A mediation analysis of schizophrenia variants: Risk alleles may affect a gene's expression differently from how the expression affects risk.

Xi Peng[1], Joel Bader[1] and Dimitrios Avramopoulos[2,3,*]

1. Department of Biomedical Engineering, Whiting School of Engineering and School of Medicine, Johns Hopkins University

2. Department of Genetic Medicine, Johns Hopkins University School of Medicine

3. Department of Psychiatry, Johns Hopkins University School of Medicine

* Corresponding author

Mailing address: 733 N. Broadway, MRB-507, Baltimore, MD 21205

Tel. 410 955-8323

Fax: 410 955-7397

Email: adimitr1@jhmi.edu

Keywords: Schizophrenia, mediation, expression quantitative trait locus (eQTL).

Running title: Mediation analysis in schizophrenia

ABSTRACT

Variants identified by Genome wide association studies (GWAS) are often expression quantitative trait loci (eQTLs), suggesting they are proxies or are themselves regulatory. Additionally eQTL analyses show that variants often affect more than one gene. Lacking data on many tissue types, developmental time points and homogeneous cell types, the extent of this one-to-many relationship is underestimated. This raises questions on whether a disease eQTL target gene explains the genetic association or is a by-stander. It also puts into question the direction of the effect of a gene's expression on the risk, since the many genes regulated by the same variant may have opposing effects, imperfectly balancing each other. We used two brain gene expression datasets (CommonMind and BrainSeq) for a mediation analysis of schizophrenia-associated variants. We find that indeed eQTL targets often mediate risk but the direction in which expression affects risk is often different from the direction in which the risk allele changes expression. Of 38 genes significant for mediation in both datasets 33 showed consistent direction (Chi^2 test $P=6*10^{-6}$) and for 15 of them (45%) the expression change associated with the risk allele was protective, which suggests the likely presence of other target genes with overriding effects. Our results identify specific risk mediating genes and suggest caution in interpreting the biological consequences of targeted modifications of gene expression, as not all eQTL targets may be relevant to disease and those that are might have different than expected directions.

INTRODUCTION

Schizophrenia (SZ) is a common and disabling mental disorder with a point prevalence of 0.5%, an onset in late adolescence or early adulthood and a life long course with significant disability (Messias et al. 2007). It has high heritability, among the highest in psychiatric disorders, consistently estimated around 80% (Kety 1987; Cardno et al. 1999). Although recognized diagnostically as a single disorder, SZ has a highly heterogeneous phenotype with symptoms that range from prominent delusions, hallucinations agitation and erratic behavior to lack of interest and motivation, apathy and disorganization. The response of patients to different treatments is also highly variable. While some of this heterogeneity might be the result of environmental effects a large proportion of the variability likely reflects the underlying genetic heterogeneity and the compromise of different biological processes that alone or in combinations lead to the phenotype we call "SZ". Once we understand the links between genetic variation and biological processes, genetic testing might predict each patient's response to treatment or liability to environmental exposures. This would be a tremendous step forward in personalized prevention and treatment, a benefit for the patient and the society.

Over the last few years, thanks to large collaborative genome wide association studies (GWAS) such as by the Psychiatric Genomics Consortium (PGC: https://pgc.unc.edu/), SZ has become the psychiatric disorder with the largest number of genetic variants robustly shown to contribute to risk. The number of SZ loci has steadily increased from 5 to 22 and currently over 100 (Schizophrenia GWAS Consortium 2011; Ripke et al. 2013; Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014; Pardinas et al. 2018). As observed with GWAS-identified loci across other complex disorders the SZ-associated variants are most often located in non-coding sequences and 40% of the time their haplotype blocks do not include coding exons (Hindorff et al. 2009; Manolio et

al. 2009; Visel et al. 2009). These variants are presumed to be regulatory, which is supported by their concentration in open chromatin DNA as marked by deoxyribonuclease I (DNase I) hypersensitive sites (DHSs) (Maurano et al. 2012) and by studies of Quantitative Trait Loci (eQTLs) (Gilad et al. 2008; Cookson et al. 2009; Hindorff et al. 2009; Majewski and Pastinen 2011; Schaub et al. 2012). Regulatory sequences however can be far from their target gene so it is difficult to assign a specific gene or genes to each variant solely by location. In fact it has been shown using chromatin interaction data that in most cases the nearest gene to the variant is not the one affected by it (Maurano et al. 2012). Further, regulatory sequences often regulate more than one gene as shown by interactions with multiple promoters (Akerborg et al. 2019) and observed in eQTL databases (GTEx: gtexportal.org, Commonmind: www.nimhgenetics.org/resources/commonmind and BrainSeq: eqtl.brainseq.org). Also, as eQTL discovery depends on the studied cell type, tissue and developmental time point and current studies are far from covering all these possibilities, it is likely that there remain undiscovered variant-gene correlations and that many more eQTLs might regulate multiple rather than one gene. In fact, because studies of eQTLs specific to development are rare (O'Brien et al. 2018) and done in bulk tissue, many eQTLs likely remain unknown. These missing eQTLs are also the most likely to be of importance for a developmental brain disorder like SZ.

Alleles changing the expression of a gene in a way that increases disease risk will be subject to selective pressure, especially if this effect is relatively large. If however they regulate multiple genes with opposing effects of on the risk (Figure 1) a large effect would be dampened and the allele may escape selection. This also means that if a disease risk allele increases the expression of the gene it does not necessarily follow that increased expression translates to higher risk. This is of great importance as

disease-modeling studies often perturb gene expression to study the outcome and understand the biology of disease (Hill et al. 2012; Yang et al. 2018; Schrode et al. 2019). It is therefore necessary to seek formal evidence that a specific gene's expression mediates disease risk and in which direction, if we want to have an accurate list of the genes and a correct understanding of their role in disease.

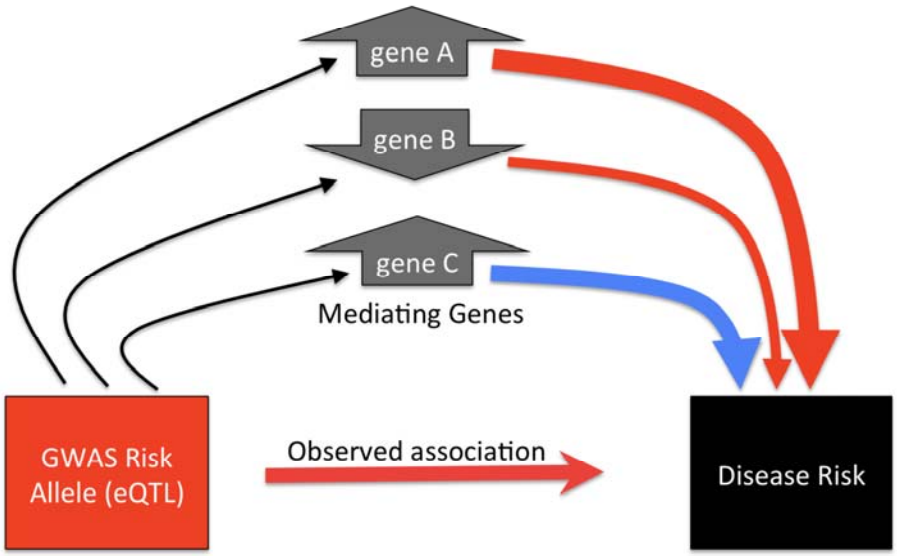

Figure 1: Mediation through multiple genes. The risk allele changes the expression of multiple genes up- or downward (indicated by the arrow box around each gene). This modification may increase risk (curved red arrows, width reflects effect size) or decrease it (blue arrow). The association of the risk allele with higher risk (straight red arrow) reflects the sum of all these effects on risk. The mediation of some genes expression change may not follow the direction of the the risk allele's effect, as in gene C here.

Mediation analysis is a statistical method developed to examine whether a relationship between two variables (genotype and SZ risk in this case) can be explained by a third, mediating variable - in this case gene expression (MacKinnon et al. 2007). Importantly the "effect to be mediated", here the effect of the variant on risk, does not need to be significant in order to test for mediation (Mackinnon and Fairchild 2009; Zhao et al. 2010). This is important for our analysis described below because, while all variants were selected to be significant by GWAS, they were not necessarily also significant in the much smaller postmortem tissue datasets. Mediation analysis has been widely used in

psychology, yet it has only rarely been applied to test whether gene expression mediates disease risk. Formally showing mediation for disease associated variants that are eQTLs is important as it can (i) validate the assumption that a gene regulated by a variant is important for the disease; (ii) point to the specific genes where this can be formally shown to be true, as opposed to other genes regulated by the same variant but irrelevant to disease; (iii) indicate the direction of the effect of the gene on disease which might not coincide with the direction of the effect of the risk allele if more than one disease genes are regulated by the same variant.

Here we hypothesize that for many of the known eQTLs that are also associated with SZ: A) The eQTL target gene frequently mediates the effect of the single nucleotide polymorphism (SNP) on the risk and B) the direction of effect of the risk allele on gene expression may not be the same with the direction of effect of the gene expression on the risk. To test these hypotheses, we perform mediation analysis on two large public datasets, Commonmind (CMC) and BrainSeq Consortium (BSC). We find evidence supporting this hypothesis and report on mediating genes and their direction of effect at multiple levels of statistical confidence.

RESULTS

We first performed an eQTL in the CMC and BSC datasets to identify our target pairs of gene-SNP group (set of SNPs in LD, see methods) for mediation analysis. The complete results are reported in detail in Supplementary Table 1. Below we provide a few metrics and highlight some of these results

CMC dataset

In eQTL analysis we found 14,258 cis and 411 trans significant SNP-gene pairs (cis-: P < 0.01, trans-: P < 1*10-7, FDR ~5%). Of the 219 independent SNP groups (groups of SNPs in LD – see methods) 106 were significantly correlated with the expression of 311 genes in cis and 12 with 18 genes in trans. Of the 106 cis-eQTLs groups, 55 (51.9%) were correlated with the expression of more than one gene with a maximum of 22 for SNP group 80. Of the 311 genes with cis-eQTLs 13 (4.2%) were correlated with 2 SNP groups at the same locus (<500 Kb). For trans-eQTLs, 3 out of 12 groups (25.0%) and the target genes were located on different chromosomes. These were group 19 on Chr1, which was an eQTL for *HECW1* (ENSG00000002746) on Chr7; Group 24 on Chr1 was an eQTL for *CDC27* (ENSG00000004897) on Chr17. Group 166 on Chr14, was a eQTL for genes *EPHA10* (ENSG00000183317) on Chr1 and *SLC35B1* (ENSG00000121073) on Chr17

Group166 was also a cis-eQTL for genes AC005477.1 and *RGS6*; genes *EPHA10* and *SLC35B1* had no nearby SZ-associated SNPs so it was not tested for cis-eQTLs. Group 24 was a cis-eQTL for *FANCL*. *HECW1* was also not tested for cis-eQTLs.

The remaining 9 trans-eQTL groups were <6Mb away from the correlated genes and 7 of them were <1Mb. Four out of the 12 trans-eQTL SNP groups (33.3%) were correlated with more than one gene. Only 1 of these 4 groups (Group166) was located on a different chromosome than the correlated genes.

BSC dataset

In the BSC dataset eQTL analysis we found 8,958 significant SNP-gene eQTL pairs in cis. That corresponded to 92 SNP groups being eQTLs for 272 genes. Of those 52 (56.5%) groups were eQTLs for more than one, with a maximum of 14. Six (2.2%) genes had two eQTL SNP groups. For trans-eQTL, 220 SNP-gene pairs were found corresponding to seven distinct SNP group-gene pairs. For 4 out of 7 (57.1%) the gene

and SNP group were on different chromosomes These were Group51 on Chr3 with gene *PSENEN* (ENSG00000205155) on Chr19 (Group51 was also cis-eQTL for gene *LRRFIP2*, *DCLK3, TRANK1*, while *PSENEN* did not have nearby SNP groups); Group11 on Chr1 with *OSBP* (ENSG00000110048) on Chr11 (Group11 was also cis-eQTL for *BRINP2*, while *OSBP* did not have nearby SNP groups); Group174 on Chr15 with *GMPS* (ENSG00000163655) on Chr3, (Group 174 was also cis- eQTL for PSMA4, *CRABP1, CTSH* while *GMPS* did not have nearby SNP groups); Group129 on Chr10, with *PPAPDC1B* (ENSG00000147535) on Chr8 (Group 129 was also cis- eQTL for genes *CNNM2*, C10orf32, *NT5C2*, RP11-18I14.10, *CALHM2*, *PSD*, *CUEDC2*, *PCGF6*, *FBXL15*, *TRIM8*, *TMEM180*, *INA*, while *PPAPDC1B* although it was within SNP group 111 did not have cis-eQTLs). All of the other groups were located in a <3 Mb region from the genes and 1_of them <1Mb.

eQTL overlaps between CMC and BSC

Of the identified cis- and trans-eQTLs, 70 SNP groups that were eQTLs for 149 genes in cis, and 2 SNP groups that were eQTLs for 2 genes in trans respectively overlapped between the CMC and BSC datasets. Most of the cis-overlaps (68 of 70 SNP groups that were eQTLs for 137 genes) and all of the trans-overlaps showed consistent direction between the two datasets (Supplementary Table 1).

Testing the mediation analysis platform via Simulations

To test whether the mediation analysis platform behaved as expected in the detection of the direction of effects in the presence of multiple mediators we performed analysis on simulated data and repeated the same mediation analysis on a number of simulated Independent (I) - Mediator (M) - Outcome (O) variable combinations (listed in Supplementary Table 2 A, B & C). When one I was mediated by multiple M in varying

directions, the analysis correctly identified both positive and negative mediations. When one I influenced two M by the same coefficient, but only one of them impacted O, we correctly saw significant mediating effect of that M but not the other (analyzed separately). In the case of the non-mediating M, the I had significant residual effect on the O and there was no significant mediating effect, correctly indicating the presence of another mediator. If two I impacted the O through two M, we saw significant mediating effects for both M and significant residual effects from each I to the O, correctly indicating the presence of an additional mediator. We concluded that the mediation analysis was performing as expected under a variety of possible underlying effects and combinations.

Permutations to calculate false discovery (FDR)

As we discuss in the methods, the analysis platform we use to test mediation only reports bootstrap based confidence intervals (CI) and only at 95%. This along with our testing all correlated SNPs in each LD group complicates assessing the true significance of positive results. To achieve higher confidence we extended the CI as described in the methods. To correctly assess mediation significance and calculate reliable FDR we permuted the link between the genotypes and either both the gene expression and phenotype, or just the phenotype (preserving genotype-expression correlations), and repeated the same mediation analysis for each permutation, counting the number of positive results and comparing with the observed results. Initial permutations showed that not preserving genotype-expression correlations greatly reduced the number of observed positives and therefore would be too liberal so we continued by permuting only the link to phenotype.

For the CMC dataset, these permutations of only phenotype data showed an average of 37.9 significant mediating effects with SD=7.4 suggesting an FDR of 34% based on our

results (below). Under the extended CI, the mean number of genes with significant mediating effects was 0.93 (SD = 1.07), thus given 15 observed significant mediating effects, the FDR was 6.2%. The results under the extended CI are therefore more reliable for the purpose of follow up.

For the BSC dataset the mean of the number of genes with significant mediating effects in permutations was 32.16 (SD= 8.50) so considering 178 observed positives (see below) the FDR was around 18.1%. Under the extended CI, the mean of number of genes with significant mediating effects was 0.64 (SD= 0.83). Therefore with 59 significant mediating effects at this level, the FDR is 1.1%.

Mediation results

In the CMC dataset, 4,156 of the 14,669 SNP-gene pairs were significant for mediation (the 95% CI calculated by bootstrapping did not include 0, see methods). This corresponds to 68 SNP groups (of the 106 tested - 64%) significantly mediated by 113 genes under the 95% CI model (Supplementary Table 3). About half (56) of the 113 genes showed a negative mediating effect, meaning that the correlation between the risk allele and the gene was in an opposite direction to the correction between the gene and the phenotype. Under the stringent extended CI model (see methods), the number of significant SNP-gene pairs was 15 which consisted of 14 SNP groups that were eQTLs for 15 genes. 7 of these genes showed a negative mediating effect.

In the BSC dataset, 5,575 out of 9,178 pairs were significant for the mediation, corresponding to 78 of the 92 tested SNP groups (85%) significantly mediated by 178 genes under 95% CI model (Supplementary Table 3). Similar to CMC, about half (93) of the 178 genes (52.9%) showed a negative mediating effect. Under the more stringent

extended CI model, the number of significant SNP-gene pairs was 60 consisting of 40 SNP groups and 59 genes. 33 of these 59 genes showed a negative mediating effect.

To validate the direction of mediation across datasets we examined the overlapping SNP group – gene pairs between datasets for consistency. Of the 153 genes included in both datasets mediation analysis 38 were significant in both under 95% the CI model. Of those 38, 33 (87%) showed consistent direction in the two studies (Chi^2 test $P=6*10^{-6}$). Of these 33, 18 had positive and 15 had negative mediating effect (i.e the risk allele assisted expression change mediated protection from SZ). Under the extended CI, four genes showed significant mediating effect in both datasets all of them with consistent directions. Three of them (*CSPG4P12, DDHD2, GOLGA2P7*) had positive and one (*ZMAT2*) had negative effects as defined above.

DISCUSSION

Our goal was to formally test whether genetic associations between common variants and disease are mediated by gene regulation and to determine whether the direction of this mediation is that expected from the direction in which the risk allele modifies the expression. To achieve this goal we performed a mediation analysis on the two largest independent gene expression datasets from the dorsolateral prefrontal cortex of SZ cases and controls - the CMC and the BSC. We tested variants associated with schizophrenia that are also eQTLs in these datasets. Due to the complexities of combining GWAS data and expression data from different datasets in the presence of LD, where different SNPs can represent the same association/eQTL signal, we report on SNP groups, groups of SNPs in LD. To avoid false positives due to the multiple testing

within and between these groups we set some initial p-value thresholds but calculated FDR based on permutations.

Our full list of eQTLs is provided on Supplementary Table 2 and as expected is not much different from previous reports on these same datasets (Fromer et al. 2016; Jaffe et al. 2018). By reducing the search space to only SZ-associated SNPs we also identify a number of trans-eQLTs described above. Interestingly many of them were also cis-eQTLs affecting multiple genes, while many of the genes affected in trans were at locations that showed no genetic associations with SZ. It is possible, especially if one accounts for the possibility of opposing effects of different genes on the risk, that while these are true risk genes there is either no local regulatory variation, or that such variation also influences other genes and the sum of effects on risk is minimal.

We find that a large number of GWAS signals show evidence that they are mediated by at least one of the genes for which they that are eQTLs, as is generally accepted by the literature. This is about 64% of SNP groups in the CMC dataset and 85% in the BSC dataset at the relaxed criteria FDR of 33.5% and 18.1% respectively. This suggest more than half of the signals are mediated by gene expression, which given power limitations is likely a low estimate.

Four genes (*CSPG4P12, DDHD2, GOLGA2P7, ZMAT2)* showed significant mediating effect in both datasets under our stringent criteria with FDR 6.2% and 1.1% in the CMC BSC respectively, all of them with consistent directions. We consider these high confidence schizophrenia genes.

Higher expression of *CSPG4P12* was found to mediate increased risk for SZ. It is a pseudogene of CSPG4, a chondroitin sulfate proteoglycan that is a marker gene for oligodendrocyte progenitor cells. According to data from GTEx *CSPG4P12* is lowly expressed in many tissues including brain where SZ-associated SNPs affect its

regulation. Higher expression of *DDHD2* was found to mediate increased risk for SZ. *DDHD2* is a principal brain triglyceride lipase known to cause a recessive form of complex hereditary spastic paraplegia (Inloes et al. 2014). A de novo frameshift mutation in this gene has been reported in an Afrikaner schizophrenia patient (Xu et al. 2012). Higher expression of *GOLGA2P7* was found to mediate increased risk for SZ. *GOLGA2P7* is a pseudogene of *GALGA2,* a Golgi apparatus related gene. *GOLGA2P7* has also been reported by Jaffe et al to be significantly developmentally regulated (Jaffe et al. 2018). Lower expression of expression of *ZMAT2* was found to mediate increased risk for SZ. A zinc-finger protein, *ZMAT2* is expressed in multiple tissues including high expression in the brain (GTEx data). In epidermal cells it is know to be an interactor of the pre-spliceosome that is required to keep cells in an undifferentiated, proliferative state (Tanis et al. 2018). It has also been reported by Jaffe et al to be significantly developmentally regulated (Jaffe et al. 2018).

We identify multiple instances where the mediation is not in the direction suggested by the effect of the risk allele on expression. This means that although the disease risk allele correlates with decreased expression of a gene, the decrease in expression mediates lower risk or the reverse. The validity of this result is supported not only by its presence in both datasets, but also by the high consistency of the overlapping signals. This observation is of great importance for the design of studies of the biological link between genetic variation and disease. For example, examining the consequences of a gene knock down under the wrong assumption that lower expression mediated higher risk would lead to wrong conclusions. We hypothesize that this commonly observed apparent discordance in direction is because it is common for variants to regulate multiple genes. We observe this one-to-many relationship in our results and it is likely to be much more widespread, if one accounts for statistical power and the study of mixed cell populations and single time points. Given the strong selective disadvantage of

schizophrenia (Power et al. 2013) it is expected that variants that have effects that counteract each other may be more likely to remain in the population. In this case the small effect sizes on disease risk that we consistently in GWAS variants may reflect their combined effect on multiple genes, the individual effect possibly being more pronounced.

Our results highlight the complexity of the interplay between population dynamics and regulatory variation, which creates unpredictable relationships between the effects of variants of gene expression and that of gene expression on the risk. As it becomes increasingly common to manipulate the genome in targeted ways in order to understand the biology behind disease risk, understanding this interplay is increasingly important. At the same time however this opens the possibility that the small effects of variants on risk might be a gross underestimate of the effects of gene expression on the risk, which might open new possibilities for significant interventions.

MATERIALS AND METHODS

Datasets

Psychiatric Genomics Consortium Data

Summary data from Psychiatric Genomics Consortium (PGC) (Consortium 2014) were downloaded from https://www.med.unc.edu/pgc and a p-value threshold of 10^{-6} was used to select variants for our analysis, a relaxed threshold as this analysis is not meant to nominate SZ risk variants but rather to test our mediation and direction hypotheses. Due to the high linkage disequilibrium (LD) of the major histocompatibility complex region, chromosome 6 was excluded leaving 13,197 SNPs meeting criteria. These included multiple SNPs per locus, often in high LD with each other. To be inclusive, rather than testing only the lead SNP, that might not be the driver of the association we

analyzed all SNPs showing association with disease at our chosen threshold. These tests are highly correlated due to LD, so the independent statistical tests performed are much less than the number of SNPs but more than the number of loci, which needs to be accounted for in the interpretation of results. To make the results easier to interpret we grouped SNPs in 219 independent groups. These were defined as groups between which no two SNPs were correlated at r^2 >0.2 (supplementary table 4), though the correlations within groups was generally much higher. As we will describe later, in order to correct significance levels since each group is still more than one test we performed permutations to calculate the study-wide expected number of positives under the null hypothesis and the standard deviation (SD) of this number. Finally, we consider two SNP groups to be independent signals at the same locus if their closest SNPs are less than 1 Mb from each other.

CMC Data

We downloaded the QCed and normalized expression data and imputed genotype data from the CMC (https://www.synapse.org/#!Synapse:syn2759792/wiki/69613). Details on the generation of this data can be found in the group's published work (Fromer et al. 2016). In brief this dataset contains the results of RNA sequencing data from postmortem human dorsolateral prefrontal cortex, genotyped on the Illumina Infinium HumanOmniExpressExome chip and imputed to the 1,000 Genomes Phase 1 reference panel. In the dataset we used the RNA data had been adjusted by the investigators' known covariates (Institution, Sex, Age at Death, PMI, RIN & RIN squared, ancestry, and one clustered experimental variable - see original publication (Fromer et al. 2016)) and hidden covariates, generated by surrogate variable analysis, using linear regression. We kept for analysis 258 individuals with diagnosis of "Schizophrenia" and 279 "Controls" with genotype and expression data. The cases included 214 Caucasians, 38

African Americans, 5 Hispanics, 1 Asian, and the controls 212 Caucasians, 45 African Americans, 18 Hispanics, 3 Asian, 1 Multiracial, respectively. For consistency with the BSC data, we removed SNPs with: genotyping rate < 0.90, minor allele frequency < 0.05, Hardy-Weinberg P value < 10^{-6}, and also removed all SNPs with multi-character allele codes or with single-character allele codes outside of A, C, T, G, or missing code using PLINK 1.9. We then extracted the genotypes of SNPs that were also present in the PGC data with consistent alleles. Finally, we excluded all strand-ambiguous SNPs (genotypes G/C or A/T). In total 9536 SNPs and 16311 genes were included in the eQTL analysis.

BrainSeq Consortium Data

The pre-imputed and QCed genotype data and the non-QCed human dorsolateral prefrontal cortex expression data of the BSC (Jaffe et al. 2018) were provided to us by Dr. Andrew Jaffee. Genotyping of postmortem tissue in this cohort was performed using the Illumina HumanHap650Y_V3, Human 1M-Duo_V3, and Omni5 chips, followed by imputation on the 1,000 Genomes Phase 3 reference set. We kept SNPs with genotyping rate > 0.90, minor allele frequency > 0.05, Hardy-Weinberg P value > $1*10^{-6}$. The Poly(A)+ RNA sequencing was performed by the original investigators using Illumina HiSeq 2000 with two hundred bp paired-end sequencing. Reads were mapped to the human genome hg19 using TopHat 2.0.4. Similar to the processing of the CMC expression data, we removed samples with RIN < 5.5. All samples in the dataset had read numbers exceeding 70 million. Reads had been normalized and transferred to log2CPM using the voom function in limma (https://www.rdocumentation.org/packages/limma/versions/3.28.14). Genes with less than 1 CPM for more than half of the samples were considered not expressed and removed as in the CMC. To identify outliers we converted raw reads to FPKM and used

hierarchical clustering to identify any sample(s) that clustered separately from the rest. We identified and removed 33 outlier samples. The R package Supervised Normalization of Microarrays (SNM) (Mecham et al. 2010) was used to adjust by known, (Sex, Age at Death, PMI, RIN & RIN squared, ancestry) and hidden covariates, generated by surrogate variable analysis (as implemented in http://bioconductor.org/packages/release/bioc/html/sva.html). To be consistent with the CMC dataset, we excluded samples with age <17 yr. We also excluded SNPs absent or reported to have different alleles than listed on the PGC file. Ambiguous SNPs were also removed. Finally, 9386 SNPs, and 9401 genes were included in the eQTL analysis. We kept 345 individuals, 151 cases and 194 controls. The cases included 83 Caucasian and 68 African American, and the controls 86 Caucasians, 108 African Americans.

eQTL analysis

We defined as Cis-eQTL analysis the analysis of SNP - gene pairs closer than 500 KB and as Trans- that of pairs at greater distance. In the trans-eQTL analysis we included all expressed genes but only SNPs in the schizophrenia loci. We performed all eQTL analyses with the matrixEQTL package using linear models (Shabalin 2012).

We performed eQTL analysis to select SNP-gene pairs that we would include to our downstream mediation analysis. Because we recognize that the top SZ associated SNP is not necessarily the one best capturing the effect on expression which may involve more than one variant, relatively loose thresholds were used to define significance in this analysis. For cis- and trans-eQTL, $p < 0.01$ and $p < 1*10^{-7}$ were defined as significant for proceeding to the next steps respectively. Our analysis showed that with these P

value thresholds, we achieved FDR near 5% in both cases (CMC: cis 6.0%, trans 3.7%; BSC: cis 8.7%, trans 6.0%). Both cis- and trans- SNP-gene pairs that passed the thresholds were included in the mediation analysis.

Mediation analysis

Only significant SNP-gene pairs in the eQTL analysis as defined above were included in the mediation analysis. The Python package PyProcessMacro 0.9.7 (https://github.com/QuentinAndre/pyprocessmacro) with a two-step linear regression model was applied for the analysis where we considered one mediator at a time for simplicity. The 1st regression step was: $M = ß_1X + b + e$; the 2nd regression was: $Y = ß_2X + ß_3M + b + e$ (were X: genotype, M: Mediator Gene expression, Y: Phenotype, ß: corresponding Coefficients, b: Intercept, e: Error). $ß_2$ is the direct effect of the SNP on the phenotype. The indirect effect of a SNP on the phenotype through the mediator is the product of $ß_1$ and $ß_3$. The total effect is the sum of the indirect and the direct effects. PyProcessMacro calculates and provides 95% CI for the indirect (mediated) effect by 5000 bootstraps to test its statistical significance. If these confidence intervals do not cross 0 there is significant mediation with $p<0.05$. Unfortunately the package does not allow for the calculation of more stringent CI. To further reduce the false positive rate, we also extended the calculated CI to 1.5 times the original in both directions. The formula we used for intervals from A to B is: $C = (A+B)/2$, $newA = C - ((B-A)/4)*3$, $newB = C + ((B-A)/4)*3$. The same process was repeated on every qualified SNP-gene pair.

Simulations and permutations

To ensure that the mediation analysis can correctly assign the direction of the effects, we analyzed simulated data where the effect of one independent variable I on outcome O was mediated by multiple mediators M in varying directions. We simulated I variable data following a standard normal distribution. Gene expression data was created based on the I data with coefficients in different directions. Then O data was constructed from either single M or combination of two M with various coefficients. We repeated the same eQTL and mediation analysis for each I-M-O combination. The combinations are shown in Supplementary Table 2

To test the validity of the calculated mediation significance we permuted either both of the gene expression and phenotype data, or just the phenotype data, and repeated the same mediation analysis. Due to computational burden we only performed enough permutations to show that the number of positives we observe is far smaller than with the original data and to calculate a FDR. For the permutations of both gene expression and phenotype 10 runs were enough to show us that eliminating the link between genotype and expression was dramatically reducing the number of positives and therefore was not the appropriate approach. For the permutations of only phenotype data we performed 100 runs.

DATA ACCESS

All data used in this work was already publicly available. No new data was generated.

ACKNOWLEDGEMENTS

This work was supported by NIH grant R01MH113215. The corresponding author is also supported by P50MH094268 and R01MH106522.

DISCLOSURE DECLARATIONS

The authors have no conflicts of interest or other disclosures

BIBLIOGRAPHY

Akerborg O, Spalinskas R, Pradhananga S, Anil A, Hojer P, Poujade FA, Folkersen L, Eriksson PP, Sahlen P. 2019. High-Resolution Regulatory Maps Connect Vascular Risk Variants to Disease-Related Pathways. *Circ Genom Precis Med* **12**: e002353.

Cardno AG, Marshall EJ, Coid B, Macdonald AM, Ribchester TR, Davies NJ, Venturi P, Jones LA, Lewis SW, Sham PC et al. 1999. Heritability estimates for psychotic disorders: the Maudsley twin psychosis series. *Arch Gen Psychiatry* **56**: 162-168.

Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. 2009. Mapping complex disease traits with global gene expression. *Nat Rev Genet* **10**: 184-194.

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. *Nat Neurosci* **19**: 1442-1453.

Gilad Y, Rifkin SA, Pritchard JK. 2008. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet* **24**: 408-415.

Hill MJ, Jeffries AR, Dobson RJ, Price J, Bray NJ. 2012. Knockdown of the psychosis susceptibility gene ZNF804A alters expression of genes involved in cell adhesion. *Hum Mol Genet* **21**: 1018-1024.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**: 9362-9367.

Inloes JM, Hsu KL, Dix MM, Viader A, Masuda K, Takei T, Wood MR, Cravatt BF. 2014. The hereditary spastic paraplegia-related enzyme DDHD2 is a principal brain triglyceride lipase. *Proc Natl Acad Sci U S A* **111**: 14924-14929.

Jaffe AE, Straub RE, Shin JH, Tao R, Gao Y, Collado-Torres L, Kam-Thong T, Xi HS, Quan J, Chen Q et al. 2018. Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nat Neurosci* **21**: 1117-1125.

Kety SS. 1987. The significance of genetic factors in the etiology of schizophrenia: results from the national study of adoptees in Denmark. *J Psychiatr Res* **21**: 423-429.

Mackinnon DP, Fairchild AJ. 2009. Current Directions in Mediation Analysis. *Curr Dir Psychol Sci* **18**: 16.

MacKinnon DP, Fairchild AJ, Fritz MS. 2007. Mediation analysis. *Annu Rev Psychol* **58**: 593-614.

Majewski J, Pastinen T. 2011. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet* **27**: 72-79.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A et al. 2009. Finding the missing heritability of complex diseases. *Nature* **461**: 747-753.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J et al. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**: 1190-1195.

Mecham BH, Nelson PS, Storey JD. 2010. Supervised normalization of microarrays. *Bioinformatics* **26**: 1308-1315.

Messias EL, Chen CY, Eaton WW. 2007. Epidemiology of schizophrenia: review of findings and myths. *Psychiatr Clin North Am* **30**: 323-338.

O'Brien HE, Hannon E, Hill MJ, Toste CC, Robertson MJ, Morgan JE, McLaughlin G, Lewis CM, Schalkwyk LC, Hall LS et al. 2018. Expression quantitative trait loci in the developing human brain and their enrichment in neuropsychiatric disorders. *Genome Biol* **19**: 194.

Pardinas 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. *Nat Genet* **50**: 381-389.

Power RA, Kyaga S, Uher R, MacCabe JH, Langstrom N, Landen M, McGuffin P, Lewis CM, Lichtenstein P, Svensson AC. 2013. Fecundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings. *JAMA Psychiatry* **70**: 22-30.

Ripke S O'Dushlaine C Chambert K Moran JL Kahler AK Akterin S Bergen SE Collins AL Crowley JJ Fromer M et al. 2013. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet* **45**: 1150-1159.

Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. 2012. Linking disease associations with regulatory information in the human genome. *Genome Res* **22**: 1748-1759.

Schizophrenia GWAS Consortium. 2011. Genome-wide association study identifies five new schizophrenia loci. *Nat Genet* **43**: 969-976.

Schizophrenia Working Group of the Psychiatric Genomics Consortium.. 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**: 421-427.

Schrode N, Ho SM, Yamamuro K, Dobbyn A, Huckins L, Matos MR, Cheng E, Deans PJM, Flaherty E, Barretto N et al. 2019. Synergistic effects of common schizophrenia risk variants. *Nat Genet* **51**: 1475-1485.

Shabalin AA. 2012. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**: 1353-1358.

Tanis SEJ, Jansen P, Zhou H, van Heeringen SJ, Vermeulen M, Kretz M, Mulder KW. 2018. Splicing and Chromatin Factors Jointly Regulate Epidermal Differentiation. *Cell Rep* **25**: 1292-1303.e1295.

Visel A, Rubin EM, Pennacchio LA. 2009. Genomic views of distant-acting enhancers. *Nature* **461**: 199-205.

Xu B, Ionita-Laza I, Roos JL, Boone B, Woodrick S, Sun Y, Levy S, Gogos JA, Karayiorgou M. 2012. De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat Genet* **44**: 1365-1369.

Yang CP, Li X, Wu Y, Shen Q, Zeng Y, Xiong Q, Wei M, Chen C, Liu J, Huo Y et al. 2018. Comprehensive integrative analyses identify GLT8D1 and CSNK2B as schizophrenia risk genes. *Nat Commun* **9**: 838.

Zhao X, Lynch J, xa, G, Chen Q, John Deighton served as e, Gavan Fitzsimons served as associate editor for this a. 2010. Reconsidering Baron and Kenny: Myths and Truths about Mediation Analysis. *Journal of Consumer Research* **37**: 197-206.