Title: A Bayesian Framework for Multiple Trait Colocalization from Summary Association Statistics

Claudia Giambartolomei[1], Jimmy Zhenli Liu[2], Wen Zhang[3], Mads Hauberg[3,4], Huwenbo Shi[5], James Boocock[1], Joe Pickrell[2], Andrew E. Jaffe[6], the CommonMind Consortium[#], Bogdan Pasaniuc*[1], Panos Roussos*[3,7,8]

[1]Department of Pathology and Laboratory Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA; Department of Human Genetics, University of California, Los Angeles, Los Angeles, CA 90095, United States of America.

[2] New York Genome Center, New York, New York, United States of America

[3]Department of Genetics and Genomic Science and Institute for Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, New York, 10029, United States of America.

[4]The Lundbeck Foundation Initiative of Integrative Psychiatric Research (iPSYCH), Aarhus University, Aarhus, 8000, Denmark.

[5]Bioinformatics Interdepartmental Program, University of California, Los Angeles, 90024

[6]Lieber Institute for Brain Development, Johns Hopkins Medical Campus; Departments of Mental Health and Biostatistics, Johns Hopkins Bloomberg School of Public Health Baltimore, MD, 21205, United States of America.

[7]Department of Psychiatry and Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, New York, 10029, United States of America.

[8]Mental Illness Research Education and Clinical Center (MIRECC), James J. Peters VA Medical Center, Bronx, New York, 10468, United States of America.

# The members of the CommonMind Consortium are listed under “Consortia”.

Correspondence:

Dr. Panos Roussos
Icahn School of Medicine at Mount Sinai
Department of Psychiatry and Department of Genetics and Genomic Science and Institute for Multiscale Biology
One Gustave L. Levy Place,
New York, NY, 10029, USA
Panagiotis.roussos@mssm.edu

Dr. Claudia Giambartolomei
University of California, Los Angeles, Los Angeles
Department of Pathology and Laboratory Medicine,
Los Angeles, CA 90095, USA
claudia.giambartolomei@gmail.com

ABSTRACT

Most genetic variants implicated in complex diseases by genome-wide association studies (GWAS) are non-coding, making it challenging to understand the causative genes involved in disease. Integrating external information such as quantitative trait locus (QTL) mapping of molecular traits (e.g., expression, methylation) is a powerful approach to identify the subset of GWAS signals explained by regulatory effects. In particular, expression QTLs (eQTLs) help pinpoint the responsible gene among the GWAS regions that harbor many genes, while methylation QTLs (mQTLs) help identify the epigenetic mechanisms that impact gene expression which in turn affect disease risk. In this work we propose **m**ultiple-trait-c**oloc** (***moloc***), a Bayesian statistical framework that integrates GWAS summary data with multiple molecular QTL data to identify regulatory effects at GWAS risk loci. We applied ***moloc*** to schizophrenia (SCZ) and eQTL/mQTL data derived from human brain tissue and identified 56 candidate genes that influence SCZ through methylation. Our method can be applied to any GWAS and relevant functional data to help prioritize diseases associated genes.

INTRODUCTION

Genome-wide association studies (GWAS) have successfully identified thousands of genetic variants associated with complex diseases[1]. However, since the discovered associations point to non-coding regions, it is difficult to identify the causal genes and the mechanism by which risk variants mediate disease susceptibility. Advancement of high-throughput array and sequencing technology has enabled the identification of quantitative trait loci (QTLs), genetic variants that affect molecular phenotypes such as gene expression (expression QTL or eQTL) and DNA methylation (methylation QTL or mQTL). Integration of molecular QTL data has the potential to functionally characterize the GWAS results. Additionally, analyzing two datasets jointly has been a successful strategy to identify shared genetic variants that affect different molecular processes, in particular eQTL and GWAS integration [2–8,13,19,26,28]. Integrating methylation data[20] could help identify epigenetic regulatory mechanisms that potentially control the identified genes and contribute to disease.

To our knowledge, a statistical approach to integrate multiple QTL datasets with GWAS is lacking. Therefore, we developed **m**ultiple-trait-**coloc** (***moloc***), a statistical method to quantify the evidence in support of a common causal variant at a particular risk region across multiple traits. We applied *moloc* to schizophrenia (SCZ), a complex polygenic psychiatric disorder, using summary statistics from the most recent and largest GWAS by the Psychiatric Genomics Consortium[9], which reported association for 108 independent genomic loci. eQTL data were derived from the CommonMind Consortium[10], which generated the largest eQTL dataset in the dorsolateral prefrontal cortex (DLPFC) from SCZ cases and control subjects (N=467). Finally, we leveraged mQTL data that were previously generated in human DLPFC tissue (N=121) to investigate epigenetic variation in SCZ[11]. Integration of multiple phenotypes helps better characterize the genes predisposing to complex diseases such as SCZ.

MATHERIALS AND METHODS

Method Description

We extended the model of Pickrell[3] and *coloc*[2] to analyze jointly multiple traits. For each variant, we assume a simple linear regression model to relate the vector of phenotypes $\vec{y}$ or a log-odds generalized linear model for the case-control dataset, and the vector of genotypes $\vec{x}$. Under this model the expectation of the trait is:

$$E[y_i] = \beta x_i$$

We define a genomic region containing *Q* variants, for example a *cis* region around expression or methylation probe. We are interested in a situation where summary statistics (effect size estimates and standard errors) are available for all datasets in a genomic region with *Q* variants.

We make two important assumptions. Firstly, that the causal variant is included in the set of *Q* common variants, either directly typed or well imputed. If the causal SNP is not present, the power to detect a common variant will be reduced depending on the LD between other SNPs included in the model and the causal SNP (see *coloc* paper[2]). Secondly, we assume at most one causal variant is present for each trait. In the presence of multiple causal variants per trait, this algorithm is not able to identify colocalization between additional association signals independent from the primary one.

We start by computing a Bayes Factor for each SNP and each of the trait (i.e GWAS, eQTL, mQTL). Using the Wakefield Approximate Bayes factors[12] (WABF), only the variance and effect estimates from regression analysis are needed, as previously described[2,3]. The computation of WABF also includes the shrinkage factor *r*, the ratio of the prior variance *W* (expected effect size under the alternative) and total variance, *r* = *W/(V + W)*. The evidence in support of one of the models with more than 1 trait is:

$$BF^{(m)} = \prod_{i \in m} WABF_i$$

We next estimate the support for possible scenarios in a given genomic region. We call “configuration” a possible combination of *n* set of binary vectors indicating whether the variant is causal for the selected trait, where *n* is the number of traits considered. For instance, if we consider three traits, there can be up to three causal variants and 15 possible configurations of how they are shared among the traits. We combine the Bayes Factor for each configuration with the priors to assess the support for each scenario. For each configuration *S* and observed data *D*, the likelihood of configuration *h* relative to the null (H_0) is given by:

$$\frac{P(H_h \mid D)}{P(H_0 \mid D)} = \sum_{S \in S_h} \frac{P(D \mid S)}{P(D \mid S_0)} \times \frac{P(S)}{P(S_0)} \qquad (1)$$

where, $P(D|S)/P(D|S_0)$ is the Bayes Factor for each configuration, and $P(S)/P(S_0)$ is the prior odds of a configuration compared with the baseline configuration S_0, and the sum is over all configurations which are consistent with a given hypothesis.

The Regional Bayes Factor (RBF) is the Bayes Factor for each configuration combined with the priors to assess the support for each scenario. If priors do not vary across SNPs under the same hypotheses, we can multiply the likelihoods by one common prior. For example, if we let the "." in the subscript denote scenarios supporting different causal variants, the RBF supporting the scenario for one causal variant shared between traits *G* and *E* (equation 1) is:

$$RBF_{GE} = \sum_{i=1}^{Q} \pi^{(1,2)} WABF_i^{(1)} WABF_i^{(2)}$$

where *π* is the prior probability according to how many traits the SNP is associated with, *I* is an indicator that evaluates to 1 if *i,j,* or *i,j,k,* are different and 0 otherwise.

While the RBF summarizing the scenario with one causal variant for traits *G* and *E*, and a different causal variant for trait *M*, is:

$$RBF_{GE.M} = \sum_{i=1}^{Q} \sum_{j=1}^{Q} \pi^{(1,2)} \pi^{(3)} WABF_i^{(1)} WABF_i^{(2)} WABF_j^{(3)} I[i \neq \; j]$$

Notably, the equations for the model with no colocalization can be re-written in terms of the model with colocalization.

For example,

$$RBF_{GE.M} = RBF_{GE} \times RBF_M - \frac{\pi^{(1,2)} \times \pi^{(3)}}{\pi^{(1,2,3)}} \times RBF_{GEM}$$

In general, the model supporting configuration *h* with *n* traits is:

$$\frac{P(H_h|D)}{P(H_0|D)} = \prod_{h \in H} \pi^{(n)} \sum_{j=1}^{Q} WABF_j^{(n)} - \frac{\prod_{h \in H} \pi^{(n)}}{\pi^{(1,2,\ldots H)}} \sum_{j=1}^{Q} \pi^{(1,2,\ldots H)} WABF_j^{(1,2,\ldots h)} \tag{2}$$

where *n* is the number of traits considered, *h* is the configuration of interest out of the *H* possible configurations and *π* is the prior probability according to how many traits the SNP is associated with.

We set the prior probability that a SNP is the causal one for each trait, to be identical ($\pi^{(1)} = \pi^{(2)} = \pi^{(3)}$) and refer to this as p1. We also set the prior probability that a SNP is associated with two traits, to be identical ($\pi^{(1,2)} = \pi^{(1,3)} = \pi^{(2,3)}$) and refer to this as *p2*. We refer to the prior probability that a SNP is causal for all traits as *p3*.

Finally, the posterior probability supporting configuration *h* among *H* possible configurations, is computed:

$$PP_h = P(H_h|D) = \frac{P(H_h|D)}{\sum_{i=0}^{H} P(H_i)} = \frac{\frac{P(H_h|D)}{P(H_0|D)}}{1 + \sum_{i=1}^{H} \frac{P(H_i|D)}{P(H_0|D)}} \qquad (3)$$

GWAS dataset

Summary statistics for genome-wide SNP association with Schizophrenia were obtained from the Psychiatric Genomics Consortium-Schizophrenia Workgroup (PGC-SCZ) primary meta-analysis (35,476 cases and 46,839 controls) [9].

Expression QTL (eQTL) analysis

This analysis used RNA sequence data on individuals of European-ancestry (*N* = 467) from post-mortem DLPFC (Brodmann areas 9 and 46), and imputed genotypes based on the Phase 1 reference panel from the 1,000 Genomes Project as previously described[10]. MatrixEQTL[14] was used to fit an additive linear model between the expression of 15,791 genes and imputed SNP genotypes within a 1 Mb window around the transcription start site for each gene, including covariates for ancestry, diagnosis, and known and hidden variables detected by surrogate variable analysis, as described elsewhere[10]. Overall, the model identified 2,154,331 significant *cis*-eQTL, (i.e., SNP–gene pairs within 1 Mb of a gene) at a false discovery rate (FDR) ≤ 5%, for 13,137 (80%) genes.

Methylation QTL (mQTL) analysis

DNA methylation of postmortem tissue homogenates of the dorsolateral prefrontal cortex (DLPFC, Brodmann areas 9 and 46) from non-psychiatric adult Caucasian control donors (age > 13, N=121) was measured using the Illumina HumanMethylation450 (“450k”) microarray (which measures CpG methylation across 473,058 probes covering 99% of RefSeq gene promoters). DNA for genotyping was obtained from the cerebella of samples with either the Illumina Human Hap 650v3,1M Duo V3, or Omni 5M BeadArrays and merged across the three platforms following imputation to the 1000 Genomes Phase 3 reference panel as previously described[11]. The mQTL analyses was then conducted using the R package MatrixEQTL[14], fitting an additive linear model up to 20kb distance between each SNP and CpG analyzed, including covariates for ancestry and global epigenetic variation.

Moloc Analysis

Previous to running the analyses, the GWAS and eQTL datasets were filtered by poorly imputed SNPs (kept only SNPs with Rsq > 0.3). The Major Histocompatibility

(MHC) region (chr 6: 25 Mb - 35 Mb) was excluded from all co-localization analyses due to the extensive linkage disequilibrium. We applied a genic-centric approach, defined *cis*-regions based on a 50kb upstream/downstream from the start/end of each gene, since our goal is to link risk variants with changes in gene expression. We evaluated all methylation probes overlapping the *cis*-region. The number of cis-regions/methylation pairs is higher than the count of genes because, on average, there are more than one methylation sites per gene. Common SNPs were evaluated in the colocalization analysis for each gene, and each methylation probe, and GWAS. In total, 14,115 *cis*-regions and 534,962 unique *cis*-regions/methylation probes were tested. Genomic regions were analyzed only if 50 SNPs or greater were in common between all the datasets. Across all of the analyses, a posterior probability equal to, or greater than, 80% for each configuration was considered evidence of colocalization.

In order to compare existing method for colocalization of two trait analyses with three traits, we applied *moloc* using the same region definitions, but with two traits instead of three, as well as a previously developed method (coloc[2]). Effect sizes and variances were used as opposed to p-values, as this strategy achieves greater accuracy when working with imputed data[2].

Simulations

We simulated genotypes from sampling with replacement among haplotypes of SNPs with a minor allele frequency of at least 5% found in the phased 1000 Genomes Project within 49 genomic regions that have been associated with type 1 diabetes (T1D) susceptibility loci (excluding the major histocompatibility complex (MHC) as previously described[15]. These represent a range of region sizes and genomic topography that reflect typical GWAS hits in a complex trait. For each trait, two, or three "causal variants" were selected at random, and a Gaussian distributed quantitative trait for which each causal variant SNP explains a specified proportion of the variance was simulated. All analyses were conducted in R.

RESULTS

Overview of the *moloc* method

In this study, we demonstrate the use of ***moloc*** on three traits that have been measured in distinct datasets of unrelated individuals, GWAS (defined as G), eQTL (defined as E) and mQTL (defined as M). If we consider three traits, there can be up to three causal variants and 15 possible scenarios summarizing how the variants are shared among the traits. We can compute a probability of the data under each hypothesis by summing over the relevant configurations. Four examples of configurations are show in **Figure 1**. The "." In the subscript denotes scenarios supporting different causal variants. For instance, GE summarizes the scenario for one causal variant shared between traits GWAS and eQTL (**Figure 1** - Right plot top panel); GE.M summarizes the scenario with one causal variant for traits GWAS and eQTL, and a different causal variant for trait mQTL (**Figure 1** - Left plot bottom panel). We then estimate the evidence in support of different scenarios using equation (3). The algorithm outputs 15 posterior probabilities. We are most interested in the scenarios supporting a shared causal variant for two and three traits, involving the eQTL trait.

Sample size requirements

We explored the posterior probability under different sample sizes. **Figure S1** illustrate the posterior probability distribution across all of the possible scenarios that includes three traits: GWAS, eQTL and mQTL. With a GWAS sample size of 10,000 and eQTL and mQTL sample sizes of 300, the method provides reliable evidence to detect a shared causal variant behind the GWAS and another trait (median posterior probability of any hypothesis >50%). Although in this paper we analyze GWAS, eQTL and mQTL, our method can be applied to any combinations of traits, including 2 GWAS traits and an eQTL dataset. We explored the minimum sample size required when analyzing two GWAS datasets (GWAS1, GWAS2) and one eQTL (**Figure S2**). The method provides reliable evidence for all hypotheses when the two GWAS sample sizes are 10,000 and eQTL sample size reaches 300.

It is instructive to observe where evidence for other hypotheses is distributed. **Figure 2A** illustrates the accuracy of our approach under different scenarios where two or three causal variants are shared. For example, under simulations of one shared variant for GWAS and eQTL and a second variant for mQTL (GE.M), on average 60% of the evidence points to the simulated scenario, while 12% point to GE, 12% to G.E.M and 7.2% to GEM.

We examined whether the inclusion of a third trait increases power to detect colocalization in comparison to running analysis with two traits on the same data. For the colocalization of three traits, we consider any scenarios where there is evidence of colocalization between the GWAS and eQTL datasets, i.e. GE, GE.M, GEM. We note that using only eQTL data recovers fewer colocalizations with GWAS loci when there is truly one single causal variant across the datasets (**Figure 2B**), providing additional support for increasing power by adding mQTLs. In this study we focus on the colocalization of GWAS with eQTL (GEM, GE.M or GE scenarios), due to smaller sample size and limited power in the mQTL dataset.

Choice of priors

The algorithm requires the definition of prior probabilities at the SNP level for the association with one (p1), two (p2), or three traits (p3). We set the priors to $p1 = 1 \times 10^{-4}$, $p2 = 1 \times 10^{-6}$, $p3 = 1 \times 10^{-7}$ based on simulations and exploratory analysis of genome-wide enrichment of GWAS risk variants in eQTLs and mQTLs. We set the prior probability that a variant is associated with one trait as 1×10^{-4} for GWAS, eQTL and mQTL, assuming that each genetic variant is equally likely a priori to affect gene expression or methylation or disease. This estimate has been suggested in the literature for GWAS[16] and used in similar methods[6]. In **Figure S3**, we find eQTLs and mQTLs to be similarly enriched in GWAS, justifying our choice of the same prior probability of association across the two traits. These values are also suggested by a crude approximation of *p2* and *p3* from the common genome-wide significant SNPs across the three dataset.

We varied the prior probability that a variant is associated with all three traits in simulations (**Table S1)**. We find that our choice of priors has good control of false positive rates under the GEM scenario (<1%) and the smallest sum across our scenarios of interests. We ran our real data analyses using different priors, and report our results under the most restrictive set of priors tested (Table S4). We note that our R package implementation allows users to specify a different set of priors.

Co-localization of eQTL, mQTL and risk for Schizophrenia

We applied our method to SCZ GWAS using eQTLs derived from 467 CMC samples and mQTL from 121 individuals. Our aim is to identify the genes important for disease through colocalization of GWAS variants with changes in gene expression and DNA methylation. We analyzed associations genome-wide, and report results both across previously identified GWAS loci, and across potentially novel loci. While we consider all 15 possible scenarios of colocalization, here we focus on gene discovery due to higher power in our eQTL dataset, by considering the combined probabilities of cases where the same variant is shared across all three traits GWAS, eQTLs and mQTLs (GEM > 0.8) or scenarios where SCZ risk loci are shared with eQTL only (GE > 0.8 or GE.M > 0.8) (**Table 1**). We identified 1,173 cis-regions/methylation pairs with posterior probability above 0.8 that are associated with all three traits (GEM), or eQTLs alone (GE or GE.M). These biologically relevant scenarios affect overall 97 unique genes. Fifty-six out of the 97 candidate genes influence SCZ, gene expression and methylation (GEM>=0.8). One possible scenario is that the variants in these genes could be influencing the risk of SCZ through methylation, although other potential interpretations such as pleiotropy should be considered.

Addition of a third trait increases gene discovery

We examined whether moloc with 3 traits enhance power for GWAS and eQTL colocalization compared to using 2 traits. Colocalization analysis of only GWAS and eQTL traits identified 11 genes with GE >= 0.8 and two genes within a previously associated SCZ LD block (*FURIN* and *PCCB*), which indicates a ~9 fold increase in the genes discovered when we consider an additional trait. The 97 genes identified with a high probability of influencing SCZ (GEM, GE, GE.M>=0.8) are listed in Table S3. The 89 additional genes that were found by adding methylation include genes such as *AS3MT* that would have been missed by only GWAS and eQTL colocalization.

Loci overlapping reported SCZ LD blocks

Psychiatric Genomics Consortium (PGC) identified 108 independent loci and annotated LD blocks around these, 104 of which are within non-HLA, autosomal regions of the genome[9]. In **Table 1** we report the number of identified gene-methylation pairs and unique genes under each scenario that overlap one of these previously defined SCZ LD blocks. We examined associations for 79 out of the 104 SCZ LD blocks. We found colocalizations in 22 (or 28%) of the SCZ LD blocks examined with an average gene density per block of 2. 12,856 gene-methylation pairs overlap the SCZ LD regions, and **Figure 3A** illustrates the average distribution of the posteriors across these regions. Cumulatively, 12.3% of the evidence points to shared variation with an eQTL (GE, GE.M and GEM). The majority of the evidence within these regions (62.2%) did not reach support for shared variation across the three traits, with 19% not reaching evidence for association with any traits, and 43.2% with only one of the three traits (35% with GWAS; 6.4% with eQTL, 1.8% with mQTLs). The lack of evidence in these regions could be addressed with greater sample sizes. **Figure 3B** shows the evidence for colocalization of GWAS with eQTL or mQTL across the forty-four candidate genes. We provide illustrative examples of SCZ association with expression and regulatory DNA region in the FURIN locus (**Figure 4** and **Figure S4**).

Potentially novel SCZ loci

We found 53 unique genes in below genome-wide significant regions (novel SCZ associations). All genes were far from a SCZ LD block (more than 50kb, **Table S4**), and contained SNPs with p-values for association with SCZ ranging from 10^{-4} to 10^{-9}. These genes will likely be identified using just the GWAS signal if the sample size is increased. *KCNN3* is among these genes which encodes an integral membrane protein that forms a voltage-independent calcium-activated channel. It regulates neuronal excitability by contributing to the slow component of synaptic afterhyperpolarization[17].

Comparison with previous findings

Our gene discovery analysis replicate several previous results[10,18–20] (**Table 2**). One recent study[20] performed mQTL analysis on 1714 individuals from three independent sample cohorts, and used colocalization between mQTLs and SCZ GWAS to identify genomic regions associated with both schizophrenia and methylation. From their analysis, 32 methylation probes have a posterior probability of colocalizing with SCZ >=0.8. We analyzed 15 out of these methylation probes, and reproduced 7 for colocalization of GWAS and methylation (combined GEM, GM, GM.E >=0.8, cg00585072, cg02951883, cg08607108, cg08772003, cg19624444, cg26732615). Hannon et al.[20] annotated these regions with 26 genes. Since we integrate eQTL information, our analysis points to specific genes responsible for these associations.

Association of gene expression with methylation

DNA methylation is one the best studied epigenetic modifications**.** Methylation can alter gene expression by disrupting transcription factor binding sites (with variable consequences to expression depending on the TF), or by attracting methyl-binding proteins that initiate chromatin compaction and gene silencing. Therefore methylation can be associated with both increased or decreased gene expression[21,22]. Increased CpG methylation in promoter regions is usually associated with silencing of gene expression[23]. However, in genome-wide expression and methylation studies, the correlation of methylation and gene expression is low or the pattern of association is mixed, even for CpG methylation within promoter[22]. One challenge of examining DNA methylation with expression is the uncertainty of linking the CpG site with a specific gene, especially for CpG sites that are distal to any coding genes. To overcome this challenge, we sought to explore direction of effects of methylation and expression, for gene expression and DNA methylation that colocalize with posterior probability above 0.8 (GEM, EM, and G.EM scenarios) (**Table 1**). Overall, we tested 2,227 DNA methylation and gene expression pairwise interactions and found a significant negative correlation between the effect sizes of methylation and expression in the proximity of the transcription start site (**Figure 5**).

DISCUSSION

In this paper, we propose a statistical method for integrating genetic data from molecular quantitative trait loci (QTL) mapping into genome-wide genetic association analysis of complex traits. The proposed approach requires only summary-level statistics and provides evidence of colocalization of their association signals. To our knowledge, a method integrating more than two traits is lacking. In contrast to other methods that attempt to estimate the true genetic correlation between traits such as LD score regression[24] and TWAS[5], *moloc* focuses on genes that are detectable from the datasets at hand. Thus, if the studies are underpowered, most of the evidence will lie in the null scenarios.

We expose one possible application of this approach in SCZ. In this application, we focus on scenarios involving eQTLs and GWAS, alone or in combination with mQTLs. Other scenarios are also biologically important. For example, colocalization of GWAS and mQTL excluding eQTLs (GM.E scenario) could unveil important methylation mechanisms affecting disease but not directly influencing gene expression in *cis*. We report these and other scenarios in our web resource and encourage further examination of these cases in future analyses. The GEM scenario provides evidence that SCZ risk association is mediated through changes in DNA methylation and gene expression. While our method does not detect causal relationships among the associated traits, i.e. whether risk allele leads to changes in gene expression through methylation changes or vice versa, there is evidence supporting the notion that risk alleles might affect transcription factor binding and epigenome regulation that drives downstream alterations in gene expression[21,25].

We provide posterior probabilities supporting respective hypotheses for each gene-methylation pair analyzed, and the SNP for each trait with the highest probability of colocalization with any other trait. For example, the SNP with the highest posterior probability of GWAS colocalization with eQTL or mQTL will be computed from PPA of GE + GE.M + GM + GM.E + GEM. However, the aim of this method is not fine-mapping of SNPs and we encourage researchers to further analyze the identified local associations with methods better suited for fine-mapping.

We assign a prior probability that a SNP is associated with one trait (1×10^{-4}), to two (1×10^{-6}), and to three traits (1×10^{-7}). We have shown with simulations that these are reasonable choices for the particular datasets at hand. Moreover, we prove that eQTL enrichment in GWAS has a similar enrichment to mQTL in GWAS, however the choices are arguable. One solution is to estimate priors for the different combinations of datasets. Pickrell et al.[3] proposed estimation of enrichment parameters from genome-wide results maximizing *a posteriori* estimates for two traits. For multiple traits, another possibility is using deterministic approximation of posteriors[26]. We leave these explorations to future research.

We note that this approach can be extended to more than three traits. However, the number of possible combinations increases exponentially as the number of traits increases, therefore computation time is a limiting factor and realistically it works well for up to four traits. Owing to the increasing availability of summary statistics from multiple datasets, the systematic application of this approach can provide clues into the molecular mechanisms underlying GWAS signals and how regulatory variants influence complex diseases.

CONFLICTS OF INTEREST

None reported.

SUPPLEMENTAL DATA

Supplemental Data include four figures and four tables.

CONSORTIA

The CommonMind Consortium includes: Menachem Fromer, Panos Roussos, Solveig K Sieberts, Jessica S Johnson, Douglas M Ruderfer, Hardik R Shah, Lambertus L Klei, Kristen K Dang, Thanneer M Perumal, Benjamin A Logsdon, Milind C Mahajan, Lara M Mangravite, Hiroyoshi Toyoshiba, Raquel E Gur, Chang-Gyu Hahn, Eric Schadt, David A Lewis, Vahram Haroutunian, Mette A Peters, Barbara K Lipska, Joseph D Buxbaum, Keisuke Hirai, Enrico Domenici, Bernie Devlin, Pamela Sklar

ACKNOWLEDGMENTS

We would like to thank Chris Wallace at the Department of Medical Genetics, NIHR Cambridge Biomedical Research Centre, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK.

This work was supported by the National Institutes of Health (R01AG050986 Roussos and R01MH109677 Roussos), Brain Behavior Research Foundation (20540 Roussos), Alzheimer's Association (NIRG-340998 Roussos) and the Veterans Affairs (Merit grant BX002395 Roussos). Additionally, this work was supported in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai.

Data were generated as part of the CommonMind Consortium supported by funding from Takeda Pharmaceuticals Company Limited, F. Hoffman-La Roche Ltd and NIH grants R01MH085542, R01MH093725, P50MH066392, P50MH080405, R01MH097276, RO1-MH-075916, P50M096891, P50MH084053S1, R37MH057881 and R37MH057881S1, HHSN271201300031C, AG02219, AG05138 and MH06692. Brain tissue for the study was obtained from the following brain bank collections: the Mount Sinai NIH Brain and Tissue Repository, the University of Pennsylvania Alzheimer's Disease Core Center, the University of Pittsburgh NeuroBioBank and Brain and Tissue Repositories and the NIMH Human Brain Collection Core. CMC Leadership: Pamela Sklar, Joseph Buxbaum (Icahn School of Medicine at Mount Sinai), Bernie Devlin, David Lewis (University of Pittsburgh), Raquel Gur, Chang-Gyu Hahn (University of Pennsylvania), Keisuke Hirai, Hiroyoshi Toyoshiba (Takeda Pharmaceuticals Company Limited), Enrico Domenici, Laurent Essioux (F. Hoffman-La Roche Ltd), Lara Mangravite, Mette Peters (Sage Bionetworks), Thomas Lehner, Barbara Lipska (NIMH).

WEB RESOURCES

We developed a web site to visualize the colocalization results of SCZ GWAS, eQTL, mQTLs under all possible scenarios (**icahn.mssm.edu/moloc**). The browser allows searches by gene, methylation probe, and scenario of interest. The *moloc* method is available as an R package from https://github.com/clagiamba/moloc.

FIGURE TITLES AND LEGENDS

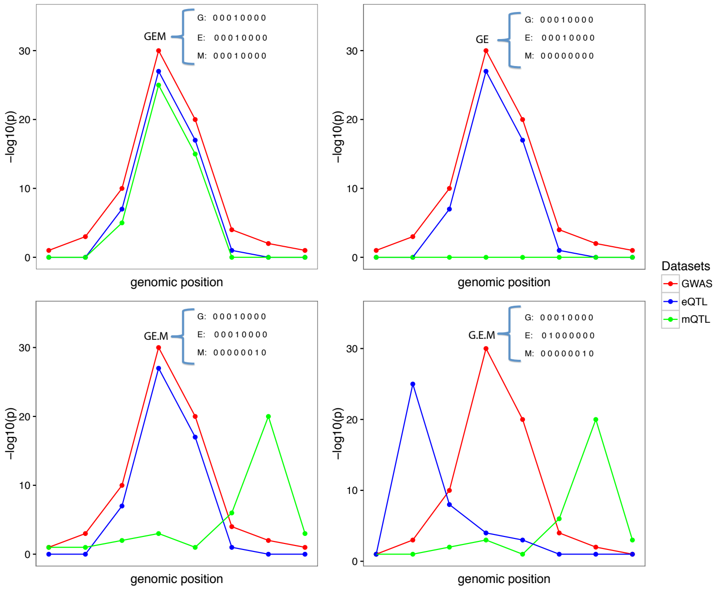

Figure 1. Graphical representation of four possible configurations at a locus with 8 SNPs in common across three traits. The traits are labeled as G, E, M representing GWAS (G), eQTL (E), and mQTL (M) datasets, respectively. Each plot represents one possible configuration, which is a possible combination of 3 sets of binary vectors indicating whether the variant is associated with the selected trait. Left plot top panel (GEM scenario): points to one causal variant behind all of the associations; Right plot top panel (GE scenario): represent the scenario with the same causal variant behind the GE and no association or lack of power for the M association; Left plot bottom panel (GE.M scenario): represents the case with two causal variants, one shared by the G and E, and a different causal variant for M; Right plot bottom panel (G.E.M. scenario): represents the case of three distinct causal variants behind each of the datasets considered.

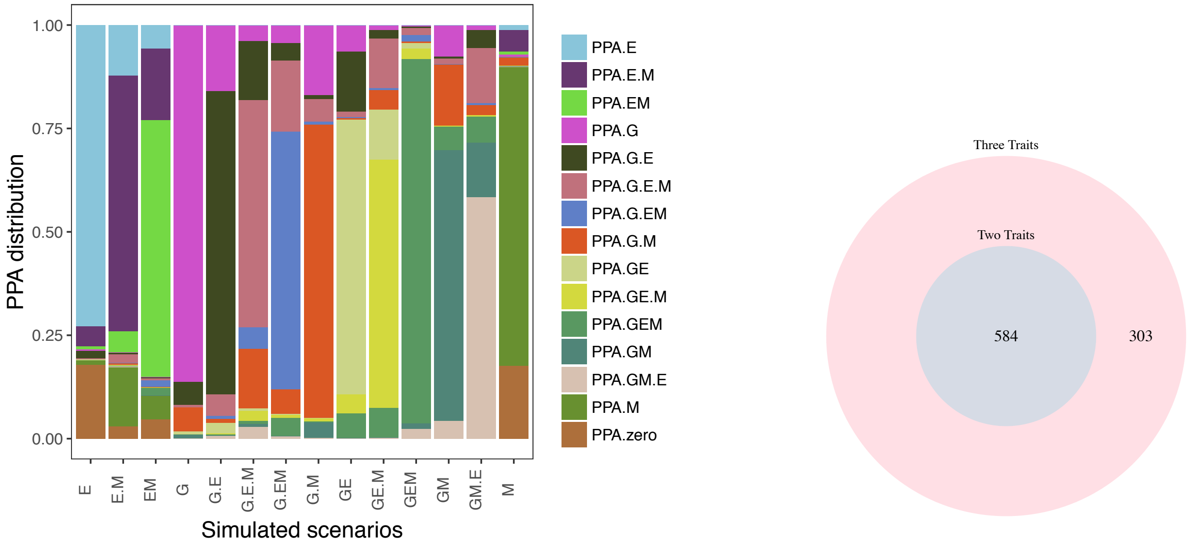

Figure 2. A. Results from simulations under colocalization/non-colocalization scenarios using a sample size of 10,000 individuals for GWAS trait (denoted as G), 300 for eQTL trait (denoted as E), and 300 for mQTL trait (denoted as M). X-axis shows all 15 simulated scenarios, e.g. G.E.M, three different causal variants for each of the three traits; G.EM, 2 different causal variants, one for G and one shared between E and M; GE, 1 shared causal variant for G and E; GE.M, 2 different causal variants, one shared between G and E and one for M; GEM, one causal variant shared between all the three traits. The x-axis shows the distribution of posterior probabilities under the simulated scenario. **B.** Venn diagram comparing number of colocalization of two traits (coloc PPA >=80%) with three traits (moloc PPA GE + GE.M + GEM) in simulations with one causal variant shared between all the three traits (GEM). Results include 887 out of 1,000 simulations passing 80% threshold for colocalization. The variance explained by the trait is 0.01 for GWAS (1%), and 0.1 (10%) for the eQTL and mQTL.

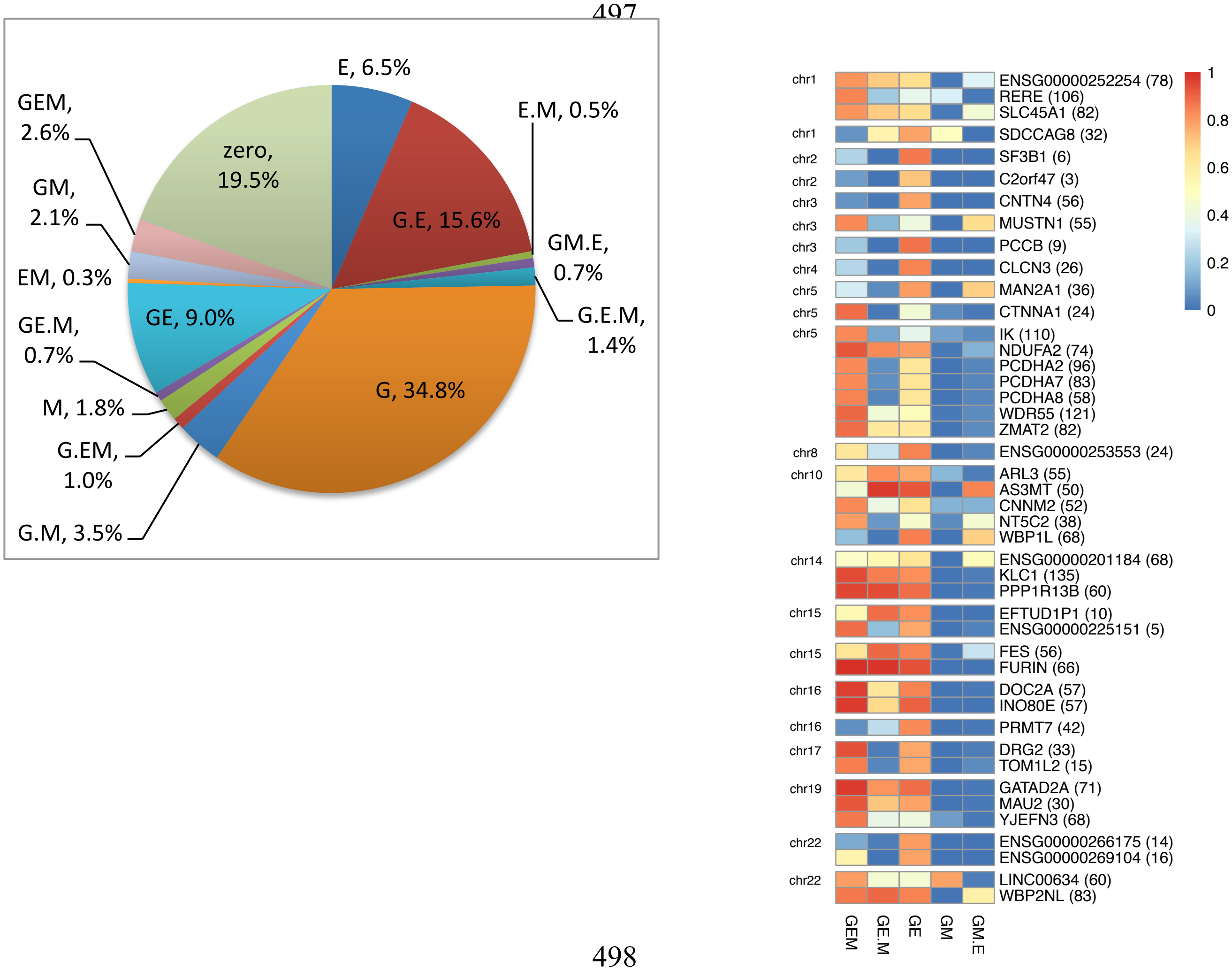

Figure 3. Summary of genes identified using three-trait colocalization within the SCZ LD blocks. **A**. Mean posterior probability for each hypotheses computed using the cis-regions overlapping the SCZ LD blocks. Sections of the pie chart represent the 15 scenarios representing the possible combination of the three traits. The "." between the traits denotes scenarios supporting different causal variants. The combined scenarios GE, GE.M, GE account for 12.3%. **B**. Heatmap displaying the maximum posterior

probabilities reached by the 44 regions overlapping known SCZ LD blocks (gene, number of methylation probes).

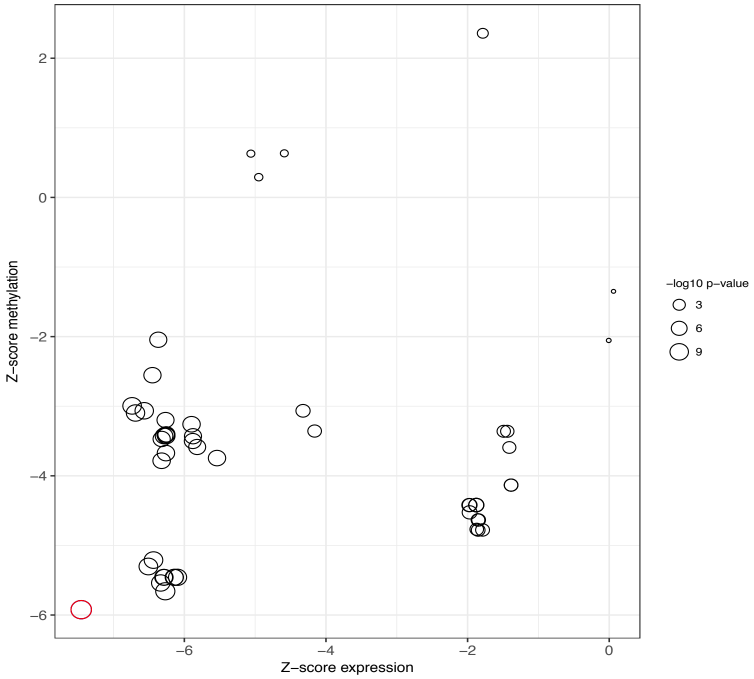

Figure 4. Illustration of one example of colocalization results with GWAS-eQTL-mQTL. *FURIN* gene and cg24888049; Shown are Z-scores (regression coefficients/standard errors) from association of expression (x-axis) and association of methylation (y-axis) at the *FURIN* locus. The red point shows the SNP with the strongest evidence for eQTL, mQTL, GWAS (rs4702).

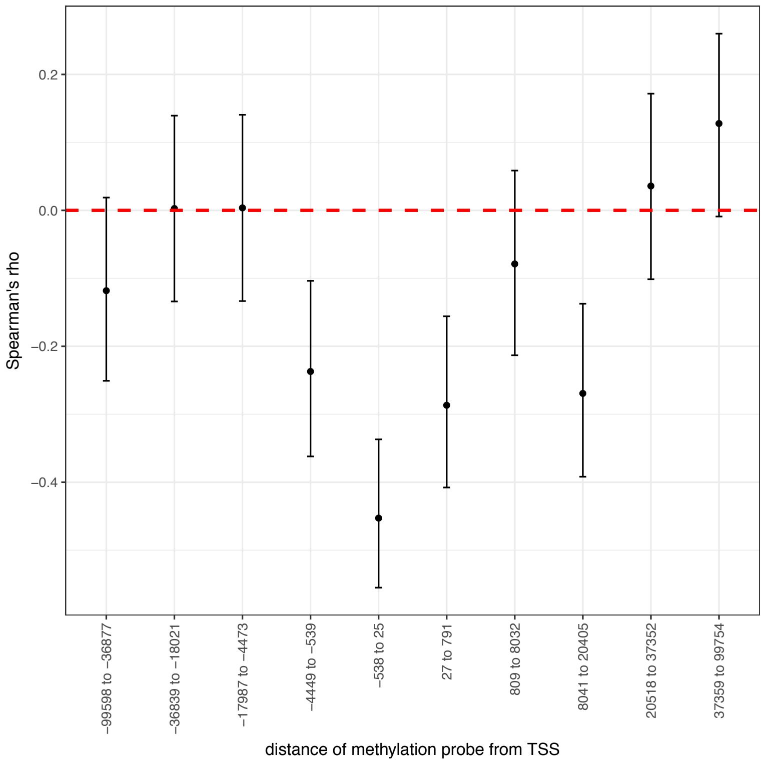

Figure 5. Spearman correlation of eQTL and mQTL effect estimates by distance from transcription start site of the gene. Intervals of methylation probe distance from TSS were estimated based on 10 equal size bins.

TABLE TITLES AND LEGENDS

Table 1. Number of genes with evidence of colocalization (PPA>=0.8) under each scenario.

Scenarios	Sharing of variant	Unique gene-methylation pairs	Unique genes		
		Total PPA>=80%	Total PPA>=80%	Overlapping SCZ LD blocks	Number of LD blocks
Null	No associations	290,850	10,667	92	56
G	GWAS only	4,445	222	149	63
E	eQTL only	116,674	4,427	16	13
M	mQTL only	23,662	6,150	41	28
G.E	GWAS not eQTL (2 causals)	1,501	77	54	27
E.M	eQTL not mQTL (2 causals)	8,713	2,324	7	6
G.M	GWAS not mQTL (2 causals)	241	81	54	26
GE	GWAS,eQTL	389	34	19	15
EM	eQTL,mQTL	1,724	893	3	3
GM	GWAS,mQTL	38	23	18	10
GM.E	GWAS,mQTL not eQTL (2 causals)	21	12	8	5
G.EM	eQTL,mQTL not GWAS (2 causals)	24	12	8	4
GE.M	GWAS,eQTL not mQTL (2 causals)	35	18	10	7
G.E.M	not GWAS not eQTL not mQTL (3 causals)	72	33	26	15
GEM	GWAS,eQTL,mQTL	127	56	27	12
GEM *or* GE.M *or* GE	**combined scenarios for GWAS,eQTL**	1,173	97	44	22
total	total	534,962	14,115	291	78

Table 2. Summary of Previous Findings integrating SCZ GWAS, CMC eQTL and methylation datasets.

Method Used **Scenarios examined in our analysis**	**CMC**[10] **22**	**SMR**[27] **9**	**SMR** [28] GWAS+eQTL: 26	**TWAS**[18] GWAS+eQTL**: 35** GWAS+eQTL+mQTL**: 8**	**COLOC**[20] GWAS+ mQTL: **15**
Validated scenarios (%) at PPA 0.8	**13 (59%)**	**4 (44.4%)**	**22 (85%)**	GWAS+eQTL**: 21 (60%)** GWAS+eQTL+mQTL**: 6 (75%)**	**7 (46%)**
Genes validated	*SF3B1, C2orf47, CNTN4, CLCN3, ENSG00000253553, PPP1R13B, EFTUD1P1, ENSG00000225151, FURIN, INO80E, TOM1L2, DRG2, MAU2, GATAD2A, WBP2NL*	*SF3B1, PCCB, C17ORF39, IRF3*	*AL022476.2, ALMS1P, CLCN3, DOC2A, DRG2, EFTUD1P1, ELAC2, EMB, FAM86B3P, FURIN, GATAD2A, GOLGA2P7, INO80E, JRK, PCCB, PCDHA7, RBBP5, RP11-45P15.4, SF3B1, SLC9B1, SLCO4C1, VPS37A*	**GWAS+eQTL:** *ALMS1P ,C2orf47, CPNE7, DOC2A, DRG2, ELOVL7, EMB, FURIN, GATAD2A, MAU2, MCHR1, NDUFA2, NT5C2, PCCB, PCDHA2, PRMT7, SEPT10, SF3B1, SLC45A1, TMEM81, ZMAT2* **GWAS+eQTL+mQTL:** *SLC45A1,PCCB,NDUFA2,PCDHA2,ZMAT2,PRMT7*	*cg00585072 and PCDHA8, PCDHA2, PCDHA7; cg01262667 and ENSG00000267629; cg02951883 and MAD1L1; cg08607108 and MAD1L1; cg19624444 and MAD1L1; cg08772003* and AS3MT,C10orf32**;** *cg26732615 and GATAD2A, YJEFN3*

REFERENCES

1. Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five Years of GWAS Discovery. Am. J. Hum. Genet. *90*, 7–24.

2. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. *10*, e1004383.

3. Pickrell, J.K., Berisa, T., Liu, J.Z., Ségurel, L., Tung, J.Y., and Hinds, D.A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. Nat. Genet. *48*, 709–717.

4. Pickrell, J.K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am. J. Hum. Genet. *94*, 559–573.

5. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W.J.H., Jansen, R., de Geus, E.J.C., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. *48*, 245–252.

6. Hormozdiari, F., van de Bunt, M., Segrè, A.V., Li, X., Joo, J.W.J., Bilow, M., Sul, J.H., Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes. Am. J. Hum. Genet. *99*, 1245–1260.

7. Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A.L., Kraft, P., and Pasaniuc, B. (2014). Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. PLoS Genet. *10*, e1004722.

8. Consortium, E.N.C.O.D.E.P., and Bernstein, B.E. et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

9. of the Psychiatric Genomics Consortium, S.W.G. (2014). Biological insights from 108

schizophrenia-associated genetic loci. Nature *511*, 421–427.

10. Fromer, M., Roussos, P., Sieberts, S.K., Johnson, J.S., Kavanagh, D.H., Perumal, T.M., Ruderfer, D.M., Oh, E.C., Topol, A., Shah, H.R., et al. (2016). Gene expression elucidates functional impact of polygenic risk for schizophrenia. Nat. Neurosci. *19*, 1442–1453.

11. Jaffe, A.E., Gao, Y., Deep-Soboslay, A., Tao, R., Hyde, T.M., Weinberger, D.R., and Kleinman, J.E. (2016). Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex. Nat. Neurosci. *19*, 40–47.

12. Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with P-values. Genet. Epidemiol. *33*, 79–86.

13. Ip, H., Jansen, R., Abdellaoui, A., Bartels, M., Boomsma, D.I., and Nivard, M.G. (2017). Stratified Linkage Disequilibrium Score Regression reveals enrichment of eQTL effects on complex traits is not tissue specific. bioRxiv.

14. Shabalin, A.A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics *28*, 1353–1358.

15. Wallace, C. (2013). Statistical testing of shared genetic control for potentially related traits. Genet. Epidemiol. *37*, 802–813.

16. Stephens, M., and Balding, D.J. (2009). Bayesian statistical methods for genetic association studies. Nat. Rev. Genet. *10*, 681–690.

17. Deignan, J., Luján, R., Bond, C., Riegel, A., Watanabe, M., Williams, J.T., Maylie, J., and Adelman, J.P. (2012). SK2 and SK3 expression differentially affect firing frequency and precision in dopamine neurons. Neuroscience *217*, 67–76.

18. Gusev, A., Mancuso, N., Finucane, H.K., Reshef, Y., Song, L., Safi, A., Oh, E., O

’donovan, M.C., Katsanis, N., Crawford, G.E., et al. TITLE: Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights.

19. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M., et al. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat. Genet. *48*, 481–487.

20. Hannon, E., Dempster, E., Viana, J., Burrage, J., Smith, A.R., Macdonald, R., St Clair, D., Mustard, C., Breen, G., Therman, S., et al. (2016). An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation. Genome Biol. *17*, 176.

21. Tak, Y.G., and Farnham, P.J. (2015). Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. Epigenetics Chromatin *8*, 57.

22. Wagner, J.R., Busche, S., Ge, B., Kwan, T., Pastinen, T., and Blanchette, M. (2014). The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. Genome Biol. *15*, R37.

23. Du, X., Han, L., Guo, A.-Y., and Zhao, Z. (2012). Features of methylation and gene expression in the promoter-associated CpG islands using human methylome data. Comp. Funct. Genomics *2012*, 598987.

24. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.-R., Duncan, L., Perry, J.R.B., Patterson, N., Robinson, E.B., et al. (2015). An atlas of genetic correlations across human diseases and traits. Nat. Genet. *47*, 1236–1241.

25. Li, Y.I., Van De Geijn, B., Raj, A., Knowles, D.A., Petti, A.A., Golan, D., Gilad, Y., and Pritchard, J.K. RNA splicing is a primary link between genetic variation and disease.

26. Wen, X., Pique-Regi, R., and Luca, F. (2017). Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. PLoS Genet. *13*,.

27. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M., et al. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat. Genet. *48*, 481–487.

28. Hauberg, M.E., Zhang, W., Giambartolomei, C., Franzén, O., Morris, D.L., Vyse, T.J., Ruusalepp, A., Sklar, P., Schadt, E.E., Björkegren, J.L.M., et al. (2017). Large-Scale Identification of Common Trait and Disease Variants Affecting Gene Expression. Am. J. Hum. Genet. *100*, 885–894.