HHS Public Access

Author manuscript

Hum Genet. Author manuscript; available in PMC 2021 February 01.

Published in final edited form as:

Hum Genet. 2020 February ; 139(2): 199–213. doi:10.1007/s00439-019-02098-2.

A powerful fine-mapping method for transcriptome-wide association studies

Chong Wu[1,*], **Wei Pan**[2,*]

[1]Department of Statistics, Florida State University, Tallahassee, FL, USA

[2]Division of Biostatistics, University of Minnesota, Minneapolis, MN, USA

Abstract

Transcriptome-wide association studies (TWAS) have been recently applied to successfully identify many novel genes associated with complex traits. While appealing, TWAS tend to identify multiple significant genes per locus and many of them may not be causal due to confounding through linkage disequilibrium (LD) among SNPs. Here we introduce a powerful fine-mapping method that prioritizes putative causal genes by accounting for local LD. We apply a weighted adaptive test with eQTL-derived weights to maintain high power across various scenarios. Through simulations, we show that our new approach yielded a well-controlled Type I error rate while achieving higher power and AUC than competing methods. We applied our approach to a schizophrenia GWAS summary dataset and successfully prioritized some well-known schizophrenia-related genes, such as *C4A*. Importantly, our approach identified some putative causal genes (e.g., *B3GAT1* and *RGS6*) that were missed by competing methods and TWAS. Our results suggest that our approach is a useful tool to prioritize putative causal genes, gaining insights into the mechanisms of complex traits.

[*]Correspondence: chongwu@stat.fsu.edu & weip@biostat.umn.edu.

Author's contributions

WP conceived the study. CW and WP developed the methods. CW performed the analysis and drafted the manuscript. WP supervised the study. All authors approved/edited the final manuscript.

Availability of data and materials

We make our software FOGS publicly available on a GitHub repository: https://github.com/ChongWu-Biostat/FOGS. TWAS software and eQTL derived weights can be downloaded at http://gusevlab.org/projects/fusion/. The software FOCUS can be obtained at https://github.com/bogdanlab/focus/. The Schizophrenia GWAS summary data [27] can be downloaded at http://walters.psycm.cf.ac.uk, and the Lung Health Study data can be obtained at dbGAP (phs000335.v2.p2). The data used in this paper can be downloaded at https://figshare.com/articles/FOGS_data/7636691.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Introduction

Transcriptome-wide association studies (TWAS) [2, 3, 14, 16, 44, 46] have garnered substantial interest and become increasingly popular within the human genetics community. TWAS integrate one or more external eQTL reference panels (including both genome-wide gene expression and genotype data) with a main GWAS to discover gene-trait associations. Although TWAS [2, 14, 16, 44] have been promising in uncovering many novel genes as possible mediators between a trait and genetic variants, it should be emphasized that the significant genes identified by TWAS are not necessarily causal [24, 39, 41]. As demonstrated by others [24, 39] and here, TWAS frequently identify multiple significant genes per locus (Figure 1b), many of which may not be causal due to confounding through linkage disequilibrium (LD) among SNPs. These phenomena are similar to single variant analysis in GWAS: the association between a gene (or a SNP) and a trait can be indirect, resulting from its expression correlations (or LD) with a nearby and truly causal gene (or SNP) [32]. It can be challenging to determine the underlying causal genes, and this is when fine-mapping helps. Fine-mapping seeks to determine genetic variants (or genes) causal for complex traits, given evidence of an association in a genomic region [32, 35]. Even though many fine-mapping methods have been proposed to deal with a single-variant association, few methods have been proposed to prioritize causal genes, especially for TWAS results.

Here, we propose a method called FOGS (Fine-mapping Of Gene Sets) to perform statistical fine-mapping over gene sets in a locus by testing whether the conditional effects of SNPs in each gene is null (Figure 1a). The conditional effect is estimated from a joint regression model including all the SNPs in a locus. First, we estimate the conditional effects for the SNPs via ridge regression. Since high LD (or correlations) among the SNPs can make standard regression estimation unstable, we add a small ridge penalty to regularize the estimation. Next, we apply an adaptive test, called aSPU [26], to combine the conditional effects and eQTL-derived weights effectively. The idea is that since we do not know how many SNPs are causal, we construct a class of sum of powered score (SPU) tests such that hopefully at least one of them would be powerful for a given situation. Then aSPU selects the most significant testing result data-adaptively with a proper adjustment for multiple testing.

By extensive simulations, we show that FOGS adequately controls Type I error rates under various null scenarios, and outperforms other competing methods. Specifically, compared to FOCUS (a Bayesian fine-mapping method [24]) and p-value ranking of TWAS results, FOGS achieves a higher AUC, identifies more causal genes at the same false positive rate, and yields a smaller number of false positives at the same true positive rate. Finally, we applied FOGS to a large schizophrenia GWAS summary dataset of 105,318 European individuals [27]. We found that FOGS prioritizes some causal genes with established roles in schizophrenia (e.g., *C4A* [33]). In addition, FOGS identifies some putative causal genes with literature support, which however would be missed by FOCUS and TWAS (e.g., *B3GAT1* [15, 43] and *RGS6* [9, 15, 22, 31]). Overall, our results showcase the power of FOGS to prioritize putative causal genes and to better understand the genetic basis of complex diseases.

Results

Overview of FOGS method

To prioritize causal genes at a TWAS significant region, we test on the *conditional effect* of each SNP in a gene after adjusting for other genes in the same locus (Figure 1a). Specifically, we model the phenotype **y** as

$$\mathbf{y} = \mathbf{V}\alpha + \mathbf{X}\beta + \epsilon,$$

where **X** and **V** are genotype matrices for gene A (of interest) and other genes in the same region, respectively (see Methods). We are interested in testing if gene A is putatively causal with a null hypothesis $H_0 : \beta = 0$. Although we refer to genes with non-zero effects ($\beta \neq 0$) as putatively causal, it is important to realize that statistical fine-mapping alone is unable to completely determine causality.

First, we construct the conditional Z score for **X** via a newly proposed summary statistics-based ridge regression (see Methods). Here, we apply ridge regression with a small penalty to avoid collinearity induced by high LD among SNPs. To account for genes without eQTL-derived weights in the relevant tissue, we include eQTL-derived weights from proxy tissues for such genes. This approximation is supported by the fact that eQTLs are largely shared across tissues [10] as also adopted in FOCUS [24]. Next, different tests will be powerful under different scenarios. For example, the SUM test will be most powerful when the conditional effect sizes of all the SNPs are the same and in the same directions. Since we do not know how many SNPs are causal and the true pattern of their effect sizes and directions, we apply an adaptive test, called weighted aSPU [26], to combine the *conditional effects* and eQTL-derived weights efficiently. Specifically, we first construct a class of sum of powered score (SPU) tests such that hopefully at least one of them would be powerful for a given situation. Then aSPU selects the most significant testing result data-adaptively with a proper adjustment for multiple testing. In the end, we use *one layer of Monte Carlo simulations* to estimate the p-values for both SPU and aSPU tests *simultaneously* and efficiently [26].

FOGS prioritizes and improves resolution for fine-mapping causal genes

We simulated GWAS summary data from the real genotype data drawn from the Lung Health Study (dbGAP: phs000335.v2.p2), and used the corresponding gene definitions. To mimic real GWAS summary data, we generated the phenotype and then ran a simple regression for each SNP to obtain its estimated effect size and variance. To obtain robust and unbiased results, we varied the number of causal SNPs, causal effect sizes, and relationship between causal SNPs and eQTL-derived weights (See Methods).

First, we confirmed that some non-causal genes were highly significant and yielded inflated Type I error rates under various scenarios, and fine-mapping was needed to prioritize causal genes (Supplementary Figure S1). This phenomenon was previously reported by others [24, 39].

Next, we simulated GWAS summary data with two causal SNPs in one causal gene (see Methods) and assessed the performance of FOGS. For comparison, we considered a

Bayesian fine-mapping method (FOCUS, as described in Mancuso et al. [24]) and the *p*-value ranking of TWAS results (denoted as TWAS ranking for simplicity, see Methods). Figure 2 shows the performance of different methods under different criteria. First, an AUC measures the whole area under a ROC curve and has been widely used for performance comparison, while the partial AUC (pAUC) is most useful when only certain regions of the ROC (e.g., high specificity) are of particular interest [12]. We calculated pAUC with specificity in the ranges of (0.95, 1) and of (0.9, 1), denoted pAUC1 and pAUC2 respectively. We further applied the Delong test [11] to compare AUCs, and the bootstrap with 10^4 bootstrap replications to compare pAUCs. Compared to TWAS and FOCUS FOGS achieved a significantly higher AUC (85.8%), pAUC1 (60.4%) and pAUC2 (65.9%) (Supplementary Table S1).

Second, both FOGS and FOCUS significantly reduced the number of false positives (FOGS: 0.33; FOCUS: 0.27; TWAS: 1.97) per locus compared to TWAS, and yielded lower numbers of true positives (FOGS: 0.39; FOCUS: 0.36; TWAS: 0.53), making it hard to compare different methods directly (Supplementary Table S1). To facilitate a fair comparison, we compared different methods under either the same false positive rates or the same true positive rate (Figures 2b and 2c). For example, at a false positive rate of 10%, the true positive rate of FOGS, FOCUS and TWAS were 69.8%, 51.5%, and 61.2%, respectively. Under various true positive rates, FOGS significantly reduced the number of false positives compared to both FOCUS and TWAS. For example, at a true positive rate of 60%, the false positive rates of FOGS, FOCUS and TWAS were 5.1%, 9.3%, and 15.1% respectively.

We considered several additional settings, including varying the number of causal SNPs (Supplementary Figures S2-S3), and the relationship between causal SNPs and eQTL constructed weights (Supplementary Figure S4). Importantly, we further considered some settings where all eSNPs in the causal gene were causal SNPs (Supplementary Figures S5-S6). Across all the above simulation settings, FOGS had consistent advantages over TWAS and FOCUS, yielding higher AUC and partial AUC, higher true positive rates and lower false positive rates under a given false/true positive rate. Interestingly, under the challenging settings such as when one casual SNP regulated at least two genes in the locus, FOGS performed either similalry to or slightly better than FOCUS (Supplementary Figure S7). Overall, FOGS performed well in prioritizing causal genes across all the settings considered; more studies are needed for other challenging settings.

FOGS is robust to the choice of penalty parameter λ

Figure 3 compares the performance of FOGS with various values of the ridge penalty parameter λ (see Methods) on the simulated data (two causal SNPs in one causal gene per locus). λ = 0 stands for the standard conditional model without penalization, which yielded similar or even slightly worse performance when compared to TWAS, and much worse performance compared to FOGS with a small penalty (λ = 0.1). For example, under a false positive rate of 5%, the power of FOGS with λ = 0 was 38.7%, much lower than the power of FOGS with λ = 0.1 (58.3%). At the true positive rate of 40%, FOGS with λ = 0 identified 179.1% more false positive genes than FOGS with λ = 0.1. These results

confirmed the importance of adding a small ridge penalty to stabilize the estimation in the presence of collinearity induced by LD among SNPs.

Importantly, FOGS was robust to the choice of λ as long as a reasonable λ value was used. The AUCs of FOGS with λ = 0.1, λ = 0.15, and λ = 0.2 were 85.2%, 85.6%, and 85.8%, respectively (Figure 3b). Under false positive rate at 5%, the power of FOGS with λ = 0.1, λ = 0.15, and λ = 0.2 were 58.3%, 57.1%, and 57.4%, respectively (Figure 3b). FOGS with λ = 0.1, λ = 0.15, and λ = 0.2 had similar false positive rates at various true positive rates (Figure 2c). Furthermore, we compared the results of FOGS with different λ values under different sample sizes of the reference panel (Supplementary Figures S8-S9). We showed that FOGS was robust to the choice of λ with different reference panels, though FOGS with a slightly larger λ (i.e., λ = 0.2) performed slightly better than that with the default setting λ = 0.1 when the sample size was small. In summary, we recommend using the default λ = 0.1 for FOGS in real data applications, partly because it has been widely used by others [16, 28].

Robustness analysis of FOGS

First, under the null (no SNP-trait association for all SNPs in the locus), the type I error rates of FOGS were satisfactorily controlled (Figure 4a) under different α levels. Perhaps due to the small bias induced by the ridge penalty, FOGS yielded a slightly conservative Type I error rate, even though the 95% confidence interval covered the truth most time under various nominal levels α.

Next, we simulated data as in the previous sections, randomly selecting two SNPs in one gene to be causal with effect size $c = 0.1$. We pruned out the causal SNPs to evaluate the robustness of FOGS. Note that FOGS estimated a conditional effect for each SNP by conditioning on all the SNPs from any of other genes in the locus to prioritize putative causal genes accordingly. Since some SNPs might be highly correlated with causal SNPs, the effect from causal SNPs can be indirectly adjusted. This property of FOGS explained why applying FOGS to the masked causal SNP data yielded a similar number of false positives as that of applying FOGS to the complete data (0.34 vs. 0.33; Figure 4b). FOCUS yielded similar numbers of false positives for both masked causal SNP data and complete data (0.27 vs. 0.26; Supplementary Figure S10). As expected, because the causal SNPs was pruned out before the analysis, the number of true positives for masked causal SNP data was smaller than that with complete data (0.28 vs. 0.39; Figure 4b). Interestingly, FOCUS was robust, though yielding a smaller number of true positives when analyzing masked causal SNP data (0.29 vs. 0.36 (for complete data); Supplementary Figure S10).

We further considered the situation where all the SNPs in the causal gene were pruned out before analysis. Since the whole causal gene had been removed, fewer SNPs remained highly associated with the causal SNPs and the effects of the causal SNPs might not be fully captured, leading to a much higher number of false positives. It is observed that applying FOGS to the masked causal gene data increased the false positives per locus from 0.33 to 0.41 (by 124%, Figure 4c). In comparison, FOCUS increased the false positives per locus from 0.27 to 0.46 (by 170%, Supplementary Figure S10). Importantly, a simple ranking of the TWAS results yielded much higher false positives per locus (Supplementary Figure S11).

In summary, we found that FOGS incurred some performance loss in the challenging settings when some causal SNPs were missing in the data. Importantly, the competing methods such as FOCUS suffered similarly, and we leave this interesting topic for future study.

Application to a schizophrenia GWAS summary dataset

Schizophrenia is a psychiatric disorder affecting about 0.7% of adults worldwide [25]. Although more than 100 risk loci have been identified [27, 31], schizophrenia, is truly complex; many risk regions contained more than one significantly associated gene, and it is hard to determine which gene is causal. This is when fine-mapping helps. Having validated our new fine-mapping method FOGS in simulations, we re-analyzed a large schizophrenia GWAS summary dataset [27] to provide insights into the genetic architecture of schizophrenia. As others [24], we defined the loci as in LDetect [6]. We conducted genome-wide analysis and then focused on the 57 loci containing at least one genome-wide significant SNP and at least one TWAS significant gene (see Methods). Specifically, we applied TWAS first and then applied fine-mapping methods to prioritize putative causal genes. For TWAS and later fine-mapping, we used the CMC eQTL weights when available; otherwise, we included the eQTL weights with the best accuracy across all other tissues, leading to 15,460 genes with eQTL weights (see Methods).

First, TWAS identified 203 significant genes at 71 independent loci after stringent Bonferroni correction ($0.05/16{,}000 \approx 3.1 \times 10^{-6}$; Figure 5a and Supplementary Table S2). Of these, 23 gene-schizophrenia associations did not overlap a genome-wide significant SNP, residing in 14 independent loci. However, many identified TWAS significant genes were located in the same locus (Supplementary Figure S12). 29 out of 71 loci contained more than one gene, and 11 loci contained more than five significant genes. For example, for one locus on chromosome 6 (31.6–32.7 Mb), it contained 42 genes, 19 of which have been identified as significant by TWAS (1b). Importantly, seven genes were highly significant (p-value $< 1 \times 10^{-17}$), and simply ranking the TWAS results was unlikely to work for this locus; more sophisticated fine-mapping methods were needed.

We applied both FOGS and FOCUS to all the loci with more than two genes. For FOGS, we used the same Bonferroni cutoff ($0.05/16{,}000 \approx 3.1 \times 10^{-6}$) as used in TWAS. For FOCUS, we used the default 90%-credible gene-sets to prioritize causal genes (see Methods). First, FOGS is quite computationally efficient: it took about 4.1 seconds on average to compute the p-value for a gene by using one core in a Minnesota Supercomputing Institute (MSI) server, and for the most time-consuming gene, it took about 12.0 minutes. Second, Figure 5c compares the different methods for prioritizing putative causal genes. Since in some loci TWAS identified more than two genes with highly significant results (unable to distinguish which one was more significant), we selected the most significant gene in each locus plus 19 genes with p-value $< 1 \times 10^{-17}$ as the results of the p-value ranking of TWAS statistics, denoted as TWAS ranking. FOCUS identified 52 putative causal genes (Supplementary Table S3), 48 (92.3%) of which have been identified by TWAS ranking. In other words, for the GWAS schizophrenia summary data analyzed here, there was not much difference between TWAS ranking and FOCUS. In contrast, FOGS identified 46 putative causal genes

(Figure 5b, Supplementary Table S4), 29 (63.0%) of which would be missed by the TWAS ranking. Compared to FOCUS, FOGS narrowed down the candidates of causal genes to a shorter list, which was very different from the results of TWAS ranking.

Since the majority of the significant genes (180 out of 203; 89%) were located in the 57 GWAS risk loci that contained at least one genome-wide significant SNPs, we discussed these loci in detail. Table 1 shows the putative causal genes identified by FOGS in these 57 loci. Interestingly, FOGS identified some putative causal genes (e.g., *RGS6*, p-value $< 1.0 \times 10^{-7}$) that were missed by both TWAS and FOCUS. Several GWAS have indicated that some genome-wide schizophrenia significantly associated SNPs in *RGS6* (rs35607894, p-value $= 5 \times 10^{-11}$ [22]; rs2332700, p-value $= 1 \times 10^{-9}$ [15]; p-value $= 5 \times 10^{-9}$ [31]). The possible role of *RGS6* in schizophrenia is further supported by the fact that *RGS6* is required for rapid deactivation of GABA-B receptor in the cerebellum and thus may play a role in schizophrenia pathogenesis [9]. Another example is *B3GAT1* (FOGS p-value $< 1 \times 10^{-7}$). Several studies have indicated that *B3GAT1* contained a significant SNP (rs893949, p-value $= 5 \times 10^{-7}$ [15]) associated with schizophrenia and rare *de novo* copy number mutations in *B3GAT1* contribute to the genetic component of schizophrenia [43]. As a positive control, we examined one locus on Chromosome 6 (Figures 1b and 1c), as this locus exhibited the strongest genetic signal while containing a well known putative causal gene, *C4A* [33]. Both FOCUS and FOGS identified the gene *C4A* (FOCUS p-value $< 1 \times 10^{-7}$ and FOCUS posterior probability > 0.99), partially validating the usefulness of both methods. One caveat is that *C4A* resides in the MHC/HLA region with complicated and pervasive LD structures, and one should pay special attention when applying FOGS to this region. Finally, FOGS identified some putative causal genes (e.g., *FAM114A2* and *PPP1R18*) without much literature support. These new findings suggest that further study of these putative causal genes is needed and may help us gain new insights into the genetic basis of schizophrenia.

In the end, we emphasized the importance of using an adaptive test like aSPU to aggregate information across conditional Z-scores of multiple SNPs. Because there is no uniformly most powerful test, any non-adaptive test may not be powerful for a given situation and thus fail to identify putative causal genes. For example, if we applied the popular weighted SUM test for FOGS, we identified eight putative causal genes, all of which have been identified by FOGS with aSPU. FOGS with the weighted SSU test identified 19 putative causal genes, 18 of which have been identified by FOGS with aSPU. In summary, by applying aSPU to select the most significant testing result data-adaptively, FOGS maintained high power to detect putative causal genes.

Taken together, these results strongly support the potential usefulness of FOGS to identify putative causal genes that could be missed by other competing fine-mapping methods like FOCUS and TWAS. The schizophrenia-related genes identified by FOGS may improve our understanding of the genetic mechanisms of schizophrenia.

Discussion

In this work, we have presented a new fine-mapping method called FOGS to prioritize putative causal genes for TWAS. As demonstrated before [24, 39] and further here, TWAS

frequently identify multiple significant genes per locus, many of which may not be causal due to confounding through LD among SNPs; in these situations, fine-mapping using FOGS would help. We have demonstrated that FOGS has several potential advantages through extensive simulations and real data analyses. First, by adjusting for the effects of other genes in the same locus, FOGS adequately controls Type I error rates under various null scenarios. Second, by data-adaptively aggregating information among multiple SNPs and eQTL-derived weights, FOGS outperforms the *p*-value ranking of TWAS results and FOCUS [24], a Bayesian fine-mapping method, in identifying causal genes with reduced numbers of false positives. Third, FOGS is robust to the choice of the ridge penalty parameter λ, and we recommend using the default value $\lambda = 0.1$ in real applications.

To better understand the genetic component of schizophrenia, we reanalyzed a large-scale schizophrenia GWAS summary dataset [27]. FOGS correctly identified a positive control gene *C4A*. Interestingly, FOGS identified some putative causal genes such as *RGS6* and *B3GAT1* that were missed by both TWAS and FOCUS, but might be related to schizophrenia as shown by other studies [9, 43].

Our proposed method FOGS has conceptual similarities with conditional and joint multiple SNP analysis [47] that aims to identify secondary associated SNPs in a locus. Our method extends the standard regression model to ridge regression to account for possibly strong LD among SNPs, then aggregates conditional effect sizes of multiple SNPs (in the same gene) efficiently. Using ridge regression is important for FOGS as applying standard regression model leads to the results similar to or even slightly worse than the *p*-value ranking of TWAS results as shown in our simulations. Our work was motivated by FOCUS [24] that prioritizes causal genes by adjusting the effects of the imputed gene expression levels of other genes in the same locus. In contrast, our approach adjusts the effects at the SNP level since a causal SNP might mediate its effect on the trait through pathways other than cis-transcriptional regulation.

We list some limitations of FOGS. First, to account for LD among SNPs from other genes in the locus, we adopt ridge regression. However, the ridge penalty introduces some small bias, leading to slightly conservative inference with potential power loss. It might be interesting to apply some non-convex penalties, such as MCP [48] and truncated Lasso penalty [34], to reduce the bias. Importantly, ridge regression can be viewed as a Bayesian method with independent normal priors on the components of $\tilde{\beta}_j$. Zhu et al. [51] introduced a "Regression with Summary Statistics" likelihood with promising performance with a spike-and-slab type prior [51]. Adopting the Bayesian framework may yield even better performance. Second, like other GWAS summary data based methods, FOGS is dependent on the use of a reference panel. Using a small reference panel such as the 1000 Genomes Project Phase 3, or a panel from a different population, might introduce some biases [5, 47]. In real data analysis, we used the Lung Health Study data with about 4,100 individuals as the reference panel. We might achieve more robust results and identify more putative causal genes if a larger reference panel (e.g., UK Biobank data) is applied. Third, we only focus on the European ancestry here. While we expect that FOGS can be directly applied to non-European GWAS summary data (with a corresponding reference), one caveat is that

applying FOGS with an European reference panel to non-Europeans GWAS summary data may lead to suspicious results. Several fine-mapping methods [17, 20, 32] have been developed by taking advantage of the variability in the genetic backgrounds across populations; we expect FOGS may provide more robust results if trans-ancestry information can be incorporated. Fourth, like FOCUS, our proposed method is based on a reference eQTL dataset from a single tissue; as shown by others in the context of TWAS (but not TWAS-fine mapping), using multiple eQTL datasets drawn from multiple tissues to construct weights may gain power for TWAS by taking advantage of shared eQTL across multiple tissues [19]; our method can be similarly extended. Fifth, there has been continuing interest in TWAS with many new methods being developed, such as CoMM [46] and MultiXcan [3]. Incorporating into FOGS these new ideas and/or more informative eQTL-derived weights may further improve the power. We leave these interesting topics to future work.

Conclusions

In this study, we have developed a new fine-mapping method, FOGS, to prioritize putative causal genes by accounting for local linkage disequilibrium (LD). We have shown that this approach maintains correct Type I error rates and achieves higher power than the competing methods through extensive simulations. We have applied our method to a schizophrenia GWAS summary datasets of 105,318 European individuals and identified 46 putative causal genes, some of which were missed by the competing methods but might be related to schizophrenia as shown by other studies.

Methods

New fine-mapping method: FOGS

Fine-mapping Of Gene Sets (FOGS) prioritizes causal genes by testing for the conditional association of each gene in a locus. Here, we focus on one gene, say gene A, and repeat the same procedure for every gene in the locus. The conditional effect is estimated from a joint regression model, in which all the SNPs in the locus are analyzed jointly. Specifically, we model a quantitative phenotype for n unrelated individuals, $\mathbf{y}$, as

$$\mathbf{y} = \mathbf{V}\alpha + \mathbf{X}\beta + \epsilon, \quad (1)$$

where $\mathbf{y} = \{y_i\}$ is a centered $n \times 1$ vector of phenotypes, $\mathbf{X} = \{x_{ij}\}$ is a centered (with mean 0) $n \times p$ genotype matrix at p SNPs with non-zero eQTL-derived weights for gene A (of interest) and $\mathbf{V} = \{v_{ij}\}$ is a centered $n \times q$ genotype matrix at q SNPs with non-zero eQTL weights for any of other genes in the same locus, and β and α are the joint effects for gene A (of interest) and other genes, respectively. To account for genes without eQTL-derived weights in the relevant tissue, we include eQTL weights from a proxy tissue with the best prediction accuracy (i.e., out of sample R^2). This approximation is supported by the fact that eQTLs are largely shared across tissues [10] as also adopted in FOCUS [24].

We test the causal effect between the phenotype y and gene A with the null hypothesis $H_0 : \beta = 0$. Testing $H_0 : \beta = 0$ is similar to fine-mapping of single variants [47]: instead of

prioritizing a secondary SNP, here we focus on prioritizing a causal gene with multiple SNPs in the locus. To test H_0, we need to address two technical questions: 1. how to estimate the conditional Z score $(\beta_j / \sqrt{\text{var}(\beta_j)})$ for each SNP in gene A? 2. how to construct an adaptive and powerful test to aggregate conditional Z scores and eQTL-derived weights efficiently? In the following, we will discuss our solutions for these two questions accordingly.

Estimating the conditional *Z* score via ridge regression.

For simplicity, we denote the conditional Z score for gene A as $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_p)'$. To have a better estimation of conditional Z score, we estimate Z_j $(j = 1, \ldots, p)$ for each SNP j in gene A separately. Under $H_0 : \beta = 0$, the effect of SNPs in gene A is zero and thus no need to adjust for them while estimating the conditional score Z_j for SNP j. For SNP j, our updated model is $\mathbf{y} = \mathbf{V}\alpha + X_j\beta_j + \epsilon$. For simplicity, we denote $\widetilde{\mathbf{X}} = (X_j, \mathbf{V})$ and $\widetilde{\beta} = (\beta_j, \alpha)$ and rewrite our model as $\mathbf{y} = \widetilde{\mathbf{X}}\widetilde{\beta} + \epsilon$. Due to LD, some SNPs in $\widetilde{\mathbf{X}}$ are highly correlated. To alleviate the multicollinearity problem and get a stable estimate, we add a small ridge penalty to the objective function as

$$\text{minimize } (\mathbf{y} - \widetilde{\mathbf{X}}\widetilde{\beta})'(\mathbf{y} - \widetilde{\mathbf{X}}\widetilde{\beta}) + \lambda\|\widetilde{\beta}\|_2^2. \tag{2}$$

By ridge regression, we can estimate the joint effects of multiple SNPs as

$$\hat{\beta} = (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} + \lambda\mathbf{I}_{p*})^{-1}\widetilde{\mathbf{X}}\mathbf{y} \text{ and } \text{var}(\hat{\beta}) = \sigma_J^2(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} + \lambda\mathbf{I}_{p*})^{-1}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} + \lambda\mathbf{I}_{p*})^{-1}, \tag{3}$$

where σ_J^2 is the residual variance in the joint analysis, λ is the penalty parameter controlling the extent of ridge regularization, and $\mathbf{I}_{p*}$ is the $(q + 1) \times (q + 1)$ identify matrix. When GWAS individual-level data is available, we can use the above formulas to estimate conditional Z score for SNP j as $Z_j = \hat{\beta}_1 / \sqrt{\text{var}(\hat{\beta}_1)}$; however, we often have only GWAS summary statistics resulting from single-SNP models as follows

$$\mathbf{y} = \widetilde{X}_j b_j + \epsilon, \tag{4}$$

where $\widetilde{X}_j$ is the jth column of $\widetilde{\mathbf{X}}$ for $j = 1, 2, \ldots, (q + 1)$. Inspired by Yang et al. [47], we estimate the conditional Z score by converting marginal effects (4) to joint effects (2) without using individual-level phenotype data.

The marginal effects of multiple SNPs $\widetilde{\mathbf{X}}$ estimated from single-SNP models can be written in matrix form as [47]

$$\hat{\mathbf{b}} = \mathbf{D}^{-1}\widetilde{\mathbf{X}}'\mathbf{y} \text{ and } \text{var}(\hat{\mathbf{b}}) = \sigma_M^2\mathbf{D}^{-1},$$

where $\mathbf{b} = \{b_j\}$ is a $(q + 1) \times 1$ vector of marginal SNP effects, $\mathbf{D} = \{D_j\}$ is a diagonal matrix with $D_j = \sum_{i=1}^{n} \tilde{x}_{ij}^2$, and σ_M^2 is a $(q + 1) \times 1$ vector of residual variances in the single-SNP analyses. By $\widetilde{\mathbf{X}}'\mathbf{y} = \mathbf{D}\hat{\mathbf{b}}$, we can rewrite equation (3) as

$$\hat{\beta} = (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} + \lambda \mathbf{I}_{p*})^{-1}\mathbf{D}\hat{\mathbf{b}} \text{ and } \text{var}(\hat{\beta}) = \sigma_J^2(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} + \lambda \mathbf{I}_{p*})^{-1}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} + \lambda \mathbf{I}_{p*})^{-1}. \quad (5)$$

We estimate the residual variance as

$$\hat{\sigma}_J^2 = \frac{\text{SSE}}{n - df} = \frac{\mathbf{y}'\mathbf{y} - \hat{\beta}'\mathbf{D}\hat{\mathbf{b}}}{n - df}.$$

SSE is the Sum of Squared Errors and df is the degrees of freedom, which can be estimated as $df = \sum_{j=1}^{q} d_j / (d_j + \lambda)$, where d_j are the eigenvalues of $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$. By Yang et al. [47], $\mathbf{y}'\mathbf{y} = (n-1)D_j S_j^2 + D_j \hat{b}_j^2$, where $S_j^2 = \hat{\sigma}_{M(j)}^2 / D_j$ is the squared standard error of $\hat{b}_j$. Then we take the median of $(n-1)D_j S_j^2 + D_j \hat{b}_j^2$ across all the SNPs [47] to estimate $\mathbf{y}'\mathbf{y}$. In the end, we use equation (5) to estimate the conditional Z score $Z_j = \hat{\beta}_1 / \sqrt{\text{var}(\hat{\beta}_1)}$ for SNP j. We repeat the procedure for all the SNPs in gene A and get the conditional Z score vector $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_p)'$.

For GWAS summary statistics, we are usually unable to obtain pooled individual-level genotype data and thus $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$ is unavailable. Following [16, 47], we use a reference sample, such as the Lung Health Study data (dbGAP: phs000335.v2.p2), to approximate $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$.

Note that the scale of measurement of a quantitative trait $\mathbf{Y}$ has no impact on the method derived above and thus some constant can be dropped from the equations. As in the standard practice, we rescaled $\widetilde{\mathbf{X}}$ to make $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$ the sample covariance matrix of $\widetilde{\mathbf{X}}$. Inspired by Pasaniuc et al. [28], we set the penalty parameter $\lambda = 0.1$ as the default, only adding a small ridge penalty to alleviate the collinearity problem. Of note, rescaling $\widetilde{\mathbf{X}}$ is important; otherwise the choice of penalty parameter λ will depend on the sample size of the reference panel.

Ideally, we may want to center and standardize/scale the genotype matrix $\widetilde{\mathbf{X}}$. However, because the GWAS summary statistics are often provided for unstandardized genotype data, we derived the model under the assumption that genotype matrices are only centered but not standardized. If we make the standardization assumption, we have to recalculate the corresponding marginal effect size for each SNP j as $\hat{b}_j = s_j(z_j / \sqrt{N_j})$, where s_j is the sign, z_j is the Z score, and N_j is the sample size for SNP j [38]. Since the sample sizes for the SNPs often vary and are unavailable, the approximation will not be accurate if the maximum or average sample size is used. Importantly, the main reason of adding a small ridge penalty is to stabilize the model and alleviate the collinearity problem. Thus, we prefer to work on the current setting and leave the study of whether to standardize genotype matrices to future work.

Furthermore, the same method described above for a quantitative trait can be applied to a binary trait, as long as the effect size $\mathbf{b}$ and its corresponding standard errors are measured on the log odds ratio (log(OR)) scale [47]. As shown by others [45, 49], a linear regression

model can be used to well approximate a logistic regression model for a binary trait due to small effect sizes of SNPs (and genes). Also, the log odds ratio scale effect size can be viewed as being an effect estimate for an underlying quantitative (continuous) liability [29, 47].

Aggregating conditional *Z* scores to prioritize causal gene

Once we get the conditional score Z for each SNP j in gene A, the next question is how to aggregate information from multiple SNPs to test $H_0 : \beta = 0$ efficiently. Because there is no uniformly most powerful test, any non-adaptive test may not be powerful for a given situation. Since we do not know the underlying truth, we construct a class of sum of powered score (SPU) tests such that hopefully at least one of them would be powerful for a given situation. Then we apply an adaptive test called aSPU [21, 26] to select the most significant testing result data-adaptively with a proper adjustment for multiple testing. To further improve statistical power, we apply the weighted aSPU with eQTL-derived weights.

First, we denote the weighted conditional score vector for β in (1) as

$$\mathbf{U} = (U_1, \ldots, U_p)' = \mathbf{WZ},$$

where $\mathbf{W} = \text{Diag}(\hat{w}_1, \ldots, \hat{w}_p)$ are the eQTL-derived weights and $\mathbf{Z}$ is the conditional Z score estimated from the previous subsection.

Second, we construct a class of SPU tests as follows:

$$T_{\text{SPU}(\gamma)} = \sum_{j=1}^{p} U_j^{\gamma},$$

where γ is a positive integer and we further define $T_{\text{SPU}(\infty)} = \max_j |U_j|$ since $T_{\text{SPU}(\gamma)} \propto (\sum_{j=1}^{p} |U_j|^{\gamma})^{1/\gamma} \to \max_j |U_j|$ as an even integer $\gamma \to \infty$. The SPU tests cover many popular tests as special cases. For example, SPU(1) equals to the SUM or burden test, which yields high power if all SNPs are causal with similar effect sizes and causal direction. Meanwhile, SPU(2) equals to the SSU, which would be powerful if the causal direction of SNPs are different. As γ increases, SPU(γ) will be more powerful for sparser signals.

Third, since the optimal value of γ is data-dependent and unknown, we propose using aSPU to combine the results from a class of SPU tests: the aSPU test statistic is defined as

$$T_{\text{aSPU}} = \min_{\gamma \in \Gamma} P_{\mathbf{SPU}(\gamma)},$$

where $P_{\mathbf{SPU}}(\gamma)$ is the p-value of the SPU(γ) test. The default setting of Γ is $\Gamma = \{1, 2, \ldots, 6, \infty\}$, which usually performs well in real applications [42, 44]. We have demonstrated that the result of aSPU is relatively robust to the choice of Γ set from both empirical and theoretical sides [18, 42].

In the end, we apply *one layer of Monte Carlo simulations* to estimate the *p*-values for both SPU and aSPU tests *simultaneously* [26]. Specially, first, we simulate B copies of the null score vectors independently, $U^{(b)} \sim N(0, \boldsymbol{\Sigma})$ for $b = 1, \ldots, B$, from its asymptotic null distribution and then calculate the null statistics $T_{\mathrm{SPU}}(\boldsymbol{\gamma})$ based on each null score vector $U^{(b)}$ accordingly. Next, the *p*-value of SPU($\boldsymbol{\gamma}$) is $P_{\mathrm{SPU}(\gamma)} = [1 + \sum_{b=1}^{B} I(\mid \mathrm{SPU}(\gamma)^{(b)} \mid \geq \mid \mathrm{SPU}(\gamma) \mid)] / (B + 1)$, and the *p*-value for aSPU is $P_{\mathrm{aSPU}} = [1 + \sum_{b=1}^{B} I(T_{\mathrm{aSPU}}^{(b)} \leq T_{\mathrm{aSPU}})] / (B + 1)$, with $T_{\mathrm{aSPU}}^{(b)} = \min_{\gamma \in \Gamma} p_{\gamma}^{(b)}$ and $p_{\gamma}^{(b_1)} = [\sum_{b \neq b_1} I(\mid T_{\mathrm{SPU}(\gamma)}^{(b)} \mid \geq \mid T_{\mathrm{SPU}(\gamma)}^{b_1} \mid)] / B$. To save computational time, we wrote highly efficient code in C and start with a small B, say $B = 10,000$, to scan a genome, then gradually increase B for the more significant genes that pass a significance criterion (*p*-value $< 5/B$) [26].

While both FOGS and TWAS share some similarities such as using the information from external transcriptome data (e.g., GTEx) to construct the test statistics, they differ in two aspects. First, TWAS and its related methods estimate the relationship between cis-SNPs and gene expression based on an external transcriptome dataset, then test the association between the predicted gene expression and phenotype. Under the stringent assumptions that all cis-eSNPs are valid instrumental variables, the standard TWAS, as a special case of instrumental variables regression, like Mendelian randomization (MR) [4], would conclude the causal relationship between the gene and the phenotype. In contrast, our proposed method FOGS conducts a weighted association test on the cis-SNPs and the phenotype. Second, TWAS fails to consider the effects of other SNPs or genes in the same locus, e.g. LD between a true causal SNP and a nearby eSNP, which may lead to the violation of the modeling assumptions for valid instrumental variables regression, and thus lead to invalid causal conclusions. In contrast, FOGS, like FOCUS, is more likely to identify a true causal relationship between the gene and the phenotype after adjusting for the effects of other SNPs and genes in the same locus, similar to multivariable MR [7]. Generally speaking, due to the likely violation of the modeling assumptions required by instrumental variables regression, we regard both TWAS and FOGS as methods to prioritize putative causal genes.

Simulations

We conducted extensive simulations to compare the performance of the proposed method with other two methods: FOCUS [24] and the *p*-value ranking of TWAS results, which, to our knowledge, are the only two existing fine-mapping methods for TWAS. FOCUS is a Bayesian fine-mapping method, and the *p*-value ranking of TWAS results simply chooses the most significant gene in the locus as the putative causal gene.

We focused on Chromosome 22 for all simulations and partitioned genotype data into 24 independent LD blocks by LDetect [6]. We used the CMC eQTL weights (from dorsolateral prefrontal cortex) when available; otherwise, we included the eQTL weights with the best accuracy across all other tissues. 21 out of 24 independent LD blocks contained more than two genes, and we focused on these 21 LD blocks in the following. For each LD block, we

considered all the genes whose transcription start sites are inside the block. Then we applied the fine-mapping methods to each LD block separately.

We first generated GWAS summary statistics from the real genotype data of the Lung Health Study (dbGAP: phs000335.v2.p2) with 4,145 subjects after quality control. Next, we generated the phenotype $\mathbf{Y}$ by $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, where ϵ follows a normal distribution with mean 0 and standard deviation 0.5, representing a random environmental noise; $\mathbf{X}$ is a centered $n \times p$ genotype matrix at p SNPs in the LD block and $\beta = (\beta_1, \ldots, \beta_p)'$ is the corresponding effect size vector.

Under the null, we set $\beta = 0$. Under the alternative, we considered several simulation settings with randomly choosing one gene to be causal gene. Setting 1: For each causal gene, we randomly selected one or two eSNPs (with non-zero eQTL weights) to be causal SNPs and assigned the corresponding $\beta_j = c$, where c is some constant. We also considered the situations where the effect size of causal SNP is proportional to the non-zero eQTL weights. Specifically, we set $\beta_j = c(\hat{w}_j / \sum_j |\hat{w}_j|)$, where $\hat{w}_j$ is the corresponding eQTL weights. Setting 2: We selected all eSNPs in the causal gene to be causal SNPs, assigning $\beta_j = c(\hat{w}_j / \sum_j |\hat{w}_j|)$. Of note, eSNPs for the causal genes are often shared by nearby genes, making setting 2 more challenging in general. In the end, we ran a univariate regression analysis of each SNP and the phenotype, then computed SNP-trait summary results (i.e. the estimated effect sizes and their standard errors, minor allele frequencies, major alleles and minor alleles). For each simulation setting, the heritability was estimated by the R-squared of a linear regression model with the phenotype as the outcome and all causal SNPs as covariates. We repeated this process 120 times for each LD block, resulting in 2,520 LD blocks in total. We then compared and evaluated different methods based on these 2,520 LD blocks.

Analysis of schizophrenia GWAS summary data

We reanalyzed a schizophrenia GWAS summary data of European ancestry with 105,318 individuals (40,675 cases and 64,643 controls) [27]. To our knowledge, this is to date the largest publicly available schizophrenia GWAS summary data.

Data pre-processing.—We removed variants that were either non-biallelic or strand ambiguous (SNPs with A/T, C/G alleles). We further filtered out the variants without rs IDs or not in the 1000 Genomes Project Phase 3 reference panel. We kept the variants on autosomal chromosomes with a minor allele frequency greater than 0.01. 1,146,939 variants passed these filters.

Next, we used the LD loci definition in LDetect [6]. We further obtained publicly available 51 sets of eQTL-derived weights from the FUSION website. Forty-seven sets of the eQTL weights were based on GTEx [10], and the other four were based on NTR [40], YFS [30], METSIM [36], and CMC [13] consortiums, respectively. We assumed the relevant tissue for driving schizophrenia is dorsolateral prefrontal cortex (CMC eQTL weights) because it controls complex, higher-level executive and cognitive functions, many of which are abnormal in schizophrenia patients [13], and because years of studies have pointed out its

well-characterized role in schizophrenia [8, 50]. Similar to Mancuso et al. [24], we used the CMC eQTL weights when available; otherwise, we included the eQTL weights with the best accuracy across all other tissues, leading to 15,460 genes with eQTL weights, 5,420 of which were taken from CMC eQTL weights directly.

Statistical analysis.—We first applied TWAS to all the loci with at least two genes and then applied both FOGS and FOCUS to prioritize putative causal genes.

First, we applied TWAS [16] to identify significantly associated genes. TWAS uses a weighted sum of the Z-scores with eQTL weights to construct its test statistic. Reference LD was estimated by the Lung Health Study (dbGAP: phs000335.v2.p2) data with 4,145 individuals. Because the size of the reference data should scale with the GWAS sample size [5, 47] to reduce the bias, we used the Lung Health Study data as the reference panel instead of the 1000 Genomes Project Phase 3 [1] with about 500 European individuals. We applied TWAS on the loci with more than two genes (14,770 genes in total) and used a stringent Bonferroni cutoff ($0.05/16,000 \approx 3.1 \times 10^{-6}$) to consider the genes and loci we ignored.

Next, we applied two fine-mapping methods, FOGS and FOCUS [24], to prioritize causal genes. We focused on the risk loci that contained at least one genome-wide significant SNP. As before, we used the Lung Health Study (dbGAP: phs000335.v2.p2) data to estimate LD reference panel and ran up to $B = 10^7$ permutations for each gene. For FOGS, we used the same Bonferroni cutoff ($0.05/16,000 \approx 3.1 \times 10^{-6}$) as used in TWAS. For FOCUS, we used the default 90%-credible gene-sets to prioritize causal genes. We also reported the results of the p-value ranking of TWAS results, which chose the most significant gene in the locus as the putative causal gene. In the end, we discussed the results in the loci that contained at least one genome-wide significant SNP in the pre-processed GWAS summary data (p-value $< 5 \times 10^{-8}$) in detail.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

We thank reviewers for helpful comments. This research was supported by the Minnesota Supercomputing Institute. We appreciate the availability of the dbGaP data.

Funding

This research was supported by NIH grants R21AG057038, R01HL116720, R01GM113250 and R01HL105397. CW was supported by a First Year Assistant Professor Grant at Florida State University.

References

[1]. 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature 526(7571), 68–74. [PubMed: 26432245]

[2]. Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, Torstenson ES, Shah KP, Garcia T, Edwards TL, et al. (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nature Communications 9, 1825.

[3]. Barbeira AN, Pividori MD, Zheng J, Wheeler HE, Nicolae DL, and Im HK (2019). Integrating predicted transcriptome from multiple tissues improves association detection. PLoS genetics 15(1), e1007889. [PubMed: 30668570]

[4]. Barfield R, Feng H, Gusev A, Wu L, Zheng W, Pasaniuc B, and Kraft P (2018). Transcriptome-wide association studies accounting for colocalization using egger regression. Genetic epidemiology 42(5), 418–433. [PubMed: 29808603]

[5]. Benner C, Havulinna AS, Järvelin M-R, Salomaa V, Ripatti S, and Pirinen M (2017). Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. The American Journal of Human Genetics 101 (4), 539–551. [PubMed: 28942963]

[6]. Berisa T and Pickrell JK (2016). Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics 32(2), 283–285. [PubMed: 26395773]

[7]. Burgess S and Thompson SG (2015). Multivariable mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. American journal of epidemiology 181 (4), 251–260. [PubMed: 25632051]

[8]. Callicott JH, Egan MF, Mattay VS, Bertolino A, Bone AD, Verchinksi B, and Weinberger DR (2003). Abnormal fmri response of the dorsolateral prefrontal cortex in cognitively intact siblings of patients with schizophrenia. American Journal of Psychiatry 160(4), 709–719. [PubMed: 12668360]

[9]. Collier DA, Eastwood BJ, Malki K, and Mokrab Y (2016). Advances in the genetics of schizophrenia: toward a network and pathway view for drug discovery. Annals of the New York Academy of Sciences 1366(1), 61–75. [PubMed: 27111133]

[10]. Consortium G (2017). Genetic effects on gene expression across human tissues. Nature 550(7675), 204–213. [PubMed: 29022597]

[11]. DeLong ER, DeLong DM, and Clarke-Pearson DL (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44 (3), 837–845. [PubMed: 3203132]

[12]. Dodd LE and Pepe MS (2003). Partial auc estimation and regression. Biometrics 59(3), 614–623. [PubMed: 14601762]

[13]. Fromer M, Roussos P, Sieberts SK, Johnson JS, Kavanagh DH, Perumal TM, Ruderfer DM, Oh EC, Topol A, Shah HR, et al. (2016). Gene expression elucidates functional impact of polygenic risk for schizophrenia. Nature Neuroscience 19(11), 1442–1453. [PubMed: 27668389]

[14]. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, Eyler AE, Denny JC, Nicolae DL, Cox NJ, et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. Nature Genetics 47(9), 1091–1098. [PubMed: 26258848]

[15]. Goes FS, McGrath J, Avramopoulos D, Wolyniec P, Pirooznia M, Ruczinski I, Nestadt G, Kenny EE, Vacic V, Peters I, et al. (2015). Genome-wide association study of schizophrenia in Ashkenazi Jews. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics 168(8), 649–659.

[16]. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW, Jansen R, De Geus EJ, Boomsma DI, Wright FA, et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nature Genetics 48(3), 245–252. [PubMed: 26854917]

[17]. Han B and Eskin E (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. The American Journal of Human Genetics 88(5), 586–598. [PubMed: 21565292]

[18]. He Y, Xu G, Wu C, and Pan W (2018). Asymptotically independent u-statistics in high-dimensional testing. arXiv preprint arXiv:1809.00411.

[19]. Hu Y, Li M, Lu Q, Weng H, Wang J, Zekavat SM, Yu Z, Li B, Gu J, Muchnik S, et al. (2019). A statistical framework for cross-tissue transcriptome-wide association analysis. Nature Genetics 51 (3), 568–576. [PubMed: 30804563]

[20]. Kichaev G and Pasaniuc B (2015). Leveraging functional-annotation data in trans-ethnic fine-mapping studies. The American Journal of Human Genetics 97(2), 260–271. [PubMed: 26189819]

[21]. Kwak I-Y and Pan W (2015). Adaptive gene-and pathway-trait association testing with GWAS summary statistics. Bioinformatics 32(8), 1178–1184. [PubMed: 26656570]

[22]. Li Z, Chen J, Yu H, He L, Xu Y, Zhang D, Yi Q, Li C, Li X, Shen J, et al. (2017). Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. Nature Genetics 49(11), 1576–1583. [PubMed: 28991256]

[23]. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, et al. (2016). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Research 45(D1), D896–D901. [PubMed: 27899670]

[24]. Mancuso N, Freund MK, Johnson R, Shi H, Kichaev G, Gusev A, and Pasaniuc B (2019). Probabilistic fine-mapping of transcriptome-wide association studies. Nature Genetics 51 (4), 675–682. [PubMed: 30926970]

[25]. McGrath J, Saha S, Chant D, and Welham J (2008). Schizophrenia: a concise overview of incidence, prevalence, and mortality. Epidemiologic Reviews 30(1), 67–76. [PubMed: 18480098]

[26]. Pan W, Kim J, Zhang Y, Shen X, and Wei P (2014). A powerful and adaptive association test for rare variants. Genetics 197(4), 1081–1095. [PubMed: 24831820]

[27]. Pardiñas AF, Holmans P, Pocklington AJ, Escott-Price V, Ripke S, Carrera N, Legge SE, Bishop S, Cameron D, Hamshere ML, et al. (2018). Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. Nature Genetics 50(3), 381–389. [PubMed: 29483656]

[28]. Pasaniuc B, Zaitlen N, Shi H, Bhatia G, Gusev A, Pickrell J, Hirschhorn J, Strachan DP, Patterson N, and Price AL (2014). Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. Bioinformatics 30(20), 2906–2914. [PubMed: 24990607]

[29]. Pirinen M, Donnelly P, Spencer CC, et al. (2013). Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. The Annals of Applied Statistics 7(1), 369–390.

[30]. Raitakari OT, Juonala M, Rönnemaa T, Keltikangas-Järvinen L, Räsänen L, Pietikäinen M, Hutri-Kähönen N, Taittonen L, Jokinen E, Marniemi J, et al. (2008). Cohort profile: the cardiovascular risk in young finns study. International Journal of Epidemiology 37(6), 1220–1226. [PubMed: 18263651]

[31]. Ripke S, Neale BM, Corvin A, Walters JT, Farh K-H, Holmans PA, Lee P, Bulik-Sullivan B, Collier DA, Huang H, et al. (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature 511 (7510), 421–427. [PubMed: 25056061]

[32]. Schaid DJ, Chen W, and Larson NB (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics 19, 491–504.

[33]. Sekar A, Bialas AR, de Rivera H, Davis A, Hammond TR, Kamitaki N, Tooley K, Presumey J, Baum M, Van Doren V, et al. (2016). Schizophrenia risk from complex variation of complement component 4. Nature 530(7589), 177–183. [PubMed: 26814963]

[34]. Shen X, Pan W, and Zhu Y (2012). Likelihood-based selection and sharp parameter estimation. Journal of the American Statistical Association 107(497), 223–232. [PubMed: 22736876]

[35]. Spain SL and Barrett JC (2015). Strategies for fine-mapping complex traits. Human molecular genetics 24 (R1), R111–R119. [PubMed: 26157023]

[36]. Stančáková A, Civelek M, Saleem NK, Soininen P, Kangas AJ, Cederberg H, Paananen J, Pihlajamäki J, Bonnycastle LL, Morken MA, et al. (2012). Hyperglycemia and a common variant of gckr are associated with the levels of eight amino acids in 9,369 finnish men. Diabetes 51 (1), 1895–1902.

[37]. The Autism Spectrum Disorders Working Group (2017). Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24. 32 and a significant overlap with schizophrenia. Molecular Autism 8(21), 1–17. [PubMed: 28070266]

[38]. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, Genovese G, Loh P-R, Bhatia G, Do R, et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. The American Journal of Human Genetics 97(4), 576–592. [PubMed: 26430803]

[39]. Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, Ermel R, Ruusalepp A, Quertermous T, Hao K, et al. (2019). Opportunities and challenges for transcriptome-wide association studies. Nature Genetics 51 (4), 592–599. [PubMed: 30926968]

[40]. Wright FA, Sullivan PF, Brooks AI, Zou F, Sun W, Xia K, Madar V, Jansen R, Chung W, Zhou Y-H, et al. (2014). Heritability and genomics of gene expression in peripheral blood. Nature Genetics 46(5), 430–437. [PubMed: 24728292]

[41]. Wu C and Pan W (2018). Integrating eQTL data with GWAS summary statistics in pathway-based analysis with application to schizophrenia. Genetic Epidemiology 42(3), 303–316. [PubMed: 29411426]

[42]. Wu C, Xu G, and Pan W (2019). An adaptive test on high-dimensional parameters in generalized linear models. Statistica Sinica.

[43]. Xu B, Roos JL, Levy S, Van Rensburg E, Gogos JA, and Karayiorgou M (2008). Strong association of de novo copy number mutations with sporadic schizophrenia. Nature Genetics 40(7), 880–885. [PubMed: 18511947]

[44]. Xu Z, Wu C, Wei P, and Pan W (2017). A powerful framework for integrating eQTL and GWAS summary data. Genetics 207, 893–902. [PubMed: 28893853]

[45]. Xue H, Pan W, Initiative ADN, et al. (2019). Some statistical consideration in transcriptome-wide association studies. bioRxiv, 812677.

[46]. Yang C, Wan X, Lin X, Chen M, Zhou X, and Liu J (2018). Comm: a collaborative mixed model to dissecting genetic contributions to complex traits by leveraging regulatory information. Bioinformatics 35(10), 1644–1652.

[47]. Yang J, Ferreira T, Morris AP, Medland SE, Madden PA, Heath AC, Martin NG, Montgomery GW, Weedon MN, Loos RJ, et al. (2012). Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits. Nature Genetics 44 (4), 369–375. [PubMed: 22426310]

[48]. Zhang C-H (2010). Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics 38(2), 894–942.

[49]. Zhao Q, Wang J, Hemani G, Bowden J, and Small DS (2018). Statistical inference in two-sample summary-data mendelian randomization using robust adjusted profile score. arXiv preprint arXiv: 1801.09652.

[50]. Zhou Y, Liang M, Jiang T, Tian L, Liu Y, Liu Z, Liu H, and Kuang F (2007). Functional dysconnectivity of the dorsolateral prefrontal cortex in first-episode schizophrenia using resting-state fmri. Neuroscience Letters 417(3), 297–302. [PubMed: 17399900]

[51]. Zhu X and Stephens M (2017). Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. The annals of applied statistics 11 (3), 1561. [PubMed: 29399241]

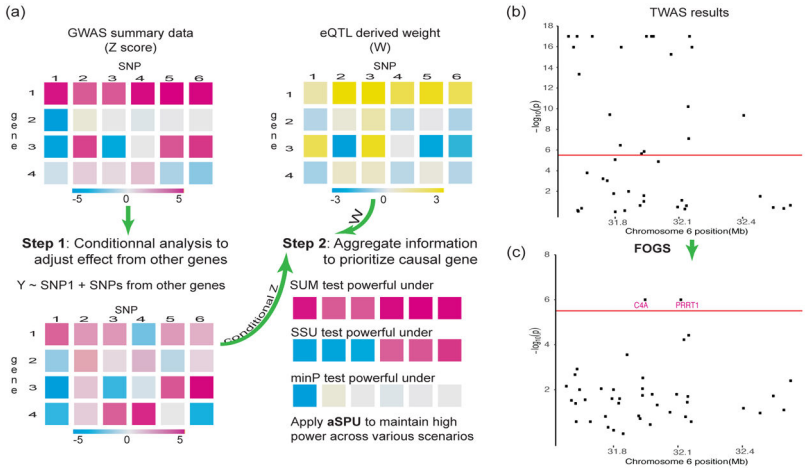

Figure 1: Overview of the fine-mapping method: FOGS.

a. FOGS prioritizes causal genes by two steps: first, FOGS conducts a conditional analysis with ridge regression to account the effects of other SNPs/genes in the locus; second, FOGS integrates eQTL-derived weights and conditional Z scores by an adaptive test to maintain high power. **b** and **c**, TWAS and FOGS were performed on summary statistics from a schizophrenia GWAS of 105,318 European individuals [27]. **b**, TWAS results for a locus on chromosome 6 (31.6–32.7 Mb). **c**, FOGS results for the same locus on chromosome 6 (31.6–32.7 Mb). 19 out of 42 genes in the locus has been identified to be significant by TWAS, prompting for fine-mapping. FOGS prioritized putative causal genes by successfully identifying the well-known schizophrenia-related gene *C4A* [33].

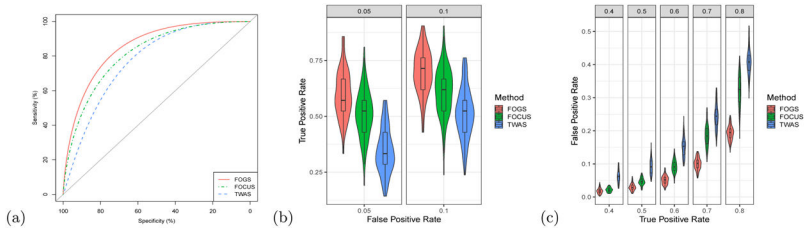

Figure 2: Performance assessment of different methods with simulated data.

We randomly selected two SNPs in one gene to be causal, and the effect size was $c = 0.1$. The estimated hertibability was about 2.5%. **a**, AUC comparison between different methods. **b**, True positive rate (Power) comparison under some specified false positive rates. **c**, False positive rate comparison under some specified true positive rates. The violin plot and box plot inside display the false/true positive rates of different methods under specified true/false positive rates.

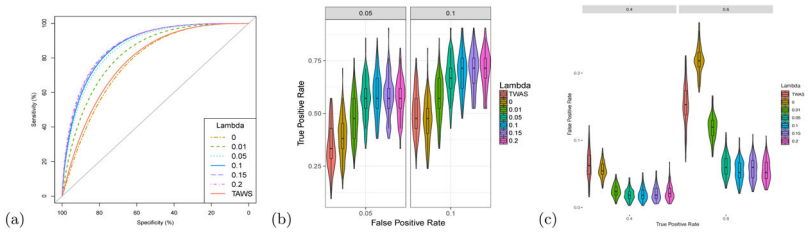

Figure 3: Sensitive analysis of ridge penalty parameter λ of FOGS with simulated data.

We randomly selected two SNPs in one gene to be causal and the effect size was $c = 0.1$. **a**, AUC comparison between different methods. **b**, True positive rate (Power) comparison under some specified false positive rates. **c**, False positive rate comparison under some specified true positive rates. The violin plot and box plot inside display the false/true positive rates of different methods under specified true/false positive rates.

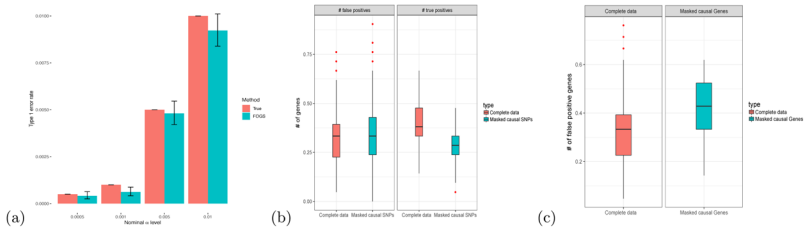

Figure 4: FOGS performance under situations with causal SNPs missing.

We considered the following three cases: **a**, no SNP-trait association for all SNPs in the locus; **b**, the two causal SNPs were missing; **c**, the causal gene (with all its SNPs) was missing (see main text for details). **a**, Histogram of Type I error rates under different α levels. Vertical line segments mark 95% confidence intervals for the estimated Type I error rates. **b**, Numbers of false positives and true positives per locus for FOGS with either complete data or data with causal SNPs missing. **c**, Number of false positives per locus for FOGS with either complete data or data with the causal gene missing.

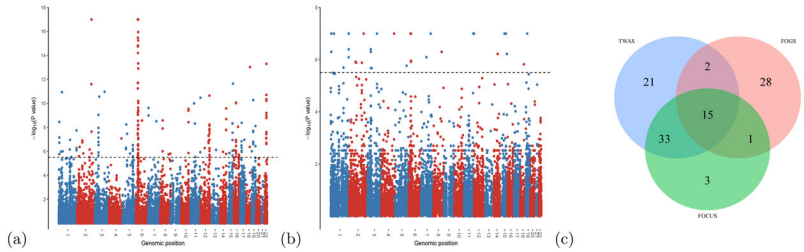

Figure 5: Applications to a GWAS schizophrenia summary data [27].

a, Manhattan plot of all TWAS associations. **b**, Manhattan plot of FOGS results (see Methods). For **a** and **b**, each point represents a single gene tested, with physical position (P0) plotted on the x axis and $-\log(p)$ values plotted on the y axis. Dashed lines indicate genome-wide significant threshold ($0.05/16{,}000 \approx 3.1 \times 10^{-6}$). **c**, Venn diagram of the putative causal genes prioritized by different methods for the risk regions that contained at least two genes. TWAS stands for the results of selecting the most significant gene in each region plus the genes with p-value $< 1e-17$ (unable to distinguish) as putative causal genes. FOGS and FOCUS represent the results of FOGS with a stringent cutoff ($0.05/16{,}000 \approx 3.1 \times 10^{-6}$) and FOCUS with 90% credible gene-sets, respectively (see Methods).

Table 1:

Putative causal genes identified by our new method FOGS in the GWAS risk loci that contained at least one genome-wide significant SNP.

FOGS identified 22 putative causal genes with their chromosomes (CHR), transcription start sites (TSS) and relevant previous studies (Reported) in the GWAS Catalog [23]. For comparison, we also list the p-values of TWAS [14, 16] and the posterior probabilities of FOCUS [24]

Gene	CHR	Position	FOGS	TWAS	FOCUS	Reported
PTPRF	1	43990858	$< 1.0 \times 10^{-7}$	2.2×10^{-2}	0.00	[15, 22, 31, 37]
MED8	1	43849578	9.0×10^{-7}	1.1×10^{-11}	1.00	[15, 31, 37]
THOC7	3	63819545	1.7×10^{-6}	2.7×10^{-11}	1.00	[15, 31, 37]
FAM114A2	5	153371268	5.0×10^{-7}	5.3×10^{-6}	0.78	
BTN2A1	6	26458152	$< 1.0 \times 10^{-7}$	2.8×10^{-4}	0.00	[37]
HIST1H2BJ	6	27093676	$< 1.0 \times 10^{-7}$	5.5×10^{-1}	0.00	
HIST1H4L	6	27840926	$< 1.0 \times 10^{-7}$	$< 1.0 \times 10^{-17}$	1.00	
TRIM27	6	28870779	$< 1.0 \times 10^{-7}$	$< 1.0 \times 10^{-17}$	0.00	
HLA-F-AS1	6	29696142	5.0×10^{-7}	8.1×10^{-1}	0.00	
MICE	6	29709508	$< 1.0 \times 10^{-7}$	1.9×10^{-1}	0.00	
ZFP57	6	29640168	$< 1.0 \times 10^{-7}$	$< 1.0 \times 10^{-17}$	1.00	[15, 37]
PPP1R18	6	30644166	$< 1.0 \times 10^{-7}$	$< 1.0 \times 10^{-17}$	1.00	
FLOT1	6	30695486	$< 1.0 \times 10^{-7}$	1.3×10^{-8}	0.00	
IER3	6	30710979	$< 1.0 \times 10^{-7}$	$< 1.0 \times 10^{-17}$	0.54	
C4A	6	31949801	$< 1.0 \times 10^{-7}$	$< 1.0 \times 10^{-17}$	1.00	
PRRT1	6	32116136	2.0×10^{-7}	7.4×10^{-2}	0.00	[37]
OPCML	11	132284874	$< 1.0 \times 10^{-7}$	1.2×10^{-6}	0.94	
IGSF9B	11	133778459	$< 1.0 \times 10^{-7}$	3.4×10^{-11}	1.00	[15]
B3GAT1	11	134248397	$< 1.0 \times 10^{-7}$	3.3×10^{-3}	0.01	[15]
RGS6	14	72398816	$< 1.0 \times 10^{-7}$	2.8×10^{-2}	0.00	[15, 31, 37]
HYKK	15	78799906	$< 1.0 \times 10^{-7}$	6.0×10^{-2}	0.27	
FURIN	15	91411884	4.0×10^{-7}	2.3×10^{-12}	1.00	[15, 31]