and corresponding genotype data. In contrast, isoform-level eQTLs are predicted based on isoform-level (or transcript-level) expression profiles and corresponding genotype data. The gene-level expression profiles are computed by averaging the expression of multiple isoforms belonging to the gene, which may omit the expression heterogeneity among these isoforms and neutralize the opposite effects and lower the power of gene-level eQTLs. Taken together, conventional gene-based approaches mainly focus on variants close to genes boundary (say +/-5kilo base pairs), thus omit remote but important association relationship between genes and variants. In the present study, we expanded the application scope of conventional gene-based approaches by using gene-level eQTLs and isoform-level eQTLs in the potentially phenotype-associated tissues to identify more candidate susceptibility genes and

115 isoforms.

116

117 ## Results

118 ### Overview of the Multi-strategy Conditional Gene-based Association (MCGA) framework

119 Our previously powerful and unified framework, DESE, proposed to estimate the potentially
120 phenotype-associated tissues of complex diseases, could iteratively run the conditional gene-based
121 association analysis with the selective expression score of genes among multiple tissues[20]. We had
122 demonstrated in DESE that the iterative operations with the selective expression analysis based on
123 expression profiles could strengthen the power of conditional gene-based analysis. In this study,
124 we proposed a new conditional gene-based framework, MCGA, which could also iteratively run
125 conditional gene-based association analysis with selective expression analysis, to systematically
126 explore the susceptibility genes associated with a complex phenotype using GWAS summary
127 statistics and eQTL summary statistics of SNPs. MCGA has two main advantages over DESE.
128 First, MCGA is based on a new effective chi-squared statistic (ECS), with which the type I error
129 could be controlled within a proper level. Second, MCGA can perform conditional gene-based
130 association analysis using different SNPs sets, i.e., physically nearby SNPs, gene-level eQTLs and
131 isoform-level eQTLs.

132

133 To evaluate the performance of MCGA, we performed extensive simulations and a real data
134 application to schizophrenia. Specifically, we organized the present study in four sequential parts
135 that cover the optimizing the exponent of chi-squared statistics to control type I error rates,

136 applying the improved chi-squared statistics to perform conditional gene-based association
137 analysis in simulation data, applying the improved conditional gene-based association analysis to
138 predict the potential susceptibility genes of schizophrenia, and finally extending the application by
139 using gene-level eQTLs and isoform-level eQTLs. Three strategies were implemented in three
140 conditional gene-based models, respectively. The model assigning a SNP to a gene according to
141 the physical distance of the SNP from the gene boundary is named MCGA_Dist. The model using
142 gene-level eQTLs is named MCGA_eQTL. The model using isoform-level eQTLs is named
143 MCGA_isoQTL (**Figure 1**). All three models of MCGA have been implemented in our integrative
144 platform KGGSEE.

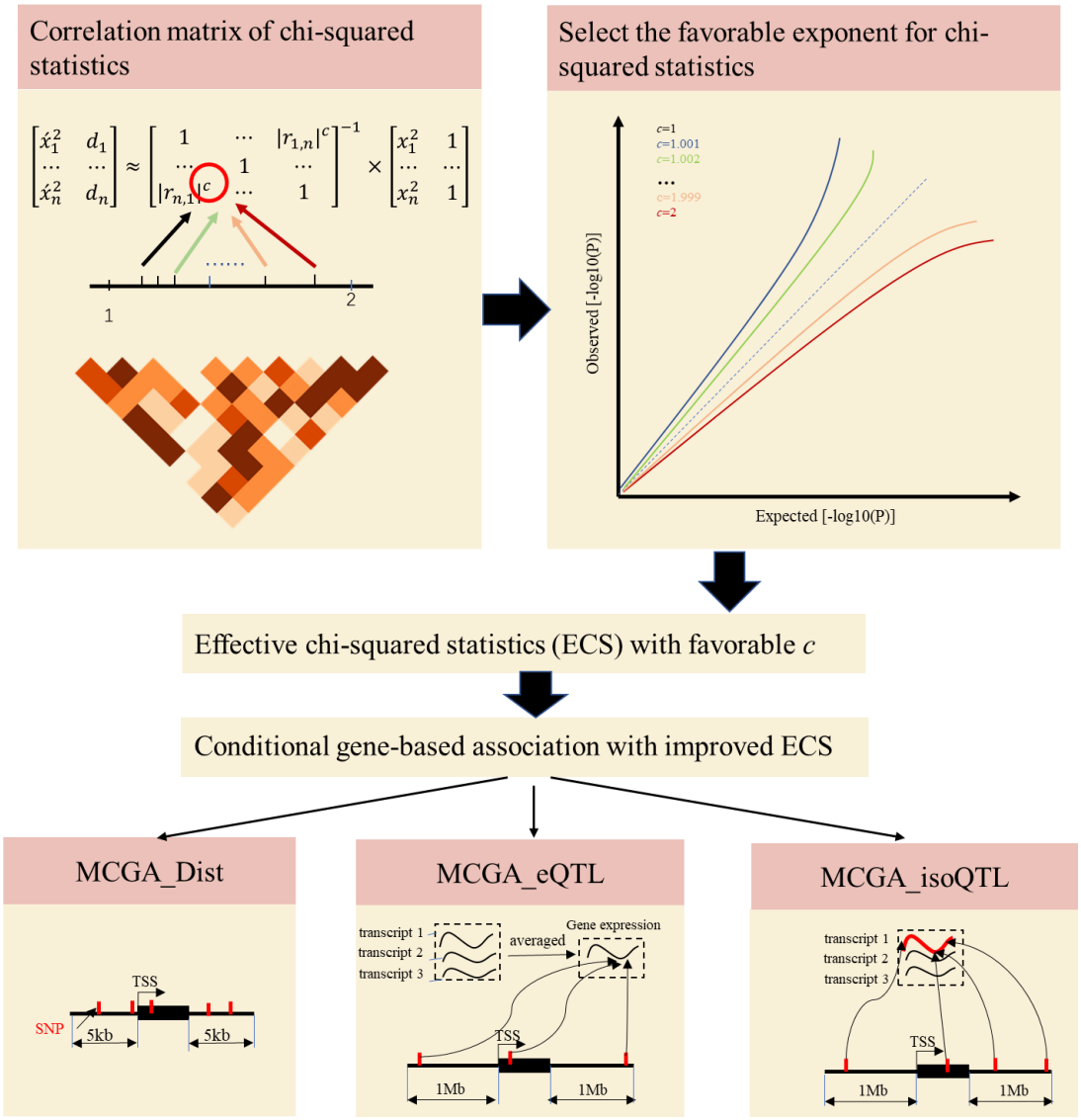


Figure 1: The simplified principle of the present study. First, choose the best exponent *c* between 1 and 2. Each time we increased *c* by an interval of 0.01. The best *c* can control the type I error within a proper level. Then we implemented the best *c* to improve the previous ECS approach and got the improved conditional gene-based approach. Third, the mapping strategies used by three conditional gene-based association models of MCGA. 5kb: 5000 base pairs. 1Mb: 10^6 base pairs. TSS: Transcription Start Site.

Our simulation results showed that the type I error rate was controlled within a reasonable level by using the new effective chi-squared statistics (ECS) with the favorable exponent. Another simulation study pointed out that association analysis based on isoform-level eQTLs was more powerful than gene-level eQTLs. As for predicting the potential susceptibility genes of

schizophrenia, MCGA_Dist, MCGA_eQTL and MCGA_isoQTL all produced a set of potentially susceptible and druggable genes. Moreover, with the help of MCGA_isoQTL, we also predicted the potential susceptibility isoforms (or transcripts) for schizophrenia. Our results also showed that the usage of isoform-level eQTLs could predict some important susceptible and druggable genes which cannot be found by MCGA_Dist and MCGA_eQTL.

The favorable exponent *c* in the correlation matrix of chi-squared statistics to control the type I error rates

We found that the exponent *c* in the correlation matrix of chi-squared statistics could determine the divergence of the *p*-values of the ECS tests from the uniform distribution (see details in **Formula (1)** of Methods). Small *c* led to a p-value larger than expected, indicating the deflated type I error rate. Large *c* led to a p-value smaller than expected, indicating an inflated type I error rate. As shown in **Figure 2**, the *c*=1.0 led to deflated *p*-values while the *c*=2.0 led to inflated *p*-values in the upper tail of the Q-Q plot against the uniform distribution. This pattern was independent of sample sizes, variant sizes and phenotype distribution (binary or continuous) (**Figure 2**). The stable trend determined by the *c* value also implied that the favorable *c*, which could properly control the type I error rate, measured by the minimal mean log fold change (MLFC), must be within sections 1 and 2. Our theoretical derivation also demonstrated *c* value should be within section 1 and 2 (see details in the **Materials and Methods**). Moreover, it seemed given the *c* value, the distributions of *p*-values were similar at different sample sizes and phenotype distributions. As shown in the Q-Q plot (**Figure 3**), most majority *p*-values at the sample size 10,000 and 40,000 of

binary or continuous phenotypes were overlapped. **Figure 3** showed the favorable *c* obtained by the grid search algorithm at 84 different scenarios. Again, the favorable *c* values were approximately independent of trait types, sample sizes and variant sizes. For the sake of simplicity, we proposed to use the averaged favorable *c* values, 1.432, for all the analyses in the present study.

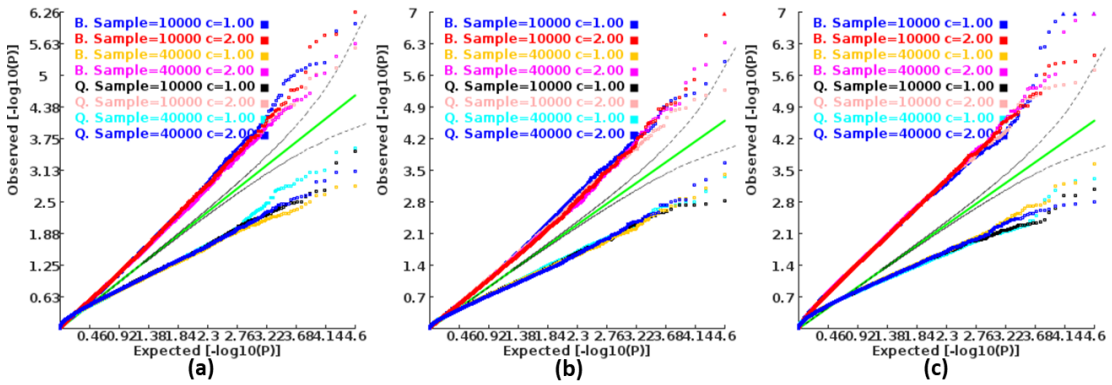


Figure 2: **Q-Q plots of the *p*-value of the ECS test under null hypothesis based on two extreme exponents (i.e., 1 and 2). a)**, **b)**, and **c)** represent the variant size of 50, 100 and 500, respectively.

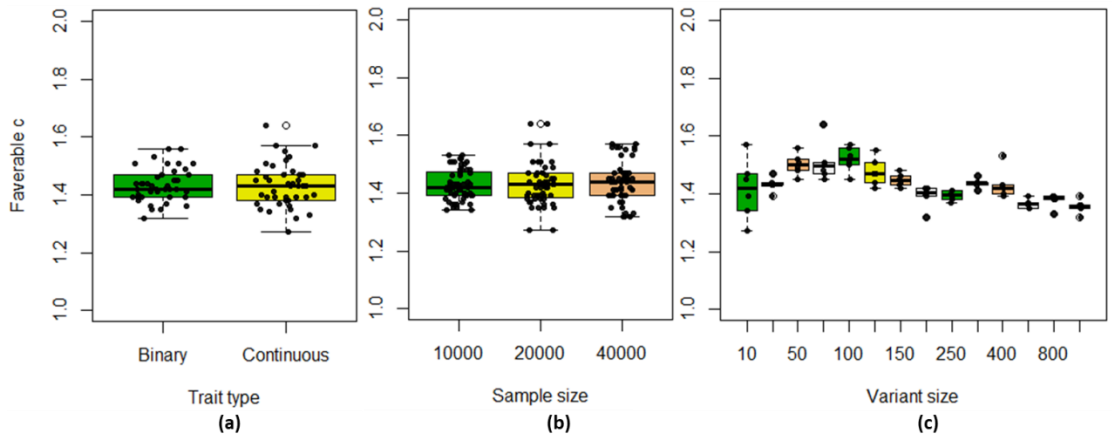


Figure 3: **The boxplot of favorable *c* values at different simulation scenarios**. **a)** at binary and continuous phenotypes; **b)** at different sample sizes; **c)** at different variant sizes.

The type I error and power of the conditional gene-based association analysis based on the new chi-squared statistics

Further, we investigated the type I error and power of the conditional gene-based association analysis based on the improved ECS with the above favorable exponent *c* value (i.e., 1.432). The basic function of the conditional gene-based test based on the improved ECS is to remove redundant associations among associated genes. As shown in **Figure 4**, in six different scenarios (see details in **Materials and Methods**), the conditional *p*-values of the genes without truly casual loci approximately followed the uniform distribution $U[0,1]$ regardless of the variance explained by its nearby genes. The distribution of conditional *p*-value was similar to that produced by the conventional likelihood ratio test for nested linear regression models. These results suggested that the conditional gene-based association analysis based on the improved ECS could produce valid *p*-values for statistical inference. In contrast, the unconditional association test produced an inflated *p*-value due to the indirect associations produced by nearby causal genes in the LD block. Concerning the statistical power, we found conditional gene-based association analysis based on the improved ECS produced smaller *p*-values than the likelihood ratio test (**Figure 5**), suggesting a higher statistical power of the former. Another advantage of conditional gene-based association analysis based on the improved ECS over the likelihood ratio test was that the former did not require individual genotypes. The reason might be that the degree of freedom in the latter was inflated by the LD among variants. Hereinafter we named the conditional gene-based association analysis based on the improved ECS with favorable exponent *c* value as MCGA.

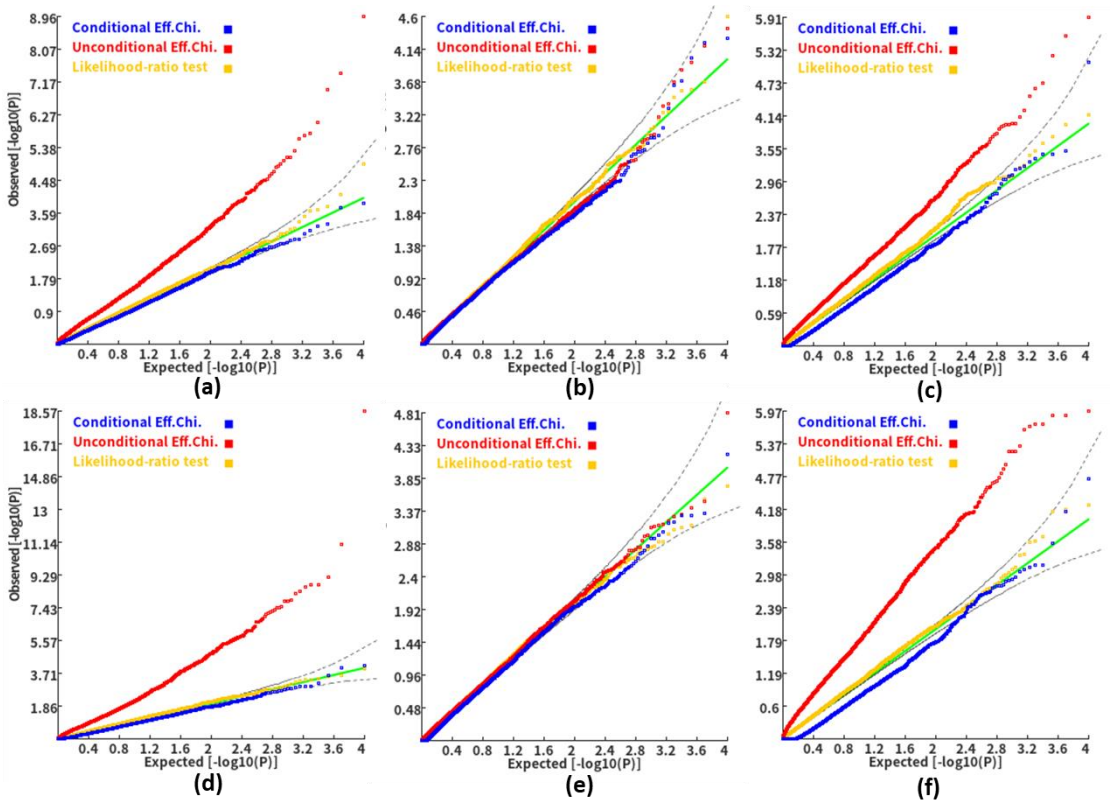


Figure 4: **Q-Q plot of the conditional, unconditional gene-based association test and likelihood-ratio test under the null hypothesis. a)** and **d)**: two gene-variant pairs with similar variant sizes (SIPA1L2 with 29 variants and LOC729336 with 30 variants). **b)** and **e)**: two gene-variant pairs with different variant sizes, and the first is larger than the second (CACHD1 with 41 variants and RAVER2 with 8 variants). **c)** and **f)**: two gene-variant pairs with different variant sizes, and the second is larger than the first (LOC647132 with 5 variants and FAM5C with 48 variants). **a)**, **b)** and **c)**: the former gene has no QTL, and QTL in the latter gene explained 0.5% of heritability. **d)**, **e)** and **f)**: the former gene has no QTL, and QTL in the latter gene explained 1% of heritability. Ten thousand phenotype datasets were simulated for each scenario. Unconditional Eff. Chi. (the red) represents unconditional association analysis at the former gene by the improved ECS. Conditional Eff. Chi (the blue) represents conditional association analysis at the former gene conditioning on the latter gene by the improved ECS. The likelihood ratio test (the yellow) was conducted based on nested linear regression models.

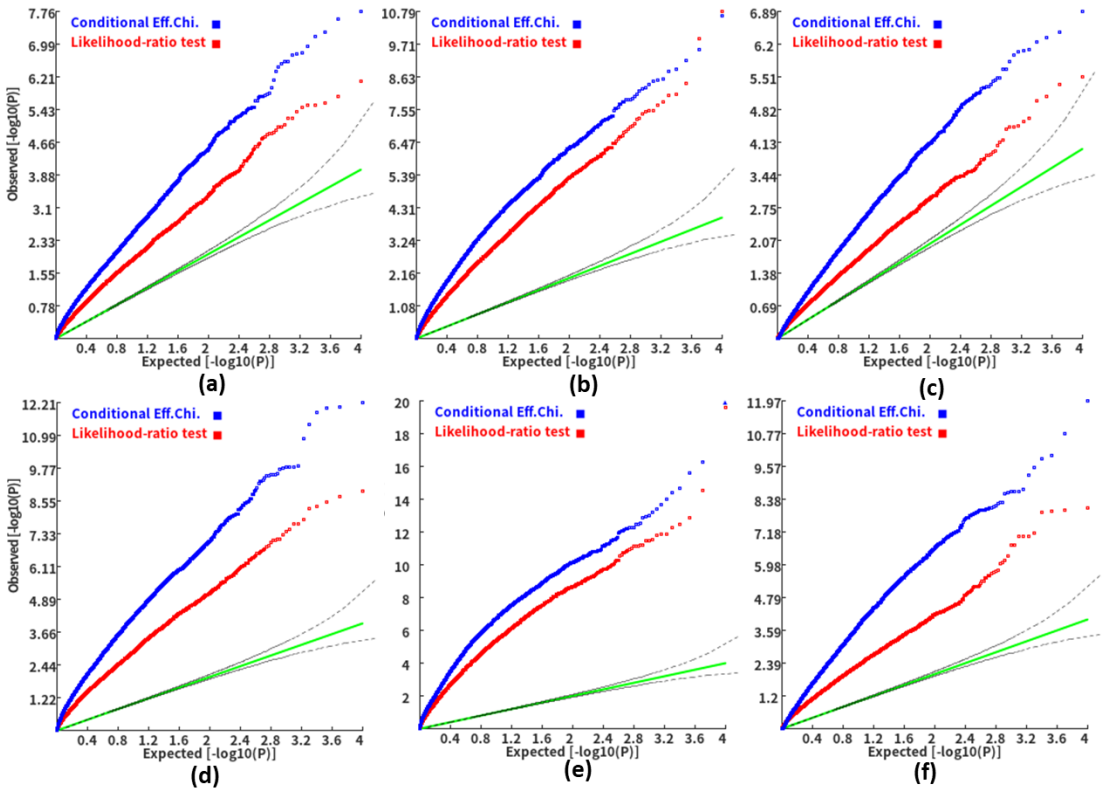


Figure 5: Q-Q plot of the conditional gene-based association test and likelihood-ratio test at different representative gene-variant pairs. a) and **d)**: two gene-variant pairs with similar variant sizes (SIPA1L2 with 29 variants and LOC729336 with 30 variants). **b)** and **e)**: two gene-variant pairs with different variant sizes, and the first is larger than the second (CACHD1 with 41 variants and RAVER2 with 8 variants). **c)** and **f)**: two gene-variant pairs with different variant sizes, and the second is larger than the first (LOC647132 with 5 variants and FAM5C with 48 variants). **a)-c)**: the QTL in either gene explained 0.25% of heritability. **d)-f)**: the QTL in either gene explained 0.5% of heritability. 1000 phenotype datasets were simulated for each scenario. Conditional Eff. Chi.: conditional association analysis at the former gene conditioning on the latter gene by the improved ECS. Likelihood Ratio Test: likelihood ratio test in which the full model included QTLs of both genes and the nested model included QTL of the latter gene.

Application of MCGA to predict the potential susceptibility genes for schizophrenia (MCGA_Dist)

In the above simulation study, we demonstrated that the conditional gene-based analysis based on the improved ECS was more powerful than the likelihood ratio test in each simulation scenario. Here to further evaluate the performance of MCGA in the real-world data, we used a recent

large-scale GWAS summary statistics dataset[20] and gene expression profiles (GTEx v8) of ~ 50 human tissues[21] to identify the susceptibility genes of schizophrenia. Here, MCGA_Dist was first used, i.e., the variants were assigned to genes if the variants were in the window of +/-5kb around the gene boundary (see details in **Materials and Methods**). We found that 221 of 34,159 genes identified by MCGA_Dist had significantly statistical *p*-values smaller than 2.5E-6 (see details in **Table S1**).

To further study the functional annotations of the 221 potential susceptibility genes, we performed Gene Ontology (GO) enrichment analysis. Interestingly, we found that most GO:BP and GO:CC enrichment terms were neuronal-, dendrite- or synaptic signaling-related terms. Besides, the GO:MF enrichment terms were all about cellular signaling transduction (see examples in **Figure 6** and see details in **Table S2**). Systematic text-mining method was used to search the PubMed database to find papers that had reported the potential susceptibility genes of schizophrenia. The results showed that 87 of the 221 (~ 39.4%) potential susceptibility genes had at least one search hit (see details in **Table S3**). The GO enrichment analysis and the text-mining results both implied the utility of MCGA_Dist.

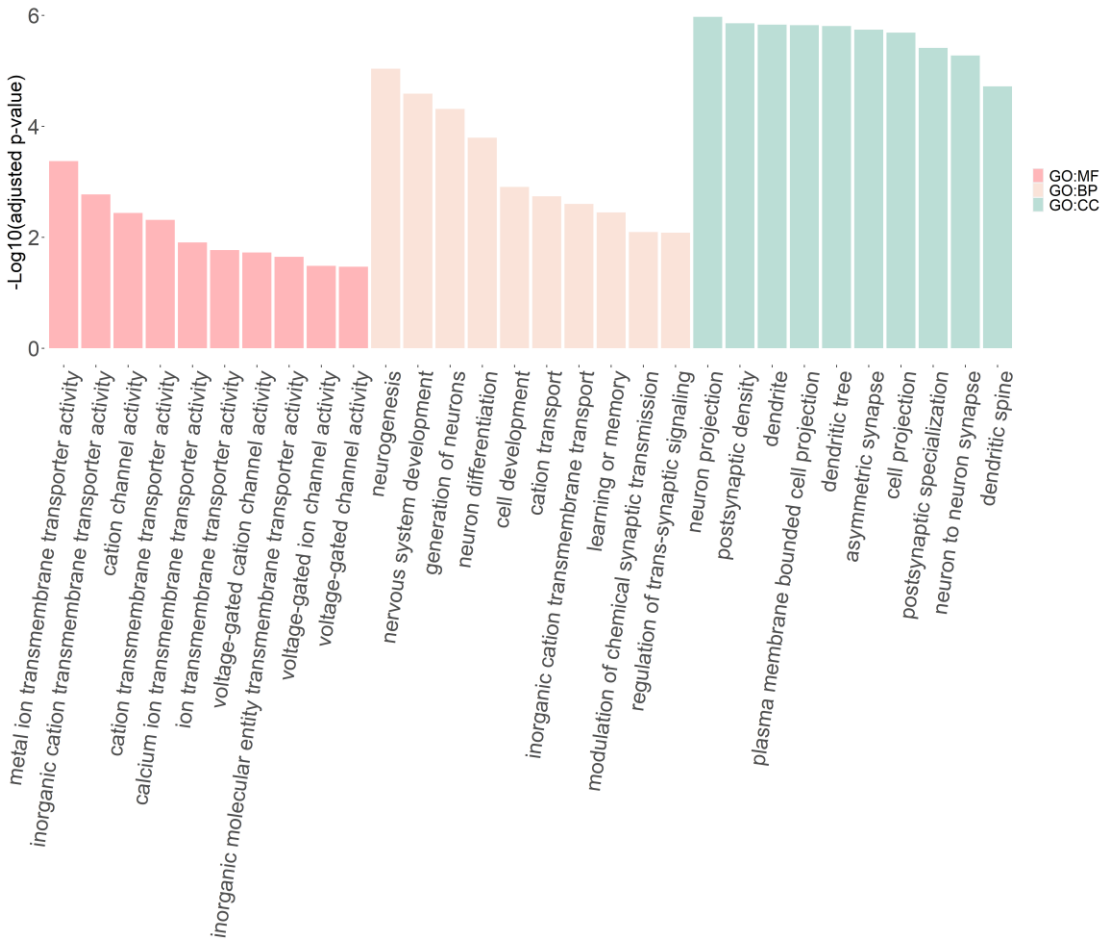


Figure 6: GO Functional annotations of the potential susceptibility genes of schizophrenia identified by MCGA_Dist. MF: Molecular Function of GO. BP: Biological Process terms of GO. CC: Cellular Component terms of GO. The X-axis represents the top 10 significant GO enrichment terms (in MF, BP and CC). The Y-axis represents the negative log10 of the adjusted *p*-value.

Evaluate the power of gene-level eQTLs and isoform-level eQTLs in association analysis by the simulation study

A common assumption is that genes close to significant variants are more likely to be the susceptibility genes, but the reality is that some potentially associated genes are not closest to the significant variants[16]. Molecular Quantitative Trait Loci (molQTL) is a genetic variant associated with a molecular trait, such as a gene-level eQTL and isoform-level eQTL, and can associate a

variant with a gene or isoform. MCGA_Dist mapped a variant to a gene if the variant was in the +/-5kb window around the gene boundary. Next, we assigned a variant to a gene (or isoform) if the variant is a gene-level or isoform-level eQTL to broaden the application of MCGA. Since the isoform-level and corresponding gene-level expression profiles were quantified based on the same RNA-sequencing data, we wanted to test whether the power of association analysis based on the gene-level eQTLs was higher than that based on isoform-level eQTLs or not. We first performed a simulation study to evaluate the power of association analysis based on gene-level eQTLs and isoform-level eQTLs, respectively.

We considered the simplest case for simplicity, i.e., variants affected phenotype only through regulating the gene expression. We simulated genotype data, isoform-level expression profiles and corresponding phenotype data (see details in **Materials and Methods**). Specifically, we simulated four scenarios, i.e., association analysis using all variants (phenotype-associated isoform-level eQTLs and the other isoform-level eQTLs, denoted as Allvar in **Table 1**), association analysis only using phenotype-associated isoform-level eQTLs (denoted as isoform eQTL in **Table 1**), association analysis using gene-level eQTLs which were computed by the gene expression profiles derived by the average value of multiple isoforms belonging to the gene. As for scenarios of genes with multiple isoforms, we specifically simulated two new scenarios (denoted as eQTL_3isoform and eQTL_6isoform in **Table 1**), i.e., a gene with three (eQTL_3isoform) and six different isoforms (eQTL_6isoform). The expression value of the gene with three isoforms was averaged by the following three isoforms, i.e., one isoform associated with phenotype and the other two random isoforms simulated by the standard normal distribution $N(0,1)$. The expression value of

the gene with six isoforms was averaged by the following six isoforms, i.e., one isoform associated with phenotype and the other five random isoforms simulated by the standard normal distribution $N(0,1)$ (see details in **Materials and Methods**). Based on the four scenarios mentioned above, we used six different parameter combinations to simulate six different cases, and each parameter combination was simulated 100 times to evaluate the statistical power. As shown in **Table 1**, the power of the association test based on phenotype-associated isoform-level eQTLs was the highest of all cases. The simulation results implied that isoform-level eQTLs were more powerful than gene-level eQTLs in association analysis.

Table 1: Different simulation scenarios and corresponding power in association analysis.

Parameter combination		Power			
Ev[a]	**Eg**	**AllVar**	**isoform eQTL**	**eQTL_3isoform**	**eQTL_6isoform**
0	0.1	0	**0.01**	0	0.01
0.1	0.1	0.03	**0.05**	0.02	0.01
0.15	0.1	0.03	**0.25**	0.04	0.02
0.05	0.2	0.26	**0.39**	0.06	0.01
0.1	0.2	0.13	**0.22**	0.02	0.03
0.15	0.2	0.68	**0.96**	0.45	0.11

[a] **Ev** denotes the effect size of independent variants on gene expression. **Eg** denotes the effect size of gene expression on phenotype. More details can be seen in Methods.

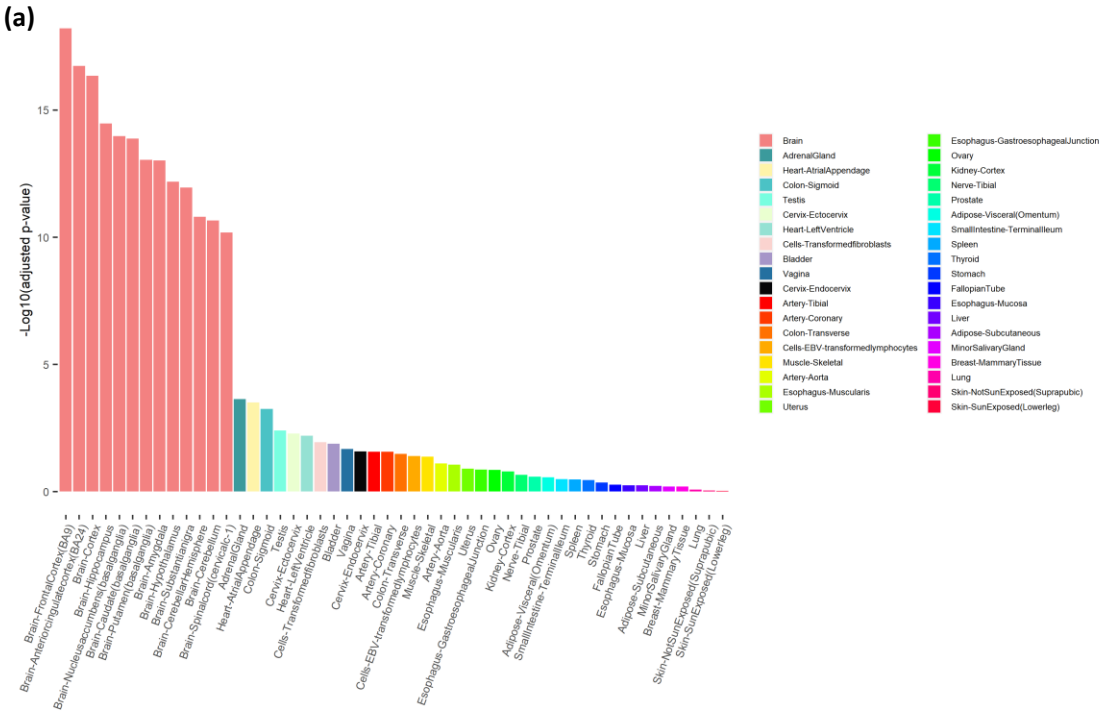
Broaden the application of MCGA by using gene-level eQTLs and isoform-level eQTLs (MCGA_eQTL and MCGA_isoQTL)

In the previous simulation study, we demonstrated that association analysis based on isoform-level eQTLs was more powerful than gene-level eQTLs in each simulation scenario. To further test this conclusion in real data and identify more potential susceptibility genes for schizophrenia, we first

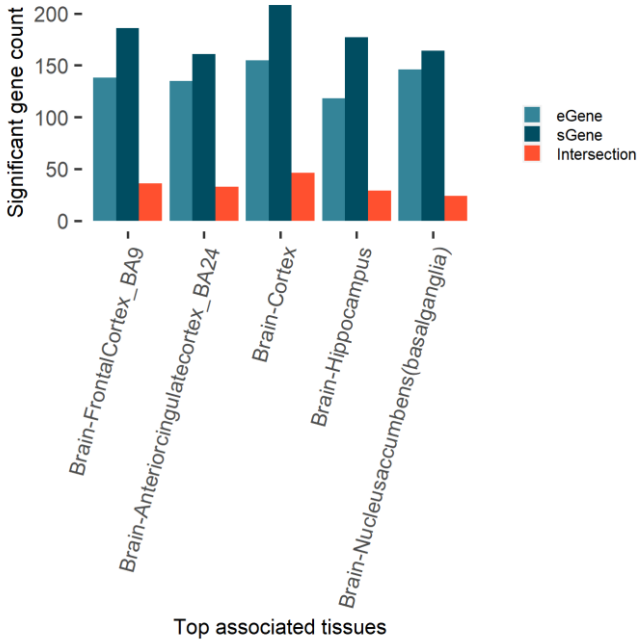
adopted the prioritized phenotype-associated tissues of schizophrenia by DESE[22] , a model to predict potentially phenotype-associated tissues based on gene selective expression analysis. The DESE results showed that all thirteen brain tissues were significantly associated with schizophrenia and ranked the top (**Figure 7a**). For the sake of simplicity, we then computed gene-level eQTLs of the top-five tissues based on gene-level expression profiles and isoform-level eQTLs of the top-five tissues based on transcript-level expression profiles, respectively. Hereinafter, the gene whose expression is associated with at least one SNP was denoted as eGene, and the gene with an isoform whose expression is associated with at least one SNP was denoted as sGene. Then we performed the improved conditional gene-based association analysis based on gene-level eQTLs and isoform-level eQTLs resulted from the corresponding tissues. In each of the top-five tissues, we found the number of potential susceptibility sGenes identified by MCGA_isoQTL was larger than that of potential susceptibility eGenes identified by MCGA_eQTL under the same filter cutoff 2.5E-6 (**Figure 7b**, see details in **Table S4** and **S5**). Besides, we found a considerable number of common genes between the estimated eGenes set and sGenes set in each of the top-five tissues (**Figure 7b**).

We also performed the GO enrichment analysis to further investigate the functional annotations of these potential susceptibility eGenes and sGenes. For the eGene set in each of the top-five associated tissues, we found only the eGenes identified based on the gene-level eQTLs of Brain-FrontalCortex (BA9) had GO enrichment terms (**Figure 7c**). For the sGene set in each of the top-five associated tissues, we found the potential susceptibility sGenes in Brain-FrontalCortex(BA9), Brain-Anteriorcingulatecortex (BA24) and Brain-Hippocampus all

had different GO enrichment terms (**Figure 7d**). We further combined the potential susceptibility eGenes and sGenes of all top-five tissues, respectively. Then we performed GO enrichment analysis based on the combined eGene set and sGene set. We found that both the eGene set and sGene set were enriched with many neuronal-, dendrite- or synaptic signaling-related GO terms (see details in **Table S6**). Then we searched the PubMed database with the combined eGene set (578 unique genes) and sGene set (696 unique genes), and found 133 of the 578 (~ 23.0%) and 168 of the 696 (~ 24.1%) potential susceptibility genes estimated by MCGA_eQTL and MCGA_isoQTL had at least one search hit, respectively (see details in **Table S7** and **S8**). The biologically sensible GO enrichment results and the PubMed search results both implied that the potential susceptibility sGenes and eGenes might have strong associations with schizophrenia.

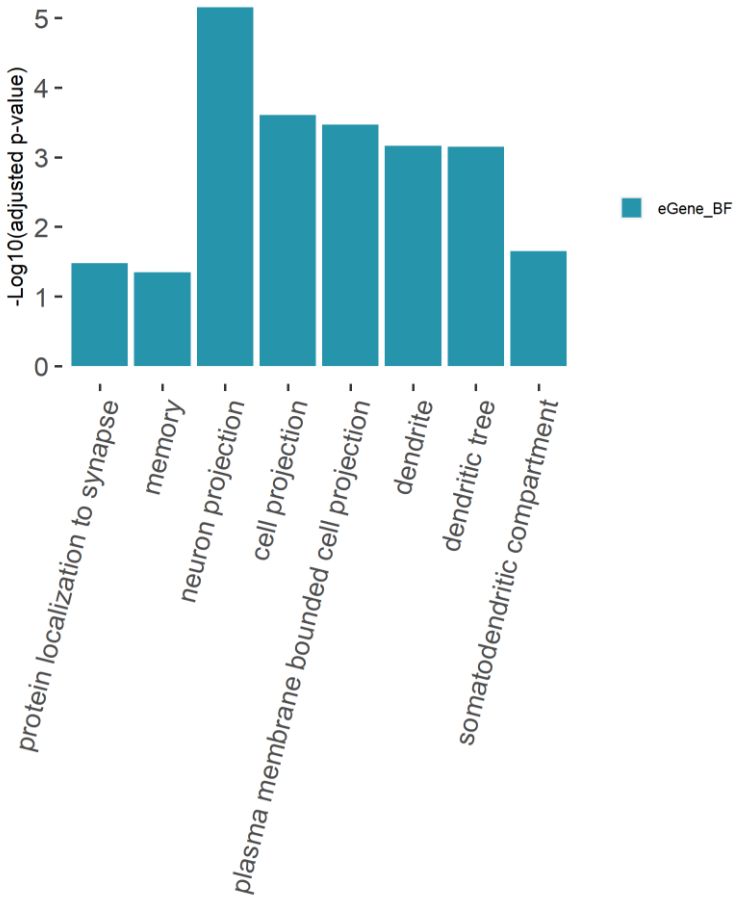


(b)

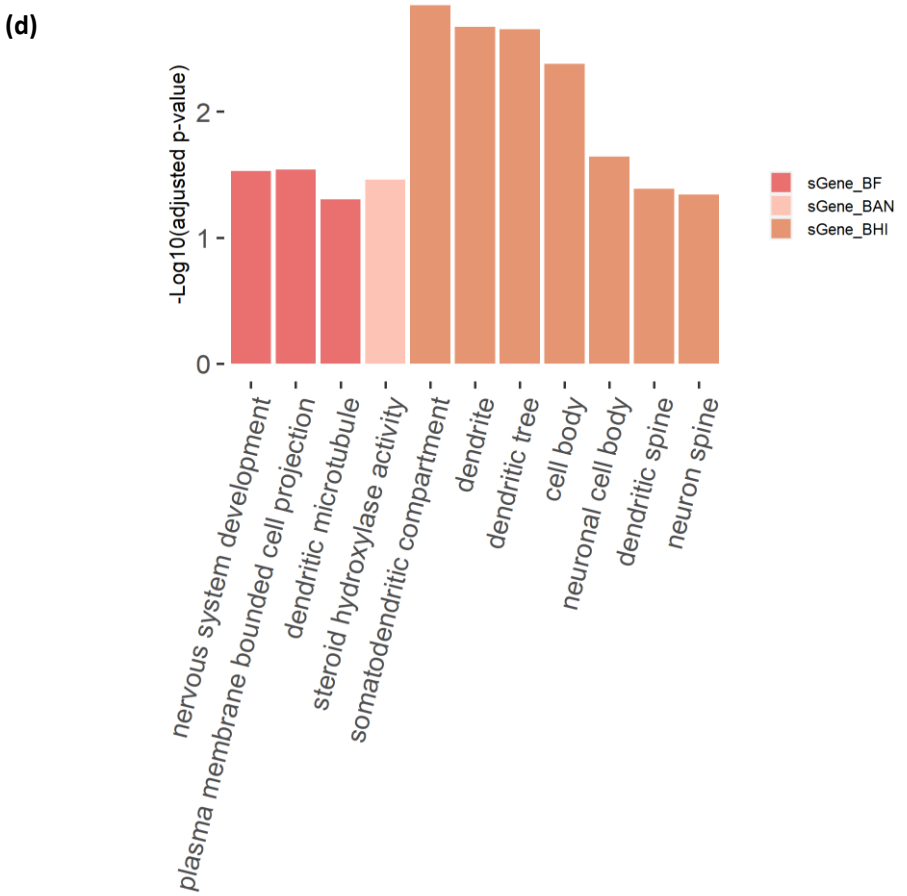


351

(c)



352



353

354 **Figure 7: Potentially associated tissues of schizophrenia and the characterization of potential**
355 **susceptibility genes in corresponding tissues. a)** The statistical significance of tissues estimated
356 to be associated with schizophrenia. The X-axis denotes the tissue names. The Y-axis means the
357 negative $\log_{10}$(p-value) of the Wilcoxon rank-sum test. **b)** The comparison of eGenes and sGenes
358 in each of the top-five associated tissues. **c**) The GO enrichment terms of eGenes in Brain-
359 FrontalCortex(BA9). BF: Brain-FrontalCortex(BA9). **d**) The GO enrichment terms of sGenes in
360 Brain-FrontalCortex(BA9), Brain-Anteriorcingulatecortex (BA24) and Brain-Hippocampus. BF:
361 Brain-FrontalCortex(BA9). BAN: Brain-Anteriorcingulatecortex(BA24). BHI:
362 Brain-Hippocampus.

363

364 **The advantages of MCGA_isoQTL versus MCGA_Dist and MCGA_eQTL**

365 The connectivity score of a gene in the weighted co-expression network might imply its real
366 association with other genes, and highly connected genes are often defined as hub genes. These
367 hub genes are located in or near the center of corresponding co-expression modules and might

play important roles in trait development[23]. We built an unsigned weighted co-expression network for each top-five tissue and investigated the normalized intra-module connectivity of potential susceptibility genes in each co-expression module. We found that the normalized intra-module connectivity scores of potential susceptibility genes in Brain-FrontalCortex_BA9 and Brain-Nucleusaccumbens(basal ganglia) identified by MCGA_eQTL were significantly larger than that of non-susceptibility genes. Interestingly, the normalized intra-module connectivity scores of potential susceptibility genes identified by MCGA_isoQTL were significantly larger than that of non-susceptibility genes in all the top-five schizophrenia-associated tissues (Wilcoxon rank-sum test $p<0.05$) (**Table 2**).

Table 2: Comparison of the normalized intra-module connectivity of potential susceptibility genes identified by MCGA_eQTL and MCGA_isoQTL.

Tissue Name	**Susceptibility vs. non-susceptibility genes by MCGA_eQTL**		**Susceptibility vs. non-susceptibility genes by MCGA_isoQTL**	
	statistic	p-value	statistic	p-value
Brain-FrontalCortex(BA9)	2.4374	**0.0148**	2.5426	**0.0110**
Brain-Anteriorcingulatecortex(BA24)	1.4303	0.1526	4.0558	**5.0000E-5**
Brain-Cortex	1.4173	0.1564	1.9946	**0.0461**
Brain-Hippocampus	1.8439	0.0652	2.6019	**0.0093**
Brain-Nucleusaccumbens(basalganglia)	2.2587	**0.0239**	3.4861	**0.0005**

We next compared the potential susceptibility genes predicted by MCGA_Dist, MCGA_eQTL and MCGA_isoQTL. As shown in **Figure 8**, twenty-three genes were collectively predicted to be susceptible to schizophrenia by the three models of MCGA. As MCGA_isoQTL could output

susceptibility gene-isoform pairs of corresponding tissues, we further got the corresponding susceptibility isoforms of twenty-three genes in corresponding tissues (see details in **Table S9)**. Interestingly, we found that susceptibility isoforms for a gene varied greatly in different tissues. For example, *ABCC8* and *LINC01415* both had one susceptibility isoform in the top-five tissues. *ENST00000529967* of *ABCC8* was significantly associated with schizophrenia only in Brain-Hippocampus, while *ENST00000587320* of *LINC01415* was significantly associated with schizophrenia in Brain-Cortex, Brain-FrontalCortex(BA9) and Brain-Nucleusaccumbens(basal ganglia). We also found that different isoforms of the same gene were predicted to be significantly associated with schizophrenia in different tissues, such as *ENST00000377600* of *BTN2A1* significantly associated with schizophrenia in Brain-Cortex and *ENST00000312541* of *BTN2A1* significantly associated with schizophrenia in Brain-FrontalCortex(BA9). MCGA_isoQTL can help predict potential susceptibility genes and isoforms of corresponding phenotype-associated tissues at a more precise level.

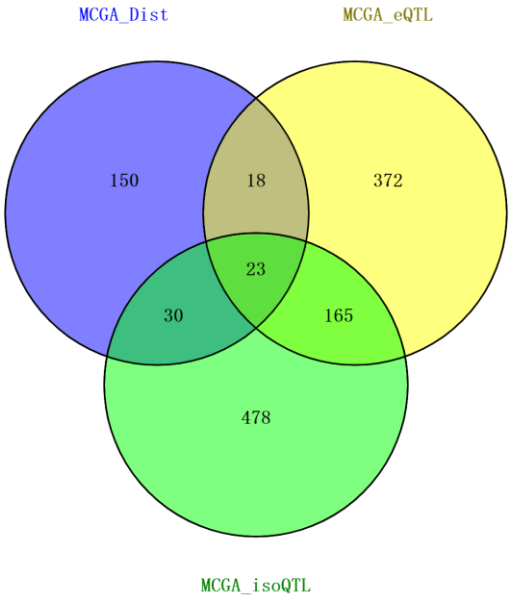


Figure 8: Comparison of the potential susceptibility genes predicted by MCGA_Dist, MCGA_eQTL and MCGA_isoQT. The venn plot was used to show the intersection genes and

401 unique genes among MCGA_Dist, MCGA_eQTL and MCGA_isoQTL.

402

403 Except for the advantage of identifying the potential susceptibility isoforms, we found that some
404 of the susceptibility genes exclusively predicted by MCGA_isoQTL were also biologically
405 sensible. We found 478 susceptibility genes exclusively predicted by MCGA_isoQTL (**Figure 8**).
406 We searched the PubMed database with these exclusive genes. The search results showed that 101
407 of the 478 (21.1%) genes each had a least one search hit which reported the associations of these
408 exclusive genes with schizophrenia. Moreover, 15 of the 478 (3.1%) exclusive genes each had at
409 least ten different supported papers in PubMed. Interestingly, transcription factor 4, i.e., *TCF4*,
410 was reported by 100 papers in the PubMed database. *TCF4* is broadly expressed and may play an
411 important role in nervous system development [provided by RefSeq, Jul 2016]. The important
412 examples of biologically sensible genes exclusively identified by MCGA_isoQTL were listed in
413 **Table 3**.

414

415 **Table 3:** The important examples of potential susceptibility genes exclusively predicted by
416 MCGA_isoQTL with at least ten search hits in PubMed.

Unique sGene[b]	# of PubMed paper hits
TCF4	100
RANGAP1	100
FUT2	97
DRD1	63
COPE	36
NCAM1	20
AS3MT	19
FGFR1	18
DPYSL2	17
PCM1	16
BRD1	15
BDNF-AS	13
GABRA2	12
ADRA1A	11

PBRM1	10

[b]**Unique sGene** means potential susceptibility gene exclusively predicted by MCGA_isoQTL.

Taken together, from the perspective of WGCNA, the statistical significance of the comparison between potential susceptibility genes and non-susceptibility genes implied that these susceptibility genes identified based on isoform-level eQTLs might play more important roles in the weighted gene co-expression network of corresponding tissues. Our results also suggested that incorporating with isoform-level eQTLs can help predict more potential susceptibility genes than gene-level eQTLs in each potentially phenotype-associated tissue. Our results pointed that MCGA_isoQTL could help find some novel and important susceptibility genes which cannot be found by MCGA_Dist and MCGA_eQTL. Moreover, based on the isoform-level eQTLs of each phenotype-associated tissue, the MCGA_isoQTL strategy can also predict the potential susceptibility isoforms in the corresponding tissues.

The druggability of the potential susceptibility genes identified by MCGA

Since drug target genes with genetic support are twice or as likely to be approved than target genes with no known genetic associations[24,25], we searched the DrugBank 5.0 database[26] and found that nine potential susceptibility genes identified by MCGA were the target genes of multiple FDA-approved antipsychotics (**Table 4**). Several most popular target genes of approved antipsychotics, i.e., *DRD2*, *DRD1* and *ADRA1A*, were identified by different MCGA models and the results suggested that the three models could complement each other to identify more potential target genes.

Table 4: The potential susceptibility genes that are also the target genes of approved antipsychotics.

Target gene	# of approved antipsychotics	Models
DRD2	32	MCGA_Dist & MCGA_eQTL
DRD1	22	MCGA_isoQTL
ADRA1A	21	MCGA_isoQTL
CHRM3	10	MCGA_eQTL
ADRA2B	10	MCGA_eQTL
HTR3A	8	MCGA_eQTL
OPRD1	1	MCGA_Dist
GABRA2	1	MCGA_isoQTL
CYP2D6	1	MCGA_eQTL & MCGA_isoQTL

To further investigate the druggability of the potential susceptibility genes, we searched the Drug Gene Interaction database (DGIdb v4.2.0) [27] and filtered the drug-gene interaction terms with at least one supported PubMed paper. After the filtration, we kept 30,072 unique drug-gene interaction terms and found 679 unique drug-gene interaction terms for 34 FDA-approved antipsychotics (see details in **Table S10**). Then we put the full list of potential susceptibility genes (by MCGA_Dist, MCGA_eQTL and MCGA_isoQTL, respectively) into DGIdb to investigate if the "antipsychotic"- "susceptibility gene" interactions were enriched in DGIdb. As shown in **Table 5**, we found that "antipsychotic" - "potential susceptibility genes" identified by the three models of MCGA were all significantly enriched in DGIdb. Moreover, as shown in **Figure 8**, 372 out of 578 and 478 out of 696 potential susceptibility genes were exclusively identified by MCGA_eQTL and MCGA_isoQTL, respectively. We found 253 unique drug-gene interaction terms for susceptibility genes exclusively predicted MCGA_eQTL (see details in **Table S11**), and 17 of 253 interaction terms were antipsychotics-gene interactions (hypergeometric distribution test *p*-value= 7.05E-5). We also found 291 unique drug-gene interaction terms for susceptibility genes

exclusively predicted MCGA_isoQTL (see details in **Table S12**), and 28 of 291 interaction terms were antipsychotics-gene interactions (hypergeometric distribution test *p*-value = 1.48E-10). We also investigated the potential druggability of the susceptibility genes identified by MCGA. Among the 42 potentially druggable gene categories in DGIdb, we found the top three potentially druggable categories for the susceptibility genes identified by MCGA_Dist, MCGA_eQTL and MCGA_isoQTL were all "DRUGGABLE GENOME" (63 vs. 129 vs. 137. The number denotes the gene set size belonging to this category, same as below), "ENZYME" (33 vs. 68 vs. 90) and "KINASE" (22 vs. 40 vs. 60) (see details in **Table S13**). Taken together, our results showed that some of the potential susceptibility genes identified by MCGA had the potential to be druggable, and the application of eQTLs (especially the isoform-level eQTLs) could aid MCGA to identify more potentially druggable genes.

Table 5: The enrichment of drug-gene interaction terms in DGIdb for susceptibility genes identified by MCGA.

Models	**# of total drug-gene interaction terms**	**# of antipsychotics-gene interaction terms**	**Enrichment *p*[c]**
MCGA_Dist	332	33	1.37E-12
MCGA_eQTL	831	64	1.39E-17
MCGA_isoQTL	792	46	6.66E-9

[c]**Enrichment *p*** denotes the *p*-value of hypergeometric distribution test.

Discussion

In this study, we proposed a multi-strategy conditional gene-based association framework, MCGA, based on a new correlation matrix of chi-squared statistics to identify the potential susceptibility

genes and isoforms for complex phenotypes. Comparing with the unconditional association test and likelihood ratio test, MCGA showed a lower type I error rate and higher statistical power. Since MCGA is a gene-based method, in this study, we adopted three strategies to map a variant to a gene, i.e., mapping based on physical position, gene-level eQTLs and isoform-level eQTLs. We implemented these three mapping strategies in corresponding three conditional gene-based association models, i.e., MCGA_Dist, MCGA_eQTL and MCGA_isoQTL, to predict the potential susceptibility genes for schizophrenia.

MCGA_Dist could output a list of genes, while MCGA_isoQTL and MCGA_eQTL could produce a list of genes for each potential phenotype-associated tissue because of the usage of gene-level eQTLs and isoform-level eQTLs of each tissue. Though MCGA_Dist predicted a relatively small size of susceptibility genes, these genes were enriched with plenty of neuronal- or synaptic signaling-related GO terms. Similar results of other research were obtained by gene-set analyses, which demonstrated that genetic variants associated with schizophrenia were enriched with synaptic pathways[28]. Besides, considerable amounts of these genes had been reported by many research papers in the PubMed database to support their associations with schizophrenia.

Since MCGA_Dist might omit some remote but important gene-variant associations, we improved MCGA_Dist with MCGA_eQTL and MCGA_isoQTL. We performed a simulation study and demonstrated that isoform-level eQTLs were more powerful than gene-level eQTLs in association analysis. Moreover, we found in real data that the size of the susceptibility gene set for schizophrenia predicted by MCGA_isoQTL was larger than MCGA_eQTL in each

phenotype-associated tissue under the same threshold. Further, we found MCGA_isoQTL had two advantages over MCGA_eQTL and MCGA_Dist. First, several important potential susceptibility genes were exclusively predicted by MCGA_isoQTL. For example, fifteen potential susceptibility genes exclusively predicted by MCGA_isoQTL each had at least ten search hits in PubMed, which implied these genes were popular in schizophrenia studies. Second, to our best knowledge, MCGA_isoQTL was the first conditional gene-based association approach to produce a list of phenotype-associated isoforms (or transcripts).

In addition, we investigated the druggability of the susceptibility genes for schizophrenia identified by MCGA. Several susceptibility genes identified by MCGA were also the popular target genes of multiple FDA-approved antipsychotics. Besides, the "susceptibility gene"-"antipsychotics" interactions were enriched in DGIdb. The druggablilty of the important susceptibility genes, especially the sGenes identified based on isoform-level eQTLs, provided more credible supports for the utility of MCGA.

Our framework might have three potential applications. First, MCGA_Dist can be used to predict potential susceptibility genes and isoforms for other complex phenotypes. Second, based on the assumption that the distribution of expression profiles of true susceptibility genes might change before and after therapeutic drug treatment, MCGA_Dist can be used to perform drug repositioning analysis based on the drug perturbed expression profile. Third, since MCGA_eQTL and MCGA_isoQTL can help predict potential susceptibility genes in each potential phenotype-associated tissue, our framework can help perform synergistic drug combination

519 prediction to screen drugs that can simultaneously perturb the expression of potential
520 susceptibility genes in each potential phenotype-associated tissue.
521
522 The present study was limited by several factors. First, the moderate sample size (ranging from
523 129 ~ 205) and mixed populations in GTEx v8 might both reduce the accuracy of
524 gene/isoform-level eQTLs. Future genetic studies based on increased sample sizes might alleviate
525 this problem. Second, the size of the susceptibility genes identified by MCGA_eQTL (578) and
526 MCGA_isoQTL (696) was a little larger than conventional studies. One of the reasons might be
527 that five brain regions were involved in the present study, and each brain region might have very
528 different dysfunctional genes associated with schizophrenia. We also used MAGMA to identify
529 the susceptibility genes of schizophrenia with the same GWAS summary statistics and found that
530 MAGMA also identified ~ 600 potential susceptibility genes with the basic parameter setup (see
531 details in **Table S14**). Susceptibility genes identified by MCGA_eQTL and MCGA_isoQTL had
532 many biologically meaningful annotations (such as neuronal- or synaptic signaling-related terms)
533 in the GO databases, and some susceptibility genes were the target genes of multiple
534 antipsychotics, and more than 20% of the susceptibility genes had been previously reported by
535 other schizophrenia research in the PubMed database. Though these potential susceptibility genes
536 were lack of systematically experimental validation, we shared the potential susceptibility genes in
537 Table **S1**, **S4** and **S5** and encouraged follow-up studies to evaluate the function and roles of these
538 susceptibility genes in the development of schizophrenia.
539
540 In conclusion, in this study, we proposed a new statistical framework to predict potential

susceptibility genes for complex phenotypes based on GWAS summary statistics and gene/isoform-level eQTLs in a multi-tissue context. The application of our framework to schizophrenia revealed many novel susceptible and druggable genes. Besides, the usage of isoform-level eQTLs can be an important supplement for the conventional gene-based approach. The framework was packaged and implemented in our integrative platform KGGSEE (http://pmglab.top/kggsee/#/). We hope our framework can facilitate researchers to gain more insights into the phenotype-associated genes and isoforms of complex phenotypes.

Materials and Methods

The new effective chi-squared statistics (ECS) for conditional gene-based association analysis

We improved our previously proposed effective chi-squared test[10] for a more efficient conditional gene-based association analysis based on a new correlation matrix of chi-squared statistics. The improved effective chi-squared statistics had two methodological advances to address the potential inflation issue, i.e., a type I error-controlled correlation matrix of the observed chi-squared statistics and a non-negative least square solution for the independent chi-squared statistics. The reasoning process was as follows. Suppose there were n loci in a set of genes. One wanted to calculate the association p-value of another physically nearby gene (containing m loci) conditioning on the set of genes (n loci). The first step of the conditional analysis was to produce effective chi-squared statistics for the set of genes (n loci) and all the genes (n+m loci in total). Each locus had a p-value for phenotype association in the GWAS. The p-values were converted to

562 corresponding chi-squared statistics with the degree of freedom 1. According to Li et al.[10], each
563 locus could be assumed to have a virtually independent chi-squared statistic. An observed
564 marginal chi-squared statistic of a locus was equal to the summation of its virtually independent
565 chi-squared statistic and the weighted virtually independent chi-squared statistic of nearby loci.
566 The weight was related to the chi-squared statistics correlation, which was a key parameter of the
567 analysis. The correlation of chi-squared statistics between two loci was approximated by the
568 absolute value of genotypic correlation to the power of *c*, i.e., $|r|^c$. Here, we derived that the key
569 parameter, i.e., exponent *c*, ranged from 1 to 2, corresponding to different non-centrality
570 parameters of a non-central chi-squared distribution (See the derivation in the next section).
571 According to Li et al.[10], the *n* virtually independent chi-squared statistics of the gene set could be
572 approximated by a linear transformation of the *n* observed chi-squared statistics (**Formula (1)**),
573

$$\begin{bmatrix} \acute{x}_1^2 & d_1 \\ \cdots & \cdots \\ \acute{x}_n^2 & d_n \end{bmatrix} \approx \begin{bmatrix} 1 & \cdots & |r_{1,n}|^c \\ \cdots & 1 & \cdots \\ |r_{n,1}|^c & \cdots & 1 \end{bmatrix}^{-1} \times \begin{bmatrix} x_1^2 & 1 \\ \cdots & \cdots \\ x_n^2 & 1 \end{bmatrix} \quad (1)$$

574 ,
575
576 where $\acute{x}_n^2(\geq 0)$, $d_n(>0)$, x_n^2 and $|r_{i,j}|$ denoted a virtually independent chi-squared statistic,
577 degree of freedom of the virtually independent chi-squared statistic, an observed chi-squared
578 statistic and the absolute value of the LD correlation coefficient (approximated by genotypic
579 correlation), respectively. The effective chi-squared statistic $\acute{S}_n$ with the degree of freedom $\acute{d}_n$
580 of the *n* loci was then obtained by **Formula (2)**:
581

$$\begin{cases} \acute{S}_n = \sum_{i=1}^{n} \acute{x}_i^2 \\ \acute{d}_n = \sum_{i=1}^{n} d_i \end{cases} \quad (2)$$

.

The effective chi-squared statistics ($\acute{S}_{n+m}$) and degree of freedom ($\acute{d}_{n+m}$) of the *n*+*m* loci could be calculated in the same way.

The effective chi-squared statistics of the *m* loci conditioning on the *n* loci was then approximated by **Formula (3)**,

$$\acute{S}_{m|n} = \acute{S}_{n+m} - \acute{S}_n \quad (3)$$

with the degree of freedom $\acute{d}_{m|n} = \acute{d}_{n+m} - \acute{d}_n$.

Because $\acute{d}_{m|n}$ was no longer an integer, we used the Gamma distribution to calculate the *p*-values. Given the above statistics and degree of freedom, the *p*-value was equal to $F(x \geq \frac{\acute{S}_{m|n}}{2}; \frac{\acute{d}_{m|n}}{2}, 2)$, where the $F(x)$ function was the cumulative distribution function of a Gamma distribution.

Because the virtually independent chi-squared statistics and degrees of freedom were expected to be larger than 0, we adopted a sequential coordinate-wise algorithm to approximate them[29]. This algorithm avoided unstable solutions in the above linear **Formula (1)** due to stochastic errors in the correlation matrix and observed chi-squared statistics.

After the above multiple approximations, it was still difficult to obtain the analytic solution for the exponent *c* in **Formula (1)**. We proposed a grid search algorithm to find a favorable value of

exponent *c* to control type I error rates of the effective chi-squared tests. The error rate was examined by divergence from a uniform distribution between an obtained and theoretical top 1% *p*-values given a *c* value, measured as mean log fold change (MLFC) [30]. In the grid search process, we increased *c* from 1.00 to 2.00 by an interval of 0.01 because it ranged from 1 to 2 (see the derivation in the **Materials and Methods**). The *c* value leading to the minimal MLFC was defined as the favorable *c* value. We considered in total 84 parameter settings, i.e. a combination of three different sample sizes (10,000, 20,000 and 40,000) and 14 different variant sizes (10, 30, 50, 80, 100, 125, 150, 200, 250, 300, 400, 500, 800, and 1000) for both binary and continuous traits, respectively. For a parameter setting, 40,000 datasets were simulated and used to produce *p*-values to determine the favorable *c* value for the setting. A region on chromosome 2 [chr2: 169428016-189671923] was randomly drawn for the simulation. The allele frequencies and LD structure of variants in the European panel of the 1000 Genomes Project were used as a reference to simulate genotype data by the HapSim algorithm [31]. According to either the Bernoulli distribution or Gaussian distribution, each subject was randomly assigned a phenotype value under the null hypothesis. The Wald test under either logistic regression or linear regression in which the major and minor allele was encoded as 0 and 1 was used to produce the association *p*-value at each variant. The *p*-values of the variants were then analyzed by the effective chi-squared test for the gene-based association analysis.

Approximate the correlation of chi-square statistics under the alternative hypothesis

Let two normal random variables $\boldsymbol{X}\sim\boldsymbol{N}(\boldsymbol{\mu_1},\boldsymbol{\sigma_1^2})$ and $\boldsymbol{Y}\sim\boldsymbol{N}(\boldsymbol{\mu_2},\boldsymbol{\sigma_2^2})$ have covariance $\boldsymbol{c}$. Note that

a squared normal random variable has non-central chi-square distribution and the squared mean of the former is called noncentrality parameter. The two variables can also be factorized as

$= \boldsymbol{\mu_1 + \sigma_1 U}$, $\boldsymbol{Y = \mu_2 + \sigma_2\,(\rho U + \sqrt{1-\rho^2}V)}$ where $\boldsymbol{U, V \overset{iid}{\sim} N(0,1)}$ and $\boldsymbol{\rho = c/\sigma_1\sigma_2}$.

Then we can calculate the co-variances of the two non-central chi-square variables $\boldsymbol{X^2}$ and $\boldsymbol{Y^2}$ by the factorized variables, $\boldsymbol{cov(X^2, Y^2) = \frac{4\mu_1\mu_2 c + 2c^2}{\sqrt{4\mu_1{}^2\sigma_1{}^2 + 2\sigma_1{}^4}\sqrt{4\mu_2{}^2\sigma_2{}^2 + 2\sigma_2{}^4}}}$. Suppose $\boldsymbol{X}$ is the Z score of the true casual variant and $\boldsymbol{Y}$ is the Z score of a non-functional variant in LD (coefficient $\boldsymbol{r}$) with the causal variant. One can assume $\boldsymbol{\mu_2 = r\mu_1, \sigma_1{}^2 = 1, \sigma_2{}^2 = 1}$ and $\boldsymbol{c = r}$. Therefore, the correlation of $\boldsymbol{X^2}$ and $\boldsymbol{Y^2}$ can be simplified as, $\boldsymbol{cor(X^2, Y^2) = \frac{4r^2\mu_1{}^2 + 2r^2}{\sqrt{4\mu_1{}^2+2}\sqrt{4\mu_1{}^2r^2+2}}}$.

Under the null hypothesis, $\boldsymbol{\mu_1 = 0}$ then $\boldsymbol{cor(X^2, Y^2) = r^2}$. Under the alternative hypothesis of large scaled sample, the $\boldsymbol{\mu_1}$ or the noncentrality parameter becomes very large, correlation of $\boldsymbol{X^2}$ and $\boldsymbol{Y^2}$ become close to $\boldsymbol{r}$. $\boldsymbol{\mu_1 \to \infty,\ cor(X^2, Y^2) = \frac{4r}{\sqrt{4+\frac{2}{\mu_1{}^2}}\sqrt{4+\frac{2}{\mu_1{}^2r^2}}} + \frac{2r^2}{\sqrt{4\mu_1{}^2+2}\sqrt{4\mu_1{}^2r^2+2}} \to r}$.

Overall, the correlation between the two (non-central) chi-square ranges from r^2 to r.

The conditional gene-based association analysis for genome-wide association study

In a GWAS, all genes were firstly calculated with the *p*-values of unconditional gene-based association test using the above effective chi-squared statistics. For a given *p*-value cutoff, the significant genes were extracted and subjected to the conditional gene-based association analysis. When there were multiple significant genes in an LD block, the genes were conditioned one by one in a pre-defined order. In the present conditional analysis, the order of the gene was defined according to the unconditional *p*-value of the gene. Here we assigned the genes within 5 Mb into

the same LD block. The conditional *p*-value of the first gene was defined as its unconditional *p*-value. The conditional *p*-value of the second gene was obtained by conditioning on the first gene, and that of the third gene was obtained by conditioning on the top two genes. The conditional *p*-values of subsequent genes were calculated according to the same procedure.

Simulations for investigating type I error and power of the conditional gene-based association analysis

Extensively independent computer simulations based on a different reference population (i.e., EAS) in different genomic regions were performed to investigate type I error and the power of the conditional gene-based association test. To approach the association redundancy pattern in realistic scenarios, we used real genotypes and simulated phenotypes. The high-quality genotypes of 2,507 Chinese subjects from a GWAS were used [32], and phenotypes of subjects were simulated according to the genotypes under an additive model. Given total variance explained by *n* independent variants, V_g, the effect of an allele at a bi-allelic variant was calculated by $a = \sqrt{V_g/[\sum_{i=1}^{n} 2P_{A_i}(1-P_{A_i})]}$, where P_{A_i} was the frequency of alternative alleles. The total expected effect *A* of a subject was equal to a*[the number of alternative alleles of all the n variants]. Each subject's phenotype was simulated by *P*=A+e, where *e* was sampled from a normal distribution $N(0, 1\text{-}V_g)$. We randomly sampled three pairs of genes, i.e., *SIPA1L2* vs. *LOC729336, CACHD1* vs. *RAVER2*, and *LOC647132* vs. *FAM5C*. The three pairs of genes represented three scenarios where the nearby gene (i.e., the first gene) had similar (*SIPA1L2* vs. *LOC729336*), larger (*CACHD1* vs. *RAVER2*) and smaller (*LOC647132* vs. *FAM5C*) variant size than the target gene (i.e., the second gene) in terms of SNP number, respectively. In the type I error investigation, the target gene had

no QTLs, while the nearby gene had one or two QTLs. In the investigation of the statistical power, both genes had QTLs.

For power comparison, the likelihood ratio test based on linear regression was adopted to perform the conditional gene-based association analysis with raw genotypes. In the full model, genotypes of all SNPs encoded as 0, 1, or 2 according to the number of alternative variants entered the regression model as explanatory variables. In the subset model, the SNPs of the nearby genes entered the regression model. The calculation of the likelihood ratio test was performed according to the conventional procedure. The R packaged "lmtest" (version 0.9.37) was adopted to perform the likelihood ratio test.

Simulations for comparing the power of gene-level eQTLs and isoform-level eQTLs in gene-based association tests

We compared the power of conventional gene-based association tests, gene-level eQTLs guided gene-based association tests and isoform-level eQTLs guided gene-based association tests by simulation studies. Assume some variants regulate gene expression, and the gene expression subsequently influences the phenotype. The same region on chromosome 2 [chr2: 169428016-189671923] was considered for the simulation. In the EUR panel of 1000 Genomes Project[33], this region contains 1600 common variants (MAF>0.05). Genotypes of the variants were simulated given allelic frequencies and LD correlation matrix according to the HapSim algorithm[31]. Phenotypes were simulated under a polygenic model of random effect[34]. According to severe LD pruning (r^2<0.01), eighty-two independent variants were extracted from the 1600

variants. The SNPs' genotypes (*s*) contributing to the phenotypes were then standardized as, $g = (s - 2q)/\sqrt{2q(1-q)}$, where q was the allele frequency of alterative allele. Phenotypes were simulated under a polygenic model of random effect[34]. We assumed 40% of the independent causal variants (m_X) regulated gene expression (total heritability h_X^2). The expression of a gene (X) was simulated according to **Formula (4)**:

$$X = \sum_{i=1}^{m_X} g_i \beta_{X,i} + \epsilon_X, \quad (4)$$

where $\beta_{1,i} \sim N(0, h_X^2/m_X)$ and $\epsilon_X \sim N(0, 1 - h_X^2)$.

The gene expression then contributed δ to a phenotype (Y). The phenotype value was simulated according to the **Formula (5)**:

$$Y = \delta X + \epsilon_Y, \quad (5)$$

where $\epsilon_Y \sim N(0, 1 - \delta^2)$. Here Y was a continuous phenotype. For a binary phenotype, a cutoff *t* was set according to a given disease prevalence *K* under a standard normal distribution and the liability threshold model [35]. Subjects with simulated Y values $\geq t$ were set as patients, and others were set as normal controls.

When a gene had multiple isoforms, we assumed one of the isoforms was associated with phenotype and simulated expression values of the isoform according to the above regulation

712 model (**Formula (5)**). The expression values of the remaining isoforms were simulated by the
713 standard normal distribution $N(0,1)$. The expression profile of a gene with multiple isoforms was
714 averaged by the expression profiles of all the isoforms belonging to the gene. The gene-level
715 eQTLs and isoform-level eQTLs were examined by the Wald test under the linear regression
716 model. The variant-phenotype association analysis was performed based on the conventional
717 association analysis procedure, and the statistical significance cutoff was set at p-value<0.001.
718

719 **Genome-wide association study of schizophrenia**

720 The GWAS summary statistics of schizophrenia included 53,386 cases and 77,258 controls of
721 European ancestry (hg19 assembly). Genotypes in the CEU panel from the 1000 Genomes Project
722 were used to correct for the relatedness of the summary statistics. To predict the potential
723 susceptibility genes of schizophrenia, the variants in the major histocompatibility complex (MHC)
724 region, i.e., chr6:27,477,797-34,448,354, were excluded because of high polymorphism in the
725 present study. Detailed descriptions of population cohorts, quality control methods and association
726 analysis methods can be found in reference[20]. The summary statistics can be accessed at the
727 Psychiatric Genomics Consortium.
728

729 **The Genotype-Tissue Expression (GTEx) project**

730 The GTEx project (release v8) created a resource including whole-genome sequence data and
731 RNA sequencing data from ~ 900 deceased adult donors[21]. Four tissues or cell types (i.e., whole
732 blood, cells-Leukemiacellline_CML, pancreas and pituitary) were filtered out and not included in

733 the following analyses because of the small sample size or weak correlation of gene expression
734 profiles with most of the other tissues.
735

736 **GO annotation of the potential susceptibility genes**

737 Functional enrichment analyses were performed by g:Profiler[36]. GO terms, i.e., biological
738 processes (BP), molecular functions (MF) and cellular components (CC), were mainly concerned.
739 g:Profiler is based on Fisher's one-tailed test, and the statistical *p*-value is multiple
740 testing-corrected. Significant GO terms were filtered by the threshold of "Padj" <0.05. The bar
741 plots of GO enrichment terms were drawn based on R-4.0.3.
742

743 **Construction of the weighted gene co-expression network in multi-brain tissues**

744 The fully processed, filtered and normalized gene-level expression profiles from GTEx v8 were
745 used to construct the weighted gene co-expression networks for the top-five brain tissues by R
746 package "WGCNA" (v1.69). WGCNA was performed to build an unsigned gene co-expression
747 network following the standard procedure, and all the parameters were used as recommended, and
748 the soft-threshold was set to 6 after testing a series of soft threshold powers (range 2 to 20). As for
749 the construction of gene co-expression modules, the hierarchical cluster tree in the co-expression
750 network was cut into gene modules using the dynamic tree cut algorithm with a minimum module
751 size of 30 genes[37]. The normalized intra-module connectivity value was computed by setting the
752 options "scaleByMax = T".
753

Drug Gene Interaction (DGIdb) database

DGIdb (v4.2.0) provides a resource of genes that have the potential to be druggable[27]. DGIdb contains two main classes of druggable genome. The first class includes genes with known drug interactions, and the other includes genes that are potentially druggable according to their membership in gene categories associated with druggability. DGIdb includes 42 potentially druggable categories and 49 interaction types (including inhibitors, activators, cofactors, ligands, vaccines and many interactions of unknown types). Only the drug-gene interaction terms with at least one supported PubMed paper were used in the present study.

PubMed search

To find supports from published research, we performed a text-mining analysis based on PubMed database on June 3rd, 2021. We searched the PubMed database with the items of "((schizophrenia[tiab]+OR+Schizophrenia[tiab]+OR+SCZ[tiab])+AND+(gene name[tiab])+AND+(gene[tiab]+OR+genes[tiab]+OR+mRNA[tiab]+OR+protein[tiab]+OR+proteins[tiab]+OR+transcription[tiab]+OR+transcript[tiab]+OR+transcripts[tiab]+OR+expressed[tiab]+OR+expression[tiab]+OR+expressions[tiab]+OR+locus[tiab]+OR+loci[tiab]+OR+SNP[tiab]))&datetype=edat&retmax=100". The java script output a file with the first column representing gene name, the second column representing the synonyms of the gene name, the last column representing the PubMed ids of hit papers.

774 **Identification of the potentially phenotype-associated tissues of schizophrenia**

775 To estimate the potentially phenotype-associated tissues, a framework called DESE (also
776 implemented in KGGSEE) proposed by our lab in a recent work was used [22]. DESE needs three
777 kinds of input datasets, i.e., the expression profiles of various tissues, reference genotype and
778 GWAS summary statistics, and outputs the estimated phenotype-associated tissues.

779

780 Specifically, the isoform-level expression profiles of 50 tissues in GTEx v8 were used. The
781 isoform-level expression profiles were preprocessed like this: the index column of the
782 preprocessed expression file was isoform symbol name, and each of 50 tissues or cell types had
783 one column representing the average expression value (i.e., mean value) of corresponding subjects
784 with the tissue. The Genotypes in the EUR panel from the 1000 Genomes Project (phase 3) were
785 downloaded from IGSR and used as reference genotype data. Three columns, i.e., chromosome
786 identifier (CHR), base-pair position (BP) and *p*-value (P) in GWAS summary statistics, were used.
787 SNPs with minor allele frequency (MAF) less than 0.05 were excluded. Only genes approved by
788 HGNC were included in the following analyses. The multiple testing adjustment method was the
789 standard Bonferroni correction, and the cutoff for the adjusted *p*-value was set as $p<0.05$. The
790 detailed commands of DESE to estimate potential phenotype-associated tissues are described on
791 the KGGSEE website. The bar plot of the rank of potential phenotype-associated tissues was
792 drawn based on R-4.0.3.

793

794 **Computation of gene-level eQTLs and isoform-level eQTLs**

795 The present study focused on the cis-eQTLs. Specifically, two files were put into our integrative
796 platform KGGSEE to produce gene/isoform-level eQTLs for each tissue, namely, expression
797 profiles and corresponding genotype data file from GTEx v8. Two levels (gene-level and
798 isoform-level) expression profiles of 50 tissues were downloaded from the GTEx v8 project, and
799 the TMP value was used in the following analyses. Genes/isoforms were selected based on
800 expression thresholds of > 0 TPM in at least 20% of all samples. The genotype data used for eQTL
801 analyses in GTEx release v8 was based on WGS from 838 donors, which all had RNA-seq data
802 available. Only variants with MAF ≥ 0.05 across all 838 samples were included in the present
803 study. GTEx v8 is based on the human reference genome GRCh38/hg38. Thus, to be consistent
804 with the GWAS results of schizophrenia (hg19 assembly), we converted the GRCh38/hg38
805 coordinates into hg19 by the UCSC LiftOver. All variants were filtered with Hardy–Weinberg
806 disequilibrium (HWD) test *p*-value <1.0E-3. The mapping window was defined as 1 Mb up- and
807 downstream of the gene boundary. If the association test *p*-value of a variant and corresponding
808 expression of gene/isoform was smaller than 0.01, the variant was regarded as a
809 gene-level/isoform-level eQTL of the gene/isoform. It should be noted that the format of the eQTL
810 file is similar to the fasta file. The eQTL data of a gene or isoform starts with the symbol ">". For
811 the gene-level eQTLs file, the symbol ">" is followed by the gene name (e.g., "*LINC00320*"), its
812 Ensembl ID ("*ENSG00000224924*") and chromosome identifier ("21"). For the isoform-level
813 eQTLs file, the symbol ">" is followed by the gene name (e.g. "*LINC00320*"), transcript Ensembl
814 ID ("*ENST00000452561*") and chromosome identifier ("21").
815

816 The gene/isoform-level eQTLs files of 50 tissues in GTEx v8 can be accessed on the KGGSEE
817 website and freely used for research purposes. The detailed commands of KGGSEE to compute
818 gene/isoform-level eQTLs of each tissue are described on the KGGSEE website.
819

820 **Estimation of the potential susceptibility genes and isoforms for schizophrenia**

821 The framework MCGA included three models, i.e., MCGA_Dist, MCGA_eQTL and
822 MCGA_isoQTL, which were all based on the improved ECS. The main difference among the
823 three models was the strategy used to map variants to genes. For MCGA_Dist, if a variant was
824 within a small window, say +/-5 kb, around the gene boundary, then the variant will be assigned
825 onto the gene according to a gene model, e.g., RefSeqGene. For MCGA_eQTL, the variant will be
826 assigned onto the gene if the variant is a gene-level eQTL of the gene. Similarly, for
827 MCGA_isoQTL, the variant will be assigned onto the isoform if the variant is an isoform-level
828 eQTL of the isoform. Another difference between MCGA_Dist and
829 MCGA_eQTL/MCGA_isoQTL was that the latter two were based on the gene/isoform-level
830 eQTLs of each tissue, thus can produce the potential susceptibility genes/isoforms in a multi-tissue
831 context.
832

833 Like our previous model DESE, MCGA contained three iterative steps. In the first step, associated
834 genes with smaller *p*-values of the ECS test were given higher priority to enter the following
835 conditional gene-based association analysis. This step could generate a list of roughly associated
836 genes by removing redundantly associated genes. It should be noted that we dealt with the

mentioned three models of MCGA in different ways. For MCGA_Dist and MCGA_eQTL, the order of a gene entering the conditional gene-based association analysis was determined by its *p*-value of the ECS test. For MCGA_isoQTL, assume gene *A* has *m* isoforms. Each isoform could get a *p*-value based on the ECS test, representing the overall statistical significance of all isoform-level eQTLs (simultaneously variants) associated with this isoform. If the isoform with the smallest *p*-value was isoform *a*, with its *p*-value $p_{a,}$ among the *m* isoforms of gene *A*, we only kept isoform *a* of gene *A* for the following analyses. The adjustment *p*-value for "gene *A* : isoform *a*" pair was adjusted to $m^* p_a$ to enter the following conditional gene-based association analysis.

The second step was to compute the selective expression score of genes/isoforms in each tissue by taking all tissues as the background (see details in reference[22]). The Wilcoxon rank-sum test was then performed by using the selective expression score of the associated gene/isoform set and not-associated gene/isoform set (generated by the first step) in each tissue.

In the third step, all genes/isoforms, including the not-associated genes/isoforms, were ranked in descending order based on the tissue-selective expression score of each gene/isoform. The tissue-selective expression score of a gene/isoform was computed based on the rank of this gene/isoform-selective expression score and the *p*-value of the Wilcoxon rank-sum test between the associated gene/isoform set and not-associated gene/isoform set in each tissue.

In the following iteration, genes/isoforms with higher tissue-selective expression scores (in the third step) were given higher priority to enter the conditional gene-based association analysis (in

the first step). The above three steps were iterated until the *p*-values of the Wilcoxon rank-sum test did not change almost, and then corresponding associated genes/isoforms were deemed to be potentially associated with the phenotype. More details about the iterative procedure can be found in the original papers[22].

MCGA is implemented in our integrative platform KGGSEE. To run MCGA_Dist, three input files were needed, i.e., GWAS summary statistics file, gene-level expression profiles of 50 tissues in GTEx v8, genotypes in EUR panel from 1000 Genomes Project (phase 3). To run MCGA_eQTL and MCGA_isoQTL, four input files were needed, i.e., GWAS summary statistics file, gene-level or isoform-level expression profiles of 50 tissues in GTEx v8, genotypes in EUR panel from 1000 Genomes Project (phase 3) and gene/isoform-level eQTLs file of each estimated disease-associated tissue. Only genes with HGNC gene symbols were considered in the present study. The output result file was a text file that contained multiple information about the association measurement of genes (or "gene: isoform" pairs) with the corresponding phenotype. Multiple testing was corrected by using Bonferroni correction. Significant genes were filtered by the "CondiECSp" threshold cutoff 2.5E-6, where "CondiECSp" meant the p-values of conditional gene-based association test based on the improved ECS. The bar plot of the comparison of potential susceptibility genes was drawn based on R-4.0.3. The venn diagram was drawn based on a web app Venny 2.1.0.

MAGMA

MAGMA is a popular tool for gene and generalized gene-set analysis based on the GWAS

summary statistics. Here the parameters and options were used as the MAGMA (v 1.08) manual recommended. Annotation analysis was firstly performed based on the SNP location file and gene location file (hg19, build 37). The SNP location information was extracted from the same GWAS summary statistics file of schizophrenia. An SNP was mapped to a gene if the SNP was in the window of +/-5kb around the gene boundary (same as MCGA_Dist). The gene analysis was performed based on the annotation results and reference data file which was created from Phase 3 of 1000 Genomes of the European population in reference to human genome build 37. Both gene location file and reference data file were downloaded from the MAGMA website. Multiple testing was corrected by using Bonferroni correction. Significant genes were filtered by the threshold of "*p*-value" 2.5E-6.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The reference genotype data are publicly available in the 1000 Genomes Project[33] in https://www.internationalgenome.org/. The genotype data are publicly available by application from dbGap (study sccession phg001219.v1) and corresponding gene/isoform-level expression

903 profiles are Genotype Tissue Expression (GTEx v8) project [21] in https://gtexportal.org/home/. The
904 summary statistics of schizophrenia are publicly available in Psychiatric Genomics Consortium
905 (PGC) in https://www.med.unc.edu/pgc/. The annotations of drug-gene interaction terms are
906 publicly available in Drug Gene Interaction (DGIdb v4.2.0) database[27] in https://www.dgidb.org/.
907 The information on FDA-approved antipsychotics can be publicly available in DrugBank 5.0[26] in
908 https://go.drugbank.com/. The functional enrichment analyses were performed by g:Profiler[36] and
909 can be publicly available in https://biit.cs.ut.ee/gprofiler. The tool used to draw the venn plot is
910 Venny in https://bioinfogp.cnb.csic.es/tools/venny/index.html. The tool MAGMA[7] and
911 corresponding reference data were downloaded from https://ctg.cncr.nl/software/magma. The
912 source code of MCGA (including MCGA_Dist, MCGA_eQTL and MCGA_isoQTL) is
913 implemented in our integrative software platform KGGSEE and publicly available in
914 http://pmglab.top/kggsee/#/.
915

916 **Competing interests**

917 The authors declare that they have no competing interests.
918

919 **Funding**

920 This work was funded by the National Natural Science Foundation of China (31771401 and
921 31970650), National Key R&D Program of China (2018YFC0910500 and 2016YFC0904300),
922 Science and Technology Program of Guangzhou (201803010116), Guangdong project
923 (2017GC010644).
924

Authors' contributions

M.L. and L.J. conceived the study. M.L. oversaw all aspects of the study. X.L., L.J., C.X, and M.L. developed the models. X.L, L.J. and C.X. performed extensive simulations and real data analyses for performance comparison. X.L., L.J. and M.L. wrote the manuscript. M.L. revised the manuscript. M.J.L. provided useful raw data and did some analysis. All authors commented on and approved the final manuscript.

Acknowledgments

We thank the GTEx Consortium and 1000 Genomes Projects for providing access to expression and sequence variant data. We also appreciate the authors of the schizophrenia study and the Schizophrenia Working Group of the Psychiatric Genomics Consortium for sharing their GWAS summary statistics.

Additional Information

Additional file: Supplemental tables

Table S1: The susceptibility genes of schizophrenia identified by MCGA_Dist.

Table S2: The results of GO enrichment analysis based on the susceptibility genes of schizophrenia identified by MCGA_Dist.

Table S3: The PubMed search hits of the susceptibility genes of schizophrenia identified by MCGA_Dist.

Table S4: The susceptibility genes of schizophrenia identified by MCGA_eQTL.

946 Table S5: The susceptibility genes of schizophrenia identified by MCGA_isoQTL.

947 Table S6: The results of GO enrichment analysis based on combined eGene set and combined
948 sGene set of schizophrenia generated by MCGA_eQTL and MCGA_isoQTL.

949 Table S7: The PubMed search hits of the susceptibility genes of schizophrenia identified by
950 MCGA_eQTL.

951 Table S8: The PubMed search hits of the susceptibility genes of schizophrenia identified by
952 MCGA_isoQTL.

953 Table S9: The potential susceptibility isoforms of the twenty-three common genes in the top five
954 phenotype-associated tissues of schizophrenia.

955 Table S10: The FDA-approved antipsychotics in DGIdb.

956 Table S11: Drug-gene interaction terms in DGIdb for susceptible genes exclusively predicted
957 MCGA_eQTL.

958 Table S12: Drug-gene interaction terms in DGIdb for susceptible genes exclusively predicted
959 MCGA_isoQTL.

960 Table S13: The number of potentially druggable categories for the susceptibility genes of
961 schizophrenia identified by MCGA.

962 Table S14: The susceptibility genes of schizophrenia identified by MAGMA.

963

964 References

965

966 1. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J.
967 (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. American journal
968 of human genetics *101*, 5-22. DOI: 10.1016/j.ajhg.2017.06.005.
969 2. Gallagher, M.D., and Chen-Plotkin, A.S. (2018). The Post-GWAS Era: From Association to

Function. American journal of human genetics *102*, 717-730. DOI: 10.1016/j.ajhg.2018.04.002.

3. Tam, V., Patel, N., Turcotte, M., Bosse, Y., Pare, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. Nat Rev Genet *20*, 467-484. DOI: 10.1038/s41576-019-0127-1.

4. Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. Genome Res *22*, 1748-1759. DOI: 10.1101/gr.136127.111.

5. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A.K., Canfield, T.K., Giste, E., Diegel, M., Bates, D., Hansen, R.S., Neph, S., Sabo, P.J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S.R., Kaul, R., and Stamatoyannopoulos, J.A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190-1195. DOI: 10.1126/science.1222794.

6. Neale, B.M., and Sham, P.C. (2004). The future of association studies: gene-based analysis and replication. American journal of human genetics *75*, 353-362. DOI: 10.1086/423901.

7. de Leeuw, C.A., Mooij, J.M., Heskes, T., and Posthuma, D. (2015). MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol *11*, e1004219. DOI: 10.1371/journal.pcbi.1004219.

8. Liu, J.Z., McRae, A.F., Nyholt, D.R., Medland, S.E., Wray, N.R., Brown, K.M., Investigators, A., Hayward, N.K., Montgomery, G.W., Visscher, P.M., Martin, N.G., and Macgregor, S. (2010). A versatile gene-based test for genome-wide association studies. American journal of human genetics *87*, 139-145. DOI: 10.1016/j.ajhg.2010.06.009.

9. Li, M.X., Gui, H.S., Kwan, J.S.H., and Sham, P.C. (2011). GATES: A Rapid and Powerful Gene-Based Association Test Using Extended Simes Procedure. American journal of human genetics *88*, 283-293. DOI: 10.1016/j.ajhg.2011.01.019.

10. Li, M., Jiang, L., Mak, T.S.H., Kwan, J.S.H., Xue, C., Chen, P., Leung, H.C., Cui, L., Li, T., and Sham, P.C. (2019). A powerful conditional gene-based association approach implicated functionally important genes for schizophrenia. Bioinformatics (Oxford, England) *35*, 628-635. DOI: 10.1093/bioinformatics/bty682.

11. Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., Weedon, M.N., and Loos, R.J. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nature genetics *44*, 369-375. DOI: 10.1038/ng.2213.

12. Hekselman, I., and Yeger-Lotem, E. (2020). Mechanisms of tissue and cell-type specificity in heritable traits and diseases. Nat Rev Genet *21*, 137-150. DOI: 10.1038/s41576-019-0200-9.

13. Gerring, Z.F., Gamazon, E.R., Derks, E.M., and Major Depressive Disorder Working Group of the Psychiatric Genomics, C. (2019). A gene co-expression network-based analysis of multiple brain tissues reveals novel genes and molecular pathways underlying major depression. PLoS Genet *15*, e1008245. DOI: 10.1371/journal.pgen.1008245.

14. Barbeira, A.N., Dickinson, S.P., Bonazzola, R., Zheng, J., Wheeler, H.E., Torres, J.M., Torstenson, E.S., Shah, K.P., Garcia, T., Edwards, T.L., Stahl, E.A., Huckins, L.M., Consortium, G.T., Nicolae, D.L., Cox, N.J., and Im, H.K. (2018). Exploring the phenotypic

consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nature communications *9*, 1825. DOI: 10.1038/s41467-018-03621-1.

15. Gamazon, E.R., Zwinderman, A.H., Cox, N.J., Denys, D., and Derks, E.M. (2019). Multi-tissue transcriptome analyses identify genetic mechanisms underlying neuropsychiatric traits. Nature genetics *51*, 933-940. DOI: 10.1038/s41588-019-0409-8.

16. Gamazon, E.R., Segre, A.V., van de Bunt, M., Wen, X.Q., Xi, H.S., Hormozdiari, F., Ongen, H., Konkashbaev, A., Derks, E.M., Aguet, F., Quan, J., Nicolae, L., Eskin, E., Kellis, M., Getz, G., McCarthy, M.I., Dermitzakis, E.T., Cox, N.J., Ardlie, K.G., and Consortium, G. (2018). Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. Nature Genetics *50*, 956-967. DOI: 10.1038/s41588-018-0154-4.

17. Huckins, L.M., Dobbyn, A., Ruderfer, D.M., Hoffman, G., Wang, W., Pardinas, A.F., Rajagopal, V.M., Als, T.D., Nguyen, H.T., and Girdhar, K. (2019). Gene expression imputation across multiple brain regions provides insights into schizophrenia risk. Nature genetics *51*, 659-674. DOI: 10.1038/s41588-019-0364-4.

18. Gandal, M.J., Zhang, P., Hadjimichael, E., Walker, R.L., Chen, C., Liu, S., Won, H., van Bakel, H., Varghese, M., Wang, Y., Shieh, A.W., Haney, J., Parhami, S., Belmont, J., Kim, M., Moran Losada, P., Khan, Z., Mleczko, J., Xia, Y., Dai, R., Wang, D., Yang, Y.T., Xu, M., Fish, K., Hof, P.R., Warrell, J., Fitzgerald, D., White, K., Jaffe, A.E., Psych, E.C., Peters, M.A., Gerstein, M., Liu, C., Iakoucheva, L.M., Pinto, D., and Geschwind, D.H. (2018). Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. Science *362*. DOI: 10.1126/science.aat8127.

19. Fan, C., Chen, K., Zhou, J., Wong, P.-p., He, D., Huang, Y., Wang, X., Ling, T., Yang, Y., and Zhao, H. (2021). Systematic analysis to identify transcriptome-wide dysregulation of Alzheimer's disease in genes and isoforms. Human Genetics *140*, 609-623. DOI: 10.1007/s00439-020-02230-7.

20. Ripke, S., Walters, J.T., O'Donovan, M.C., and Consortium, S.W.G.o.t.P.G. (2020). Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. MedRxiv. DOI: 10.1101/2020.09.12.20192922.

21. Consortium, G. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318-1330. DOI: 10.1126/science.aaz1776.

22. Jiang, L., Xue, C., Dai, S., Chen, S., Chen, P., Sham, P.C., Wang, H., and Li, M. (2019). DESE: estimating driver tissues by selective expression of genes associated with complex diseases or traits. Genome Biol *20*, 233. DOI: 10.1186/s13059-019-1801-5.

23. Bakhtiarizadeh, M.R., Hosseinpour, B., Shahhoseini, M., Korte, A., and Gifani, P. (2018). Weighted Gene Co-expression Network Analysis of Endometriosis and Identification of Functional Modules Associated With Its Main Hallmarks. Front Genet *9*, 453. DOI: 10.3389/fgene.2018.00453.

24. King, E.A., Davis, J.W., and Degner, J.F. (2019). Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. PLoS genetics *15*, e1008489. DOI: 10.1371/journal.pgen.1008489.

25. Nelson, M.R., Tipney, H., Painter, J.L., Shen, J., Nicoletti, P., Shen, Y., Floratos, A., Sham, P.C., Li, M.J., Wang, J., Cardon, L.R., Whittaker, J.C., and Sanseau, P. (2015). The support of human genetic evidence for approved drug indications. Nat Genet *47*, 856-860. DOI:

10.1038/ng.3314.

26. Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., Pon, A., Knox, C., and Wilson, M. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic acids research *46*, D1074-D1082. DOI: 10.1093/nar/gkx1037.

27. Freshour, S.L., Kiwala, S., Cotto, K.C., Coffman, A.C., McMichael, J.F., Song, J.J., Griffith, M., Griffith, O.L., and Wagner, A.H. (2021). Integration of the Drug–Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts. Nucleic acids research *49*, D1144-D1151. DOI: 10.1093/nar/gkaa1084.

28. Legge, S.E., Santoro, M.L., Periyasamy, S., Okewole, A., Arsalan, A., and Kowalec, K. (2021). Genetic architecture of schizophrenia: a review of major advancements. Psychological medicine, 1-10. DOI: 10.1017/S0033291720005334.

29. Franc, V., Hlaváč, V., and Navara, M. (2005). Sequential coordinate-wise algorithm for the non-negative least squares problem. (Springer), pp. 407-414.

30. Tokheim, C.J., Papadopoulos, N., Kinzler, K.W., Vogelstein, B., and Karchin, R. (2016). Evaluating the evaluation of cancer driver genes. Proceedings of the National Academy of Sciences of the United States of America *113*, 14330-14335. DOI: 10.1073/pnas.1616440113.

31. Montana, G. (2005). HapSim: a simulation tool for generating haplotype data with pre-specified allele frequencies and LD coefficients. Bioinformatics *21*, 4309-4311. DOI: 10.1093/bioinformatics/bti689.

32. Kung, A.W., Xiao, S.M., Cherny, S., Li, G.H., Gao, Y., Tso, G., Lau, K.S., Luk, K.D., Liu, J.M., Cui, B., Zhang, M.J., Zhang, Z.L., He, J.W., Yue, H., Xia, W.B., Luo, L.M., He, S.L., Kiel, D.P., Karasik, D., Hsu, Y.H., Cupples, L.A., Demissie, S., Styrkarsdottir, U., Halldorsson, B.V., Sigurdsson, G., Thorsteinsdottir, U., Stefansson, K., Richards, J.B., Zhai, G., Soranzo, N., Valdes, A., Spector, T.D., and Sham, P.C. (2010). Association of JAG1 with bone mineral density and osteoporotic fractures: a genome-wide association study and follow-up replication studies. American journal of human genetics *86*, 229-239. DOI: 10.1016/j.ajhg.2009.12.014.

33. Consortium, G.P. (2015). A global reference for human genetic variation. Nature *526*, 68. DOI: 10.1038/nature15393.

34. Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics, C., Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet *47*, 291-295. DOI: 10.1038/ng.3211.

35. Gillett, A.C., Vassos, E., and Lewis, C.M. (2018). Transforming Summary Statistics from Logistic Regression to the Liability Scale: Application to Genetic and Environmental Risk Scores. Hum Hered *83*, 210-224. DOI: 10.1159/000495697.

36. Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic acids research *47*, W191-W198. DOI: 10.1093/nar/gkz369.

37. Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics (Oxford, England) *24*, 719-720. DOI: 10.1093/bioinformatics/btm563.