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 52 candidate genes that influence SCZ through methylation. Our method can be applied to any GWAS and relevant functional data to help prioritize disease 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[2–5] and mQTL and GWAS[6–9] integration. All these previous efforts have focused on pairwise trait integration (e.g. eQTL and GWAS or mQTL and GWAS). Analyzing multiple datasets jointly has the potential to increase power for functional characterization of non-coding, risk variants. In addition, it has the potential to link changes in methylation with specific transcripts, avoiding the issues of arbitrary annotating GWAS-mQTL associations to the nearby genes.

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[10], which reported association for 108 independent genomic loci. eQTL data were derived from the CommonMind Consortium[2], 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[6]. Integration of multiple phenotypes helps better characterize the genes predisposing to complex diseases such as SCZ.

MATERIALS AND METHODS

Overview of the Method

We introduce an approach (***moloc***) to detect colocalization among any number of traits concurrently rather than pairwise colocalizatin (i.e. ***coloc***[11]) . We first derive our model using three traits, then generalize to any number of traits. The input of the model is the set of summary statistics derived from three traits measured in distinct datasets of unrelated individuals, {1} GWAS (which we denote as G), {2} eQTL (E) and {3} mQTL (M). 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 the genomic region. 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. Each hypothesis can be represented by a set of index set according to which of the traits each SNP is associated with: H_0- no association to any trait, H_1- association only for trait {1}, H_2- association only for trait {2}, H_3- association only for trait {3}, H_4-

association for traits {1 and 2}, H_5- association for traits {2 and 3}, H_6- association for traits {1 and 3}, H_7- association for trait 1 and 2, but different causal variants for {1} and {2}, H_8- association for traits 2 and 3, but different causal variants for {2} and {3}, H_9- association for traits 1 and 3, but different causal variants for {1} and {3}, H_{10}- association for traits 1, 2 and 3, but different causal variants for {1} and {2,3}, H_{11}- association for traits 1, 2 and 3, but different causal variants for {3} and {1,2}, H_{12}- association for traits 1, 2 and 3, but different causal variants for {2} and {1,3}, H_{13}- association for traits 1, 2 and 3, but different causal variants for {1}, for {2} and for {3}, and H_{14} SNP is associated with all 3 traits (GEM).

Our aim is to find the evidence supporting 15 possible scenarios $(H_0 \cdots H_{14})$, of sharing of SNPs among traits in the given genomic region. We compute the posterior probability of a configuration by weighting the likelihood of the data *D* given a configuration *S*, *P(D|S)*, by the prior probability of a configuration, *P(S)*. We obtain Bayes factors by dividing each likelihood to the baseline likelihood supporting the first model of no association with any trait H_0. The Regional Bayes Factor (RBF) is then the sum over all configurations *Sh* which are consistent with the given hypothesis:

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

where, $P(D|S)/P(D|S_0)$ is the Bayes Factor for each configuration compared to the baseline configuration of no association with any trait S_0, $P(S)/P(S_0)$ is the prior odds of a configuration compared with the baseline configuration S_0, and the sum is over Sh, the set of configurations supporting hypothesis $H_0$to H_{14}. Similar to pairwise

colocalization[11] we then estimate the evidence in support of different scenarios in a given genomic region using the posterior probability supporting hypothesis *h* among *H* possible hypothesis, is computed from:

$$PP_h = P(H_h|D) = \frac{P(H_h|D)}{\sum_{i=0}^{H} P(H_i)} = \frac{\frac{P(H_h|D)}{P(H_0|D)}}{1 + \sum_{i=1}^{H} \frac{P(H_i|D)}{P(H_0|D)}} \qquad (2)$$

Therefore, in our application, the algorithm outputs 15 posterior probabilities. We are most interested in the scenarios supporting a shared causal variant for two and three traits. All the computations are analytical and no iterative methods are required.

We make three important assumptions in *moloc*, the same that are made in our previous *coloc* methodology. 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 linkage disequilibrium (LD) between other SNPs included in the model and the causal SNP (**Figure S5**). 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. Thirdly, as we do not explicitly model LD between SNPs, we assume the samples are drawn from the same ethnic population and therefore have identical allele frequencies and patterns of LD.

Configurations and Notations

To compute the likelihood of the data, we need to first introduce the concepts of configurations and sets of configurations supporting each hypothesis listed above. Each trait at a region of *Q* variants can be described using a binary vector of (0,1) values, where a 1 means that the SNP is associated with the trait and only 1 SNP is non-zero. We call a configuration ***S*** a possible combination of ***M*** binary vectors, an array with dimensions *Q* X *M*, where *M* is the number of traits. The columns represent a particular sharing of any one SNP across the traits. We can combine the configurations into sets corresponding to different scenarios of SNP sharing to assess the evidence in support of each model. Thus, each hypothesis can be represented by a set of index set containing the traits each causal SNP associates with.

To illustrate our notation, consider a region with 8 SNP. 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).

Formally, let Sh denote the set of configurations supporting hypothesis h, then Sh can be written as:

$$S_h = \left\{ \cup_{k=1}^{|h|} \left\{ i_{|h|} \right\} \times h_k \mid \forall\, i_1 ... i_{|h|} \in \{1, ..., Q\}, i_1 \neq \cdots \neq i_{|h|} \right\}, \qquad (3)$$

where hk represents the k-th element (an index set) in the set representation of the hypothesis h.

We can then assess the support for each hypothesis by assigning the possible combinations of causal configurations (binary vectors with at most 1 entry equal to 1) to one of the following 15 hypotheses listed above.

Bayes Factor of a SNP with one trait

We start by computing a Bayes Factor for each SNP and each of the trait (i.e GWAS, eQTL, mQTL). We assume a simple linear regression model to relate the phenotypes or a log-odds generalized linear model for the case-control dataset, and the genotypes. Using the Wakefield Approximate Bayes factors[12] (WABF), only the variance and effect estimates from regression analysis are needed, as shown below and previously described[11,13]:

$$WABF_i^j = \frac{1}{\sqrt{1-r}} \times exp\left[-\frac{Z_i^{2(j)}}{2}r\right] \quad (4)$$

where Zij = β̂/√V is the usual Z statistic and the shrinkage factor r is the ratio of the variance of the prior and total variance (r = W/(V + W)).

The WABF requires specifying the variance W of the normal prior. In the moloc method we set W to 0.15 for a continuous trait and 0.2 for the variance of the log-odds ratio parameter, as previously described [11]. Another possibility is to average over Bayes factors computed with W = 0.01, W = 0.1, and W = 0.5 [14]. We provide this as an option that can be specified by the user.

If the variance of the estimated effect size V is not provided, it can be approximated using the allele frequency of the variant f, the sample size N (and the case control ratio

s for binary outcome). For example, in the case-control setting, V is estimated using the following equation[11]:

$$V_i = Var(\hat{\beta}_i) = \frac{1}{Ns_i(1-s_i) \times 2f_i(1-f_i)} \quad (5)$$

Bayes factor of a SNP across more than one trait

To compute the BF where a SNP i associates with more than one trait, we use:

$$BF_{i,\mathbf{s}} = \prod_{j \in S} WABF_i^j \quad (6)$$

Where **s** is the set of trait indices for which SNP i is associated with.

Note that the computations under >1 trait multiply the individual Bayes Factors together. This is equivalent to the Bayes Factor under the maximum heterogeneity model used in Wen and Stephens[15]. Two key assumptions are necessary for the following computations. Firstly that the traits are measured in unrelated individuals, and secondly that the effect sizes for the two traits are independent. We will consider when these assumptions are appropriate in Discussion.

Prior probabilities that SNP i associates with traits in s

The prior probability that SNP i associates with all traits indexed in a set in our three trait model is: π_φSNP i associates with no trait, with one trait, pairs or traits or all traits $\pi_{\{1,2,3\}}$ such that they sum to 1: $\pi_\varphi + \pi_{\{1\}} + \pi_{\{2\}} + \pi_{\{3\}} + \pi_{\{1,2\}} + \pi_{\{1,3\}} + \pi_{\{2,3\}} + \pi_{\{1,2,3\}} = 1$

Prior Probabilities of each configuration

In our model (Equation 1), *P(S)* is the prior probability of a configuration. We can define these priors from the prior probability that a SNP i associates with traits indexed in π_s (section above). Additionally, since the prior probability *P(S)* of any one configuration in the different sets do not vary across SNPs that belongs to the same set S_h, we can multiply the likelihoods by one common prior supporting the different hypothesis[11]. P(S) depends on a ratio of π_s and on Q, the number of SNPs in the region. Specifically, the prior probability of the configuration associated with no traits is $P_0 = \pi_{\varphi}^{Q}$, with one trait is e.g. $P_1 = Q \times \pi_0^{Q-1}\pi_{\{1\}}$, with a pair of traits with a causal variant shared by trait {1} and {2} is e.g. $P_4 = Q \times \pi_0^{Q-1}\pi_{\{1,2\}}$, with a pair of traits with a different causal variant for trait {1} and {2} is e.g. $P_7 = Q(Q-1) \times \pi_0^{Q-2}\pi_{\{1\}}\pi_{\{2\}}$, and with all three traits, each with a different causal variant is $P_{13} = Q(Q-1)(Q-2) \times \pi_0^{Q-3}\pi_{\{1\}}\pi_{\{2\}}\pi_{\{3\}}$.

Regional Bayes factor (the 3-trait special case)

The Regional Bayes Factor (RBF) can be computed by summing the probability of all the causal configurations consistent with a particular hypothesis (Equation 1).

For example, the RBF for hypothesis 4 is:

- $RBF_4 = \frac{P(H_4|D)}{P(H_0|D)} = \pi_{\{1,2\}} \times \sum_{i=1}^{Q} BF_{i,\{1,2\}} = \pi_{\{1,2\}} \times \sum_{i=1}^{Q} WABF_i^1 WABF_i^2$

For example, the RBF for hypothesis 11 is:

- $RBF_{11} = \frac{P(H_{11}|D)}{P(H_0|D)} = \pi_{\{3\}} \times \pi_{\{1,2\}} \times \sum_{i=1}^{Q}\sum_{j=1}^{Q} BF_{i,\{3\}}BF_{j,\{1\}}BF_{j,\{2\}}$
  $= \pi_{\{3\}} \times \pi_{\{1,2\}} \times \sum_{i=1}^{Q}\sum_{j=1}^{Q} WABF_i^3 WABF_j^1 WABF_j^2 I[i \neq j]$

where WABF are Wakefield's Asymptotic Bayes Factors for each SNP and each trait 1, 2 and 3,I[i ≠ j] is an indicator that evaluates to 1 if SNP i and j are different and to 0 otherwise, π are the prior probabilities that SNP i is the causal SNP under a specific model.

Notably, the RBF for non-colocalized signals [7 . . . n − 1] can be derived from the colocalized signals. This generalization speeds up computation significantly:

- $\frac{P(H_{11}|D)}{P(H_0|D)} = \pi_{\{3\}} \times \pi_{\{1,2\}} \times \sum_{i=1}^{Q} BF_{i,\{3\}} \sum_{j=1}^{Q} BF_{j,\{1,2\}} - \left[\frac{\pi_{\{3\}} \times \pi_{\{1,2\}}}{\pi_{\{1,2,3\}}} \times \frac{P(H_{14}|D)}{P(H_0|D)} \right]$

In practice, we collapsed the prior probabilities to a smaller set for each kind of configuration. We set the prior probability that a SNP is causal in each trait to be identical (π{1} = π{2} = π{3}) and refer to this a p1. We also set the prior probability that is associated with two traits to be identical (π{1,2},π{2,3},π{1,3}) and refer to this as p2. We refer to the prior probability that SNP i the causal for all traits (π{1,2,3}) as p3.

Regional Bayes factor (the general case)

Across a set of M traits {1,2,3...,M}, for a hypothesis h, where there are more than one independent associations among the M traits (i.e. |h| > 1), we can compute the Regional Bayes factor supporting h using a generalization which speeds up computation significantly:

$$RBF_h = \frac{P(H_h|D)}{P(H_0|D)} = \prod_{s \in h} \pi_s \sum_{i=1}^{Q} BF_{i,s} - \frac{\prod_{s \in h} \pi_s}{\pi_t} \sum_{i=1}^{Q} \pi_t BF_{i,t}, \qquad (7)$$

where **s** is the set of trait indices for which SNP i is associated with, and **t** is the union of the index set in ***h***, i.e. $t = \cup_{j=1}^{|h|} h_j$.

For example, let there be three traits {1,2,3} with two independent associations. The first signal colocalizes with traits 1 and 2 while the second is an independent association with trait 3. Hence, M = 3, h = {1,2},{3} and:

$$\frac{P(H_h|D)}{P(H_0|D)} = \pi_{\{1,2\}}\pi_{\{3\}} \sum_{i=1}^{Q} BF_{i\{1,2\}} \sum_{j=1}^{Q} BF_{j\{3\}} - \frac{\pi_{\{1,2\}}\pi_{\{3\}}}{\pi_{\{1,2,3\}}} \sum_{i=1}^{Q} \pi_{\{1,2,3\}} BF_{i\{1,2,3\}}$$

We pre-compute the single terms in the above equation, so that in practice we have:

$$\frac{P(H_h|D)}{P(H_0|D)} = RBF_{\{1,2\}} \times RBF_{\{3\}} - \frac{\pi_{12} \times \pi_3}{\pi_{123}} \times RBF_{\{1,2.3\}}$$

Similarly, let there be five traits {1,2,3,4,5} with two independent associations. First one is common to traits 1, 2, the second one common to traits 4 and 5. Hence, M = 5, h = {{1, 2}, {4, 5}} and:

$$\frac{P(H_h|D)}{P(H_0|D)} = \pi_{\{1,2\}}\pi_{\{4,5\}} \sum_{i=1}^{Q} BF_{i\{1,2\}} \sum_{j=1}^{Q} BF_{j\{4,5\}} - \frac{\pi_{\{1,2\}}\pi_{\{4,5\}}}{\pi_{\{1,2,4,5\}}} \sum_{i=1}^{Q} \pi_{\{1,2,4,5\}} BF_{i\{1,2,4,5\}}$$

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) [10].

Expression QTL (eQTL) analysis

279 For the eQTL associations, we used pre-QC'ed expression and imputed
280 genotypes, on case-control of European-ancestry (N=467) collected by the
281 CommonMind Consortium. Briefly, gene expression was assayed data from post-
282 mortem DLPFC (Brodmann areas 9 and 46) with RNA-seq and mapped to human
283 Ensembl gene reference (v70) using TopHat version 2.0.9 and Bowtie version 2.1.0.
284 Genes with less than 1 CPM (counts per million) in at least 50% of the samples were
285 discarded and RNA-seq data for a total of 16,423 Ensembl genes were considered for
286 downstream analyses. After RNA sample QC, and retaining only genetically-identified
287 European-ancestry individuals, a total of 467 samples were used for downstream
288 analyses. These comprised 209 SCZ cases, 52 AFF (Bipolar, Major depressive
289 disorder, or Mood disorder, unspecified) cases, and 206 controls. The expression data
290 was voom-adjusted for both known covariates (RIN, library batch effect, institution,
291 diagnosis, post-mortem interval, age and sex) and hidden confounds identified based
292 on surrogate variable analysis.

293 Samples were genotyped at 958,178 markers using the Illumina Infinium
294 HumanOmniExpressExome array. Genotyped variants were QC'd, phased, and
295 imputed based on the Phase 1 reference panel from the 1,000 Genomes Project.
296 Variants with INFO < 0.8 or MAF < 0.05 were filtered out, and approximately 6.4 million
297 variants were included in further analysis. Further details on processing of raw data,
298 normalization, and quality control procedures, can be found in [2].

299 The R package MatrixEQTL [16] was used to fit an additive linear model between
300 the expression of 16,423 genes and SNP dosage data at 6.4 million markers within a 1
301 Mb window around the transcription start site for each gene, including covariates for

302 diagnosis, five ancestry covariates and known and hidden variables detected by
303 surrogate variable analysis.
304 Following eQTL identification, summary statistics on 72,441,005 SNPs and 15,791
305 autosomal genes were retained for downstream analyses. Overall, the model identified
306 2,154,331 significant cis-eQTL, (i.e., SNP–gene pairs within 1 Mb of a gene) at a false
307 discovery rate (FDR) ≤ 5%, for 13,137 (80%) genes.
308
309 **Methylation QTL (mQTL) analysis**
310 For the mQTL dataset, we used pre-QC'd DNA methylation data and imputed
311 genotypes on non-psychiatric adult Caucasian control donors (age > 13, N=121). We
312 briefly summarize here the procedures to process the raw data. More details on data
313 processing, normalization, and QC can be found in [6].
314 DNA methylation was assessed on postmortem tissue homogenates of the dorsolateral
315 prefrontal cortex (DLPFC, Brodmann areas 9 and 46) on 534 unique subjects using the
316 Illumina HumanMethylation450 ("450k") microarray (which measures CpG methylation
317 across 473,058 probes covering 99% of RefSeq gene promoters). The average intensity
318 was used to check for low quality samples, intensities from the sex chromosomes were
319 used to drop samples that had predicted sex which was different from its recorded
320 value, and across-sample quantile normalization was used. DNA for genotyping was
321 obtained from the cerebella of samples with either the Illumina Human Hap 650v3, 1M
322 Duo V3, or Omni 5M BeadArrays. Genotypes were called, QC'd, phased and imputed to
323 the 1000 Genomes Phase 3 variant set, separately by genotyping platform. Imputed
324 genotypes were then merged across the three platforms, and SNPs with MAF > 5% and

325 imputation quality above 90% were retained across the samples. Multidimensional
326 scaling (MDS) components were obtained for quantitative measures of ancestry.
327 For the mQTL association analysis, we only considered 121 adult Caucasian
328 individuals (age > 13). The analysis was conducted using the R package MatrixEQTL [16],
329 fitting an additive linear model up to 20kb distance between each SNP and CpG
330 analyzed, including as covariates the first five MDS components from the genetic data
331 and the first 8 PCs (based on the 100,000 most inter-individual variable probes for
332 computational efficiency). Summary statistics on 47,675,913 SNPs and 473,058
333 Illumina 450k probes were used for downstream analyses. Overall, the model identified
334 3,601,681 significant cis-mQTL, (i.e., SNP–probe pairs within 20kb distance of a probe)
335 at a false discovery rate (FDR) ≤ 5%, for 127,237 (27%) probes.
336
337 **Moloc Analysis**
338 The GWAS, eQTL, mQTL datasets were filtered by minor allele frequency
339 greater than 5% and had individually been filtered by imputation quality (Methods).. The
340 Major Histocompatibility (MHC) region (chr 6: 25 Mb - 35 Mb) was excluded from all co-
341 localization analyses due to the extensive linkage disequilibrium. We applied a genic-
342 centric approach, defined *cis*-regions based on a 50kb upstream/downstream from the
343 start/end of each gene, since our goal is to link risk variants with changes in gene
344 expression. We evaluated all methylation probes overlapping the *cis*-region. The
345 number of cis-regions/methylation pairs is higher than the count of genes because, on
346 average, there are more than one methylation sites per gene. Common SNPs were
347 evaluated in the colocalization analysis for each gene, and each methylation probe, and

348 GWAS. In total, 12,003 *cis*-regions and 481,995 unique *cis*-regions/methylation probes
349 were tested. Genomic regions were analyzed only if greater than 50 SNPs were in
350 common between all the datasets. Across all of the analyses, a posterior probability
351 equal to, or greater than, 80% for each configuration was considered evidence of
352 colocalization.

353 In order to compare existing methods for colocalization of two trait analyses with
354 three traits, we applied *moloc* using the same region definitions across the same set of
355 SNPs, using a previously developed method (coloc[11]). Effect sizes and variances were
356 used as opposed to p-values, as this strategy achieves greater accuracy when working
357 with imputed data[11].

358 **Simulations**

359 We simulated genotypes from sampling with replacement among haplotypes of
360 SNPs with a minor allele frequency of at least 5% found in the phased 1000 Genomes
361 Project within 49 genomic regions that have been associated with type 1 diabetes (T1D)
362 susceptibility loci (excluding the major histocompatibility complex (MHC) as previously
363 described[17]. These represent a range of region sizes and genomic topography that
364 reflect typical GWAS hits in a complex trait. For each trait, two, or three "causal
365 variants" were selected at random. We have simulated continuous traits, and assume
366 that causal effect follow a multivariate Gaussian distribution, with each causal variant
367 explaining 0.01 variance of the trait in the GWAS data, and 0.1 in the eQTL and mQTL
368 datasets. For the null scenario, the causal variants explain zero variance of the traits.
369 Each scenario was simulated 1,000 times. To quantify false positive rates on a large

370 number of tests, we simulated the null 500,000 times. We simulated the 15 possible
371 scenarios with different sharing patterns between the GWAS, eQTL, and mQTL
372 datasets. We used sample sizes of 82,315, 467, and 121 individuals to reflect our true
373 sample sizes. We also used different combinations of sample sizes to explore power to
374 detect the correct hypothesis.

375 We estimated the number of false positives within each simulated scenario, by
376 counting the proportion of simulations under the null that passed a posterior probability
377 supporting each of the 14 hypothesis at a particular threshold (PPA>=threshold). We
378 also report the false positives using the sum of the posteriors (PPA.ab + PPA.abc +
379 PPA.abc). The false positive rate is the number of false positives over 1,000
380 simulations. We repeated this procedure using 500,000 simulations under our true
381 sample sizes.

382 We next sought to compare the misclassification rates, and power to detect the
383 correct hypothesis. To compute the number of misclassified calls within each simulated
384 scenario, we counted the proportion of simulations that passed a posterior probability
385 supporting a different hypothesis from the one simulated at a particular threshold
386 (PPA>=threshold). We estimated power to distinguish a particular hypothesis from the
387 others by counting the proportion of simulations that did not reach a posterior probability
388 supporting each scenario at a particular threshold (PPA(true)<threshold).

389 To explore the effect of linkage disequilibrium (LD) on estimated posterior
390 probability, we first computed an LD score for each SNP in the region, defined as the
391 sum of the squared correlation between a SNP and all the SNPs in the region. To
392 assess the degree of LD at a locus we took the average of these scores.

All analyses were conducted in R.

394 **RESULTS**

395 **Sample Size Requirements**

396 We explored false positive rates and the posterior probability under different
397 sample sizes. **Figure S1** illustrate the posterior probability distribution across all of the
398 possible scenarios that includes three traits: GWAS, eQTL and mQTL. With a GWAS
399 sample size of 10,000 and eQTL and mQTL sample sizes of 300, the method provides
400 reliable evidence to detect a shared causal variant behind the GWAS and another trait
401 (median posterior probability of any hypothesis >50%). Although in this paper we
402 analyze GWAS, eQTL and mQTL, our method can be applied to any combinations of
403 traits, including 2 GWAS traits and an eQTL dataset. We explored the minimum sample
404 size required when analyzing two GWAS datasets (GWAS1, GWAS2) and one eQTL
405 (**Figure S2**). The method provides reliable evidence for all hypotheses when the two
406 GWAS sample sizes are 10,000 and eQTL sample size reaches 300. False positive
407 rates are below 0.05 even if a threshold of 0.3 for posteriors is used, and where the
408 causal variant is masked (**Tables S4**, **S5**, **S6**). When samples are greater than 10000
409 for GWAS and 300 for eQTL and mQTL, our misclassification rates are all below 0.05
410 (**Table S8**). Where the causal variant is masked, sample sizes need to reach 10000 for
411 GWAS and 500 for eQTL and mQTL, for our misclassification rates to be below 0.05
412 (**Table S9**). Given the small sample size for the mQTL data, the algorithm has trouble
413 detecting a different causal variant for the mQTL dataset (**Table S7**). For example,
414 evidence pointing to two different causal variants between GWAS and eQTL could be
415 generated by the presence of three causal variants in reality, but the causal variant for

416 mQTL remains undetected. For this reason, we focused on cases with shared casual
417 variants between GWAS, eQTL, with or without mQTL.

418 It is instructive to observe where evidence for other hypotheses is distributed. **Figure 2**
419 illustrates the accuracy of our approach under different scenarios where two or three
420 causal variants are shared. For example, under simulations of one shared variant for
421 GWAS and eQTL and a second variant for mQTL (GE.M), on average 60% of the
422 evidence points to the simulated scenario, while 12% point to GE, 12% to G.E.M and
423 7.2% to GEM.
424

425 **Choice of priors**

426 The algorithm requires the definition of prior probabilities at the SNP level for the
427 association with one (p1), two (p2), or three traits (p3). We set the prior probability that a
428 variant is associated with one trait as 1×10^{-4} for GWAS, eQTL and mQTL, assuming
429 that each genetic variant is equally likely a priori to affect gene expression or
430 methylation or disease. This estimate has been suggested in the literature for GWAS[18]
431 and used in similar methods[19]. We set the priors $p2 = 1 \times 10^{-6}$, $p3 = 1 \times 10^{-7}$ based on
432 sensitivity and exploratory analysis of genome-wide enrichment of GWAS risk variants
433 in eQTLs and mQTLs. In **Figure S3**, we find eQTLs and mQTLs to be similarly enriched
434 in GWAS, justifying our choice of the same prior probability of association across the
435 two traits. These values are also suggested by a crude approximation of *p2* and *p3* from
436 the common genome-wide significant SNPs across the three dataset.

437 We performed sensitivity analyses using different priors. Specifically, we fixed p1 to
438 $1x10^{-4}$ and tested a range of priors for p2 and p3 from $1x10^{-5}$ to $1x10^{-8}$, with increasing
439 difference between p1, p2 and p3. We used a form of internal empirical calibration to
440 compare our prior and posterior expectations. We considered a range of values for p2
441 and p3 while keeping p1 fixed, and find that the posterior expectation of colocalisation
442 most closely resembled the prior expectation under our choice of priors (**Table S1**).
443 We note that our R package implementation allows users to specify a different set of
444 priors.
445
446 **Co-localization of eQTL, mQTL and risk for Schizophrenia**

447 We applied our method to SCZ GWAS using eQTLs derived from 467 samples
448 and mQTL from 121 individuals. Our aim is to identify the genes important for disease
449 through colocalization of GWAS variants with changes in gene expression and DNA
450 methylation. We analyzed associations genome-wide, and report results both across
451 previously identified GWAS loci, and across potentially novel loci. While we consider all
452 15 possible scenarios of colocalization, here we focus on gene discovery due to higher
453 power in our eQTL dataset, by considering the combined probabilities of cases where
454 the same variant is shared across all three traits GWAS, eQTLs and mQTLs (GEM >
455 0.8) or scenarios where SCZ risk loci are shared with eQTL only (GE > 0.8 or GE.M >
456 0.8) (**Table 1**). We identified 1,053 cis-regions/methylation pairs with posterior
457 probability above 0.8 that are associated with all three traits (GEM), or eQTLs alone
458 (GE or GE.M). These biologically relevant scenarios affect overall 84 unique genes and

459 include 39 genes that fall within the previously identified SCZ LD blocks (**Table S2**) and
460 45 potentially novel genes outside of these regions (**Table S3**). Fifty-two out of the 84
461 candidate genes influence SCZ, gene expression and methylation (GEM>=0.8). One
462 possible scenario is that the variants in these genes could be influencing the risk of SCZ
463 through methylation, although other potential interpretations such as pleiotropy should
464 be considered.
465

466 **Addition of a third trait increases gene discovery**

467 We examined whether moloc with 3 traits enhance power for GWAS and eQTL
468 colocalization compared to using 2 traits. In simulations to compare coloc and moloc
469 under one causal variant and our true sample sizes for all three datasets, we observe a
470 fold increase of 1.5 for gene discovery using moloc versus coloc. Moloc with three traits
471 recovers all the genes discovered using *coloc* with eQTL and mQTL, and additional
472 genes from the inclusion of the third layer. In our real data, colocalization analysis of
473 only GWAS and eQTL traits identified 45 genes with a posterior probability, PP4 in
474 *coloc*, of >= 0.8. The 39 additional genes that were found by adding methylation include
475 genes such as *CALN1,* a neuronal transcript associated with abnormalities in
476 sensorimotor gating in humans[20], that would have been missed by only GWAS and
477 eQTL colocalization.
478

479 **Loci overlapping reported SCZ LD blocks**

480 Psychiatric Genomics Consortium (PGC) identified 108 independent loci and
481 annotated LD blocks around these, 104 of which are within non-HLA, autosomal regions

482 of the genome[10]. In **Table 1** we report the number of identified gene-methylation pairs
483 and unique genes under each scenario that overlap the SCZ-associated LD blocks. Out
484 of the 78 SCZ-associated LD blocks we examined in our analysis, we found
485 colocalizations in 20 of them with an average gene density per block of 2.4 unique
486 genes. **Figure 3A** illustrates the average distribution of the posteriors across these
487 regions. Cumulatively, 12% of the evidence points to shared variation with an eQTL
488 (GE, GE.M and GEM). The majority of the evidence within these regions (64%) did not
489 reach support for shared variation across the three traits, with 20% not reaching
490 evidence for association with any traits, and 44% with only one of the three traits (36%
491 with GWAS; 6% with eQTL, 2% with mQTLs). The lack of evidence in these regions
492 could be addressed with greater sample sizes. **Figure 3B** shows the evidence for
493 colocalization of GWAS with eQTL or mQTL across the 39 candidate genes. We
494 provide illustrative examples of SCZ association with expression and DNA methylation
495 in the FURIN locus (**Figure 4** and **Figure S4A**).
496

497 **Potentially novel SCZ loci**

498 We found 45 unique genes that have a high posterior for SCZ and eQTL, but fall
499 in regions not previously identified to be associated with SCZ (at p-value of 5×10^{-8}).
500 All genes were far from a SCZ LD block (more than 150kb, **Table S3**), and contained
501 SNPs with p-values for association with SCZ ranging from 10^{-4} to 10^{-8}. These genes will
502 likely be identified using just the GWAS signal if the sample size is increased. *KCNN3* is
503 among these genes which encodes an integral membrane protein that forms a voltage-
504 independent calcium-activated channel. It regulates neuronal excitability by contributing

505 to the slow component of synaptic afterhyperpolarization[21]. A plot of the associations
506 with the three datasets within this locus is shown in **Figure S4B.**
507
508 **Comparison with previous findings**
509 We compare our gene discovery results to previous studies that assess GWAS-
510 eQTL[2–5] or GWAS-mQTL[5,7–9] colocalization using the same or similar datasets (**Table**
511 **2**). A substantial proportion of genes detected in our study (range 44%-85%) was
512 validated with four studies [2–5] that used eQTL and GWAS integration to prioritize genes
513 important for schizophrenia. Several studies have also linked methylation data with
514 schizophrenia [7–9]. Two recent studies[7,9] used blood mQTL data from 639 samples and
515 identified colocalization of SCZ loci with 32 and 200 methylation probes by applying
516 COLOC and SMR, respectively. A substantial proportion of SCZ-mQTL colocalization
517 was validated in our study (COLOC: 46%; SMR: 18%, **Table 2**). Another study[8] used
518 mQTL data from 166 fetal brain samples and identified 297 methylation probes
519 important for schizophrenia. We analyzed 184 of those and found evidence for 13
520 probes. We note that our methylation data did not included fetal brain samples. Finally,
521 a recent study[5] identified 44 genes involved in schizophrenia through TWAS, followed
522 by integration with chromatin data in blood that resulted in 11 genes associated with
523 GWAS, eQTL and epigenome QTL. We analyzed 8 out of the 11 associations and
524 confirmed 6 of these genes that, in our study, influence SCZ through eQTL and mQTL.
525
526 **Association of gene expression with methylation**
527 We explored direction of effects of methylation and expression, for gene

528 expression and DNA methylation that colocalize (PPA.GEM + PPA.EM + PPA.G.EM >=
529 0.8). Overall, we tested 1,947 DNA methylation and gene expression pairwise
530 interactions and found a significant negative correlation between the effect sizes of
531 methylation and expression in the proximity of the transcription start site (**Figure 5**, p-
532 value:<2.2e-16).

533 **DISCUSSION**

534 In this paper, we propose a statistical method for integrating genetic data from
535 molecular quantitative trait loci (QTL) mapping into genome-wide genetic association
536 analysis of complex traits. The proposed approach requires only summary-level
537 statistics and provides evidence of colocalization of their association signals. To our
538 knowledge, a method integrating more than two traits is lacking. In contrast to other
539 methods that attempt to estimate the true genetic correlation between traits such as LD
540 score regression[22] and TWAS[23], *moloc* focuses on genes that are detectable from the
541 datasets at hand. Thus, if the studies are underpowered, most of the evidence will lie in
542 the null scenarios. We note that our model is the same as *gwas-pw* in Pickrell et al.[13]
543 under specific settings. Precisely, *gwas-pw* averages over Bayes factors computed with
544 W = 0.01, W = 0.1, and W = 0.5 (Methods). We provide this as an option that can be
545 specified by the user. Additionally, *gwas-pw* estimates the prior parameters genome-
546 wide using a maximization procedure. However, we note that, unlike *gwas-pw* that
547 focuses on genome-wide estimation across pairs of traits, our approach focuses on one
548 locus at a time with multiple traits.

549 We expose one possible application of this approach in SCZ. In this application,
550 we focus on scenarios involving eQTLs and GWAS, alone or in combination with
551 mQTLs. Other scenarios are also biologically important. For example, colocalization of
552 GWAS and mQTL excluding eQTLs (GM.E scenario) could unveil important methylation
553 mechanisms affecting disease but not directly influencing gene expression in *cis*. We
554 report these and other scenarios in our web resource and encourage further
555 examination of these cases in future analyses. The GEM scenario provides evidence

556 that SCZ risk association is mediated through changes in DNA methylation and gene
557 expression. While our method does not detect causal relationships among the
558 associated traits, i.e. whether risk allele leads to changes in gene expression through
559 methylation changes or vice versa, there is evidence supporting the notion that risk
560 alleles might affect transcription factor binding and epigenome regulation that drives
561 downstream alterations in gene expression[24,25]. Importantly, our method helps to link a
562 specific gene to a methylation probe. DNA methylation is one the best studied
563 epigenetic modifications. Methylation can alter gene expression by disrupting
564 transcription factor binding sites (with variable consequences to expression depending
565 on the TF), or by attracting methyl-binding proteins that initiate chromatin compaction
566 and gene silencing. Therefore methylation can be associated with both increased or
567 decreased gene expression[24,26]. Increased CpG methylation in promoter regions is
568 usually associated with silencing of gene expression[27]. However, in genome-wide
569 expression and methylation studies, the correlation of methylation and gene expression
570 is low or the pattern of association is mixed, even for CpG methylation within promoter
571 regions[26]. We found a negative correlation between the effect sizes of methylation and
572 expression in the proximity of the transcription start site. However, we note that the
573 current assay (Illumina 450K) cannot discriminate DNA methylation from
574 hydroxymethylation sites. DNA methylation is generally associated with the inhibition of
575 gene expression, while DNA hydroxymethylation, the oxidized form of methylation, has
576 been associated with enhancers and increased gene expression.

577 We provide posterior probabilities supporting respective hypotheses for each
578 gene-methylation pair analyzed, and the SNP for each trait with the highest probability

579 of colocalization with any other trait. For example, the SNP with the highest posterior
580 probability of GWAS colocalization with eQTL or mQTL will be computed from PPA of
581 GE + GE.M + GM + GM.E + GEM. However, the aim of this method is not fine-mapping
582 of SNPs and we encourage researchers to further analyze the identified local
583 associations with methods better suited for fine-mapping.

584 We assign a prior probability that a SNP is associated with one trait (1×10^{-4}), to
585 two (1×10^{-6}), and to three traits (1×10^{-7}). We find support for our choice of priors in the
586 data using two methods. The first uses stratified QQ plots (Figure S3). We find that
587 eQTL enrichment in GWAS has a similar enrichment to mQTL in GWAS. The second is
588 a form of empirical calibration as in Guo et al.[28]. We find that the prior and posterior
589 expectations of colocalisation matched more closely under our choice of priors (Figure
590 S6). However the choices for prior beliefs for each hypothesis are always arguable. One
591 could estimate priors for the different combinations of datasets. Pickrell et al.[13] proposed
592 estimation of enrichment parameters from genome-wide results maximizing *a posteriori*
593 estimates for two traits. For multiple traits, another possibility is using deterministic
594 approximation of posteriors[29]. We leave these explorations to future research.

595 The results of this analysis apply to a particular brain region (dorsolateral
596 prefrontal cortex), age period (adulthood), and ancestry (Caucasians only). We justify
597 using functional data from DLPFC to examine the colocalization with schizophrenia
598 variants as it is the most proximal to disease etiopathogenesis tissue [4,5]. In addition, due
599 to the tissue specificity of gene expression regulatory mechanisms, we matched eQTL
600 and mQTL data on the same brain region and age period. The genotypes and gene
601 expression data reported by the CommonMind Consortium (CMC) comprises the largest

602 existing postmortem brain genomic resource from individuals with psychiatric disorders
603 as well as unaffected controls. The mQTL dataset used here is the largest dataset in the
604 adult brain measured in the brain region and age period considered. Although this was
605 out of the scope of this paper, we encourage application of similar methods to functional
606 data derived from different tissues and age periods to elucidate the spatiotemporal
607 regulatory mechanisms of gene expression that are involved in the etiopathogenesis of
608 schizophrenia.

609 The method makes use of LD to weigh out the posteriors for the different models,
610 therefore the statistics (priors and posteriors of configurations) will depend on the
611 pattern of association (LD) and the number of SNPs in the region (Q). Specifically,
612 increased LD in a region with one causal variant in the decreases our chance to detect
613 a shared causal effect (**Figure S5A**). This is because in loci where the extent of LD is
614 strong (a large number of variants with the same strength of association), there is
615 uncertainty on whether we are dealing with a single variant causal for both traits, or two
616 independent variants in the same LD block. Similarly, the chance of the same SNP
617 being causal for all three traits decreases as Q increases (**Figure S5B**).

618 The Bayes Factors under the models of colocalization (for the case of a single
619 variant affecting more than one trait) are computed by multiplying the individual WABFs
620 together, and this assumes independence of the effect sizes for the traits.
621 Independence is appropriate in the case of analyzing different phenotypes, or eQTLs in
622 different tissues. On the other hand, this assumption may be inappropriate in other
623 cases, for example if we were looking at colocalization of a variant for the same trait in
624 two different studies, or the same eQTLs across different tissue type, or the same trait

625 in two different studies. A more general BF computation which includes cases where the
626 effects are correlated has been proposed in [15] and could be used to extend our
627 approach to these situations. Additionally, correlation among the SNP effect sizes may
628 result if the studies contain overlapping individuals. In this case we may overestimate
629 the evidence in favor of colocalization. We could adjust SNP Bayes factors to account
630 for this overlap as in [14]. The datasets we used in the current analysis does not contain
631 overlapping individuals, however we provide the code to use this approach.

632 We note that this approach can be extended to more than three traits. Since the
633 calculations are analytical and no recursive method is used, computation time for a
634 region with 1000 SNPs is less than one second. However, time increases exponentially
635 as number of traits increases. For four traits it is about 3 seconds, for five traits it is
636 greater than 22 minutes.

637 Overall, owing to the increasing availability of summary statistics from multiple
638 datasets, the systematic application of this approach can provide clues into the
639 molecular mechanisms underlying GWAS signals and how regulatory variants influence
640 complex diseases.

641 **CONFLICTS OF INTEREST**

642 None reported.

643 **SUPPLEMENTAL DATA**

644 Supplemental Data include four figures and four tables.

645

646 **CONSORTIA**

647 The CommonMind Consortium includes: Menachem Fromer, Panos Roussos, Solveig K
648 Sieberts, Jessica S Johnson, Douglas M Ruderfer, Hardik R Shah, Lambertus L Klei,
649 Kristen K Dang, Thanneer M Perumal, Benjamin A Logsdon, Milind C Mahajan, Lara M
650 Mangravite, Hiroyoshi Toyoshiba, Raquel E Gur, Chang-Gyu Hahn, Eric Schadt, David
651 A Lewis, Vahram Haroutunian, Mette A Peters, Barbara K Lipska, Joseph D Buxbaum,
652 Keisuke Hirai, Enrico Domenici, Bernie Devlin, Pamela Sklar
653

654 **ACKNOWLEDGMENTS**

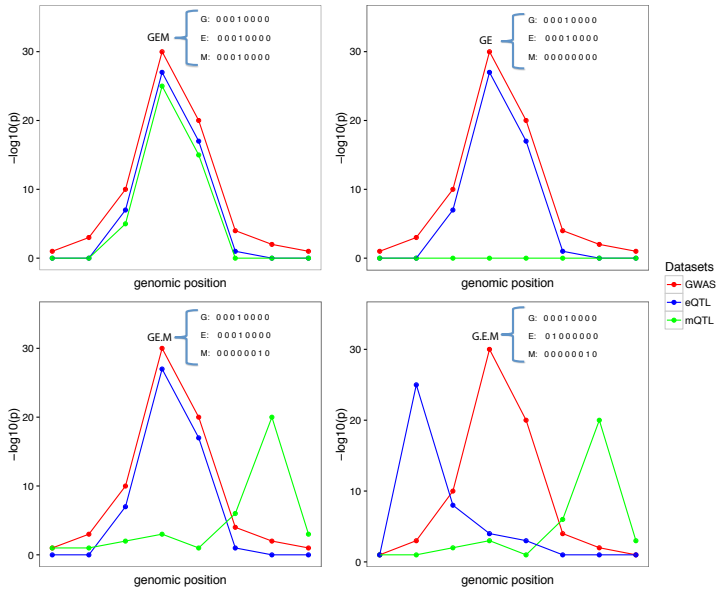
655 We would like to thank Chris Wallace at the Department of Medicine and MRC
656 Biostatistics Unit, Cambridge Biomedical Campus, University of Cambridge, Cambridge,
657 UK.
658 This work was supported by the National Institutes of Health (R01AG050986 Roussos
659 and R01MH109677 Roussos), Brain Behavior Research Foundation (20540 Roussos),
660 Alzheimer's Association (NIRG-340998 Roussos) and the Veterans Affairs (Merit grant
661 BX002395 Roussos). Additionally, this work was supported in part through the
662 computational resources and staff expertise provided by Scientific Computing at the
663 Icahn School of Medicine at Mount Sinai.
664 Data were generated as part of the CommonMind Consortium supported by funding
665 from Takeda Pharmaceuticals Company Limited, F. Hoffman-La Roche Ltd and NIH
666 grants R01MH085542, R01MH093725, P50MH066392, P50MH080405,
667 R01MH097276, RO1-MH-075916, P50M096891, P50MH084053S1, R37MH057881
668 and R37MH057881S1, HHSN271201300031C, AG02219, AG05138 and MH06692.
669 Brain tissue for the study was obtained from the following brain bank collections: the

670 Mount Sinai NIH Brain and Tissue Repository, the University of Pennsylvania
671 Alzheimer's Disease Core Center, the University of Pittsburgh NeuroBioBank and Brain
672 and Tissue Repositories and the NIMH Human Brain Collection Core. CMC Leadership:
673 Pamela Sklar, Joseph Buxbaum (Icahn School of Medicine at Mount Sinai), Bernie
674 Devlin, David Lewis (University of Pittsburgh), Raquel Gur, Chang-Gyu Hahn (University
675 of Pennsylvania), Keisuke Hirai, Hiroyoshi Toyoshiba (Takeda Pharmaceuticals
676 Company Limited), Enrico Domenici, Laurent Essioux (F. Hoffman-La Roche Ltd), Lara
677 Mangravite, Mette Peters (Sage Bionetworks), Thomas Lehner, Barbara Lipska (NIMH).
678

679 **WEB RESOURCES**

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

684 **FIGURE TITLES AND LEGENDS**



685

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

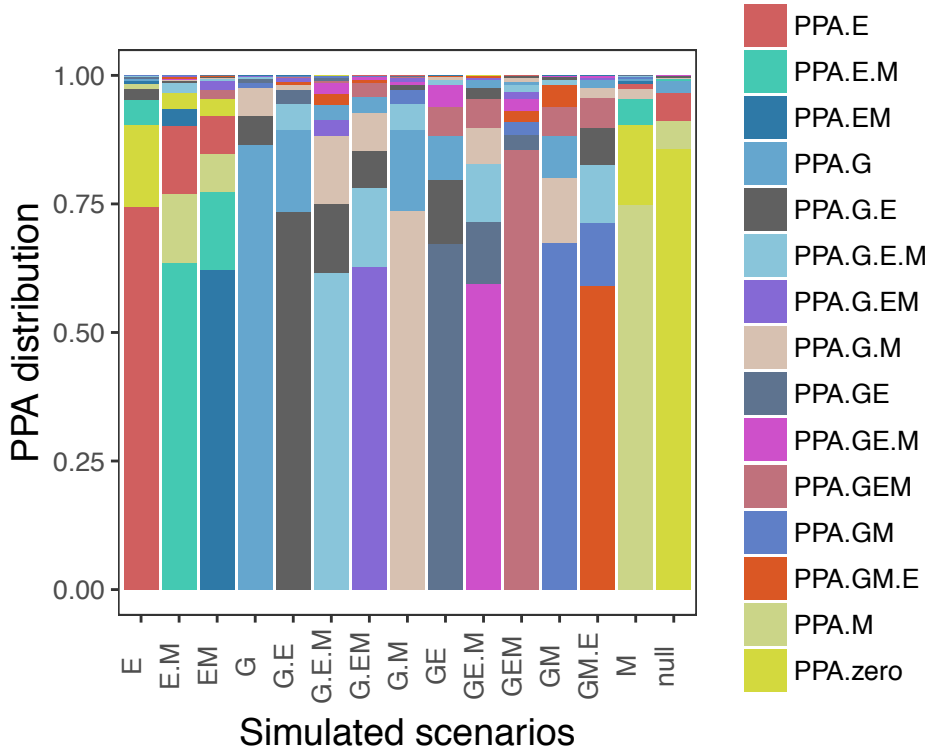
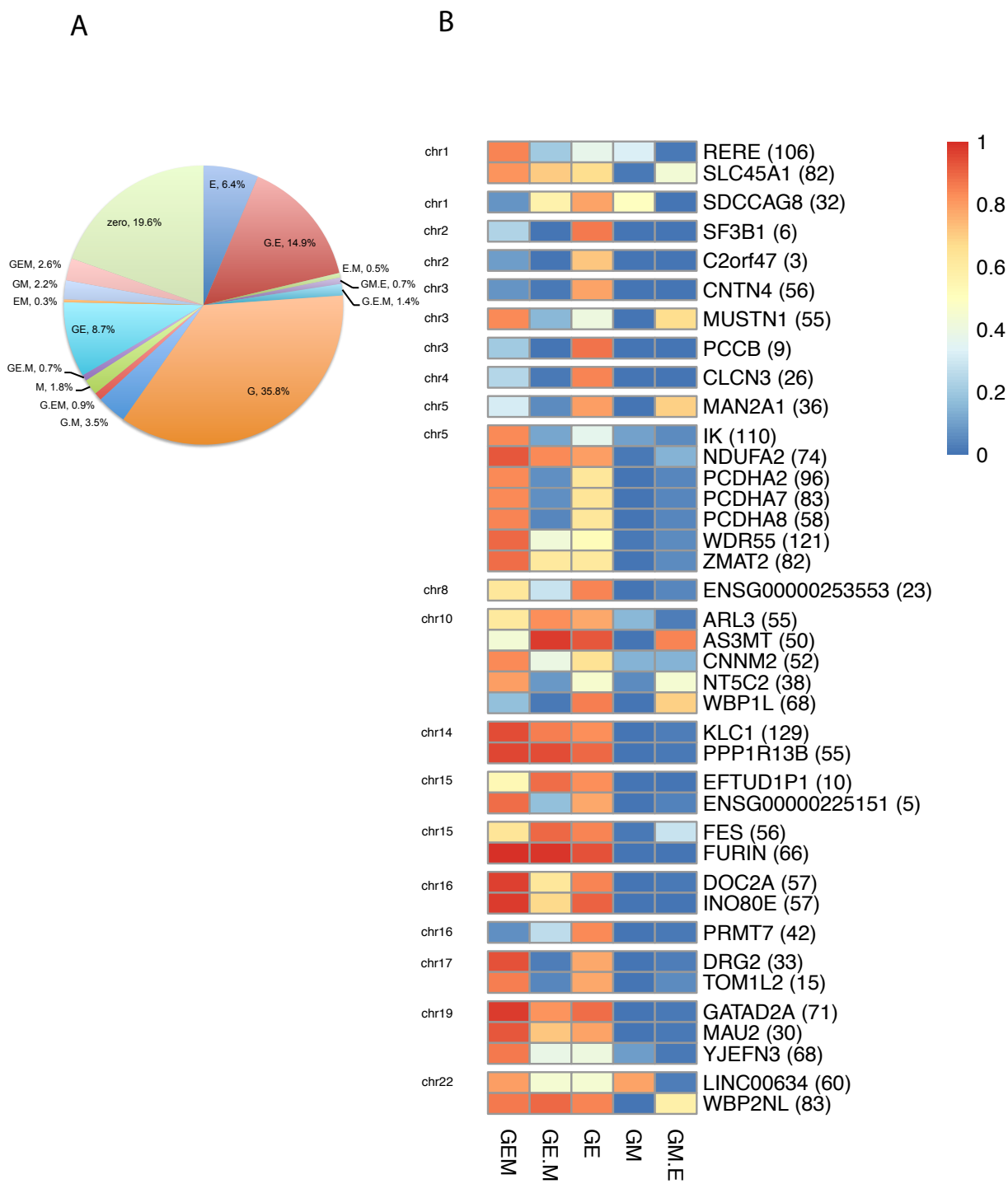
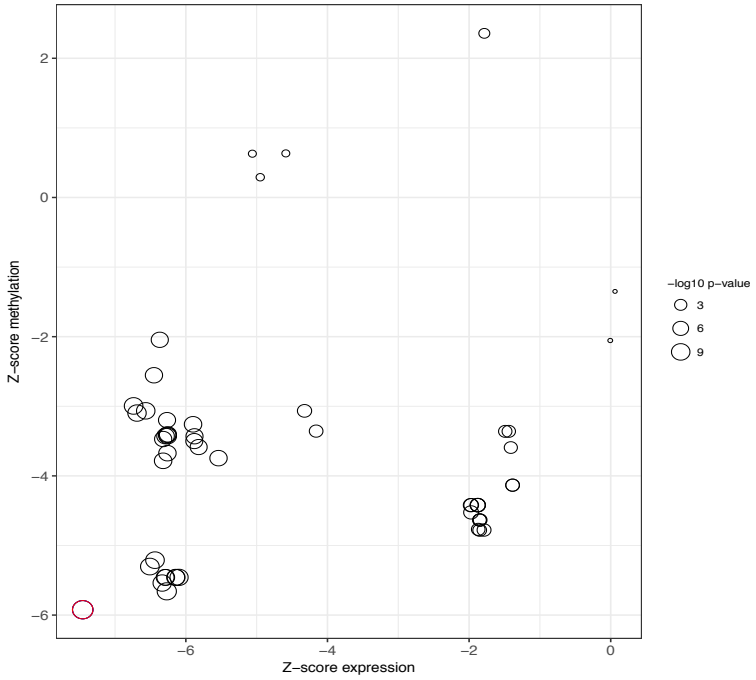


Figure 2. Results from simulations under colocalization/non-colocalization scenarios using a sample size of 10,000 individuals for GWAS trait (denoted as G), 300 for eQTL trait (denoted as E), and 300 for mQTL trait (denoted as M). X-axis shows all 15 simulated scenarios, e.g. G.E.M, three different 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. Y-axis shows the distribution of posterior probabilities under the simulated scenario. The height of the bar represents the mean of the PPA for each configuration across simulations.


A
B
E, 6.4%
G.E, 14.9%
E.M, 0.5%
GM.E, 0.7%
G.E.M, 1.4%
G, 35.8%
G.M, 3.5%
G.EM, 0.9%
M, 1.8%
GE.M, 0.7%
GE, 8.7%
EM, 0.3%
GM, 2.2%
GEM, 2.6%
zero, 19.6%
chr1 RERE (106)
SLC45A1 (82)
chr1 SDCCAG8 (32)
chr2 SF3B1 (6)
chr2 C2orf47 (3)
chr3 CNTN4 (56)
chr3 MUSTN1 (55)
chr3 PCCB (9)
chr4 CLCN3 (26)
chr5 MAN2A1 (36)
chr5 IK (110)
NDUFA2 (74)
PCDHA2 (96)
PCDHA7 (83)
PCDHA8 (58)
WDR55 (121)
ZMAT2 (82)
chr8 ENSG00000253553 (23)
chr10 ARL3 (55)
AS3MT (50)
CNNM2 (52)
NT5C2 (38)
WBP1L (68)
chr14 KLC1 (129)
PPP1R13B (55)
chr15 EFTUD1P1 (10)
ENSG00000225151 (5)
chr15 FES (56)
FURIN (66)
chr16 DOC2A (57)
INO80E (57)
chr16 PRMT7 (42)
chr17 DRG2 (33)
TOM1L2 (15)
chr19 GATAD2A (71)
MAU2 (30)
YJEFN3 (68)
chr22 LINC00634 (60)
WBP2NL (83)
GEM
GE.M
GE
GM
GM.E
1
0.8
0.6
0.4
0.2
0

712 **Figure 3.** Summary of genes identified using three-trait colocalization within the SCZ LD
713 blocks. **A**. Mean posterior probability for each hypotheses computed using the cis-
714 regions overlapping the SCZ LD blocks. Sections of the pie chart represent the 15
715 scenarios representing the possible combination of the three traits. The "." between the
716 traits denotes scenarios supporting different causal variants. The combined scenarios
717 GE, GE.M, GE account for 12.0%. **B**. Heatmap displaying the maximum posterior
718 probabilities reached by the 39 regions overlapping known SCZ LD blocks (gene,
719 number of methylation probes).



720

721 **Figure 4.** Illustration of one example of colocalization results with GWAS-eQTL-mQTL.
722 *FURIN* gene and cg24888049; Shown are Z-scores (regression coefficients/standard
723 errors) from association of expression (x-axis) and association of methylation (y-axis) at
724 the *FURIN* locus. Each point represents a SNP and the size of the dot represent the

level of association using Z-scores. The red point shows the SNP with the strongest evidence for eQTL, mQTL, GWAS (rs4702).

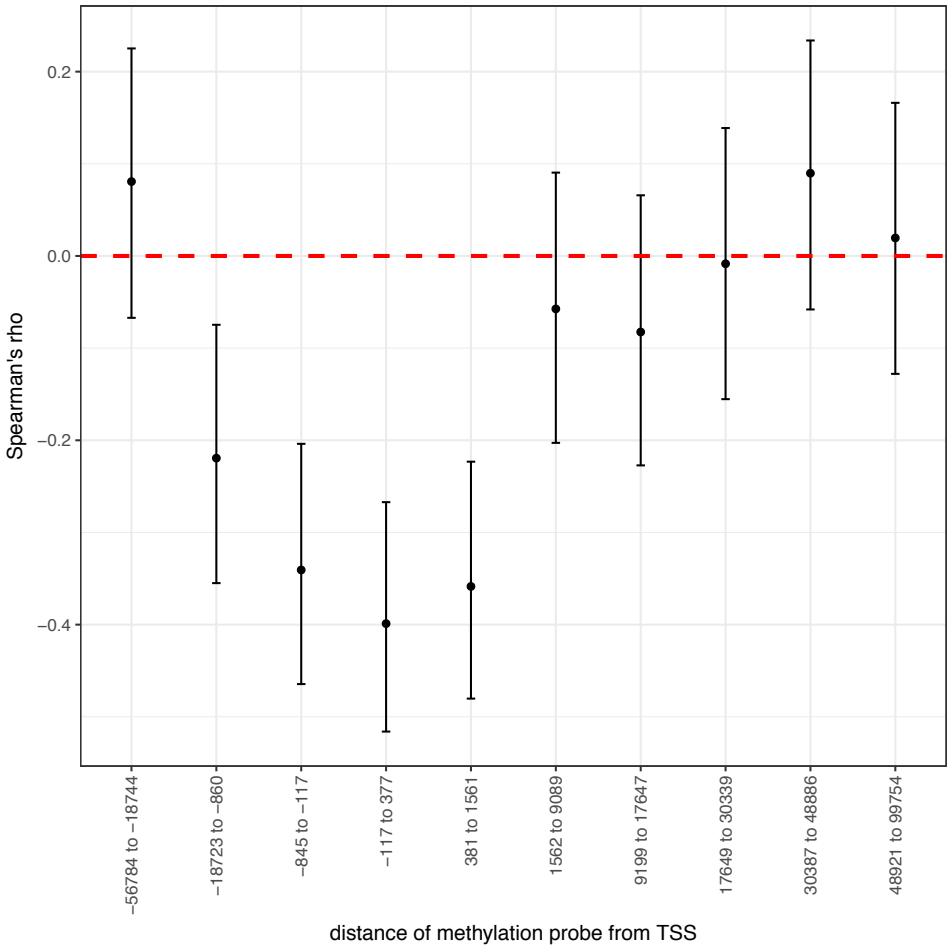


Figure 5. Spearman correlation of eQTL and mQTL effect estimates by distance from transcription start site of the gene. Intervals of methylation probe distance from TSS were estimated based on 10 equal size bins.

734 **TABLE TITLES AND LEGENDS**

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

736 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	268569	9317	90	55
G	GWAS only	4331	215	146	63
E	eQTL only	97852	3597	14	12
M	mQTL only	21914	5588	42	28
G.E	GWAS not eQTL (2 causals)	1257	66	47	27
E.M	eQTL not mQTL (2 causals)	7542	1952	7	6
G.M	GWAS not mQTL (2 causals)	219	75	52	26
GE	GWAS,eQTL	359	30	18	14
EM	eQTL,mQTL	1494	769	3	3
GM	GWAS,mQTL	38	23	18	10
GM.E	GWAS,mQTL not eQTL (2 causals)	20	11	7	5
G.EM	eQTL,mQTL not GWAS (2 causals)	22	10	7	4
GE.M	GWAS,eQTL not mQTL (2 causals)	31	17	10	7
G.E.M	not GWAS not eQTL not mQTL (3 causals)	63	29	23	15
GEM	GWAS,eQTL,mQTL	123	52	25	11
GEM *or* GE.M *or* GE	**combined scenarios for GWAS,eQTL**	1053	84	39	20
total	total	481995	12003	273	78

737 **Table 2**. Summary of Previous Findings integrating SCZ GWAS, CMC eQTL and

738 methylation datasets.

Reference	Fromer et al. [2]	Zhu et al. [3]	Hauberg et al. [4]	Gusev at al. [5]	Hannon et al. [7]	Hannon et al. [8]	Hannon[9] et al.
Method Used	**Sherlock**	**SMR**	**SMR**	**TWAS**	**COLOC**	**COLOC**	**SMR**
Datasets	GWAS SCZ (N=82315); eQTL in brain (N=537)	GWAS SCZ (N=82315); eQTL in blood (N=5311)	GWAS SCZ (N=82315); eQTL in brain (N=467)	GWAS SCZ (N=82315); eQTL in brain (N=621); chromatin QTL in blood (N=45)	GWAS SCZ (N=82315); mQTL in blood (N=639)	GWAS SCZ (N=82315); mQTL in fetal brain (N=166)	GWAS SCZ (N=82315); mQTL in blood (N=639)
Scenarios examined in our analysis	**22**	**9**	26	GWAS+eQTL**: 35** GWAS+eQTL+mQTL**: 8**	**15**	184	49
Validated scenarios (%) at PPA 0.8	**15 (68%)**	**4 (44.4%)**	**22 (85%)**	GWAS+eQTL**: 21 (60%)** GWAS+eQTL+mQTL**: 6 (75%)**	**7 (46%)**	**13 (7%)**	**9 (18%)**
Genes and methylation sites validated*	**GWAS+eQTL***: SF3B1, C2orf47, CNTN4, CLCN3, ENSG00000253553, PPP1R13B, EFTUD1P1, ENSG00000225151, FURIN, INO80E, TOM1L2, DRG2, MAU2, GATAD2A, WBP2NL*	**GWAS+eQTL**: *SF3B1, PCCB, C17ORF39, IRF3*	**GWAS+eQTL***: 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*	**GWAS+mQTL:** *cg00585072, cg01262667, cg02951883, cg08607108, cg08772003, cg19624444, cg26732615*	**GWAS+mQTL***: cg08772003, cg19624444, cg08607108, cg02951883, cg00585072, cg26732615, cg02206767, cg26395211, cg25066665, cg04140906, cg04398451, cg10596483, cg24634471*	**GWAS+mQTL***: cg25066665, cg02951883, cg07143303, cg08772003, cg14047540, cg04140906, cg12863693, cg04398451, cg26732615*

739 * Genes validated (GWAS+eQTL+mQTL and GWAS+eQTL) include the genes with
740 PPA.GEM+PPA.GE+PPA.GE.M>= 80%; methylation sites validated (GWAS+mQTL)
741 include PPA.GEM+PPA.GM+PPA.GM.E>= 80%.

742

743 **REFERENCES**

744 1. Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five Years of
745 GWAS Discovery. Am. J. Hum. Genet. *90*, 7–24.
746 2. Fromer, M., Roussos, P., Sieberts, S.K., Johnson, J.S., Kavanagh, D.H., Perumal,
747 T.M., Ruderfer, D.M., Oh, E.C., Topol, A., Shah, H.R., et al. (2016). Gene expression
748 elucidates functional impact of polygenic risk for schizophrenia. Nat. Neurosci. *19*,
749 1442–1453.
750 3. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery,
751 G.W., Goddard, M.E., Wray, N.R., Visscher, P.M., et al. (2016). Integration of summary
752 data from GWAS and eQTL studies predicts complex trait gene targets. Nat. Genet. *48*,
753 481–487.
754 4. Hauberg, M.E., Zhang, W., Giambartolomei, C., Franzén, O., Morris, D.L., Vyse, T.J.,
755 Ruusalepp, A., Sklar, P., Schadt, E.E., Björkegren, J.L.M., et al. (2017). Large-Scale
756 Identification of Common Trait and Disease Variants Affecting Gene Expression. Am. J.
757 Hum. Genet. *100*, 885–894.
758 5. Gusev, A., Mancuso, N., Finucane, H.K., Reshef, Y., Song, L., Safi, A., Oh, E., O
759 ’donovan, M.C., Katsanis, N., Crawford, G.E., et al. TITLE: Transcriptome-wide
760 association study of schizophrenia and chromatin activity yields mechanistic disease
761 insights.
762 6. Jaffe, A.E., Gao, Y., Deep-Soboslay, A., Tao, R., Hyde, T.M., Weinberger, D.R., and
763 Kleinman, J.E. (2016). Mapping DNA methylation across development, genotype and
764 schizophrenia in the human frontal cortex. Nat. Neurosci. *19*, 40–47.
765 7. Hannon, E., Dempster, E., Viana, J., Burrage, J., Smith, A.R., Macdonald, R., St

766 Clair, D., Mustard, C., Breen, G., Therman, S., et al. (2016). An integrated genetic-
767 epigenetic analysis of schizophrenia: evidence for co-localization of genetic
768 associations and differential DNA methylation. Genome Biol. *17*, 176.
769 8. Hannon, E., Spiers, H., Viana, J., Pidsley, R., Burrage, J., Murphy, T.M., Troakes, C.,
770 Turecki, G., O'Donovan, M.C., Schalkwyk, L.C., et al. (2016). Methylation QTLs in the
771 developing brain and their enrichment in schizophrenia risk loci. Nat. Neurosci. *19*, 48–
772 54.
773 9. Hannon, E., Weedon, M., Bray, N., O'Donovan, M., and Mill, J. (2017). Pleiotropic
774 Effects of Trait-Associated Genetic Variation on DNA Methylation: Utility for Refining
775 GWAS Loci. Am. J. Hum. Genet. *100*, 954–959.
776 10. of the Psychiatric Genomics Consortium, S.W.G. (2014). Biological insights from
777 108 schizophrenia-associated genetic loci. Nature *511*, 421–427.
778 11. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D.,
779 Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of
780 genetic association studies using summary statistics. PLoS Genet. *10*, e1004383.
781 12. Wakefield, J. (2009). Bayes factors for genome-wide association studies:
782 comparison with P-values. Genet. Epidemiol. *33*, 79–86.
783 13. Pickrell, J.K., Berisa, T., Liu, J.Z., Ségurel, L., Tung, J.Y., and Hinds, D.A. (2016).
784 Detection and interpretation of shared genetic influences on 42 human traits. Nat.
785 Genet. *48*, 709–717.
786 14. Pickrell, J.K. (2014). Joint analysis of functional genomic data and genome-wide
787 association studies of 18 human traits. Am. J. Hum. Genet. *94*, 559–573.
788 15. Wen, X., and Stephens, M. (2011). Bayesian methods for genetic association

789 analysis with heterogeneous subgroups: From meta-analyses to gene-environment
790 interactions.
791 16. Shabalin, A.A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix
792 operations. Bioinformatics *28*, 1353–1358.
793 17. Wallace, C. (2013). Statistical testing of shared genetic control for potentially related
794 traits. Genet. Epidemiol. *37*, 802–813.
795 18. Stephens, M., and Balding, D.J. (2009). Bayesian statistical methods for genetic
796 association studies. Nat. Rev. Genet. *10*, 681–690.
797 19. Hormozdiari, F., van de Bunt, M., Segrè, A.V., Li, X., Joo, J.W.J., Bilow, M., Sul,
798 J.H., Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of GWAS
799 and eQTL Signals Detects Target Genes. Am. J. Hum. Genet. *99*, 1245–1260.
800 20. Roussos, P., Giakoumaki, S.G., Zouraraki, C., Fullard, J.F., Karagiorga, V.-E.,
801 Tsapakis, E.-M., Petraki, Z., Siever, L.J., Lencz, T., Malhotra, A., et al. (2016). The
802 Relationship of Common Risk Variants and Polygenic Risk for Schizophrenia to
803 Sensorimotor Gating. Biol. Psychiatry *79*, 988–996.
804 21. Deignan, J., Luján, R., Bond, C., Riegel, A., Watanabe, M., Williams, J.T., Maylie, J.,
805 and Adelman, J.P. (2012). SK2 and SK3 expression differentially affect firing frequency
806 and precision in dopamine neurons. Neuroscience *217*, 67–76.
807 22. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.-R.,
808 Duncan, L., Perry, J.R.B., Patterson, N., Robinson, E.B., et al. (2015). An atlas of
809 genetic correlations across human diseases and traits. Nat. Genet. *47*, 1236–1241.
810 23. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W.J.H., Jansen, R., de
811 Geus, E.J.C., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for

812 large-scale transcriptome-wide association studies. Nat. Genet. *48*, 245–252.
813 24. Tak, Y.G., and Farnham, P.J. (2015). Making sense of GWAS: using epigenomics
814 and genome engineering to understand the functional relevance of SNPs in non-coding
815 regions of the human genome. Epigenetics Chromatin *8*, 57.
816 25. Li, Y.I., Van De Geijn, B., Raj, A., Knowles, D.A., Petti, A.A., Golan, D., Gilad, Y.,
817 and Pritchard, J.K. RNA splicing is a primary link between genetic variation and
818 disease.
819 26. Wagner, J.R., Busche, S., Ge, B., Kwan, T., Pastinen, T., and Blanchette, M.
820 (2014). The relationship between DNA methylation, genetic and expression inter-
821 individual variation in untransformed human fibroblasts. Genome Biol. *15*, R37.
822 27. Du, X., Han, L., Guo, A.-Y., and Zhao, Z. (2012). Features of methylation and gene
823 expression in the promoter-associated CpG islands using human methylome data.
824 Comp. Funct. Genomics *2012*, 598987.
825 28. Guo, H., Fortune, M.D., Burren, O.S., Schofield, E., Todd, J. a, and Wallace, C.
826 (2015). Integration of disease association and eQTL data using a Bayesian
827 colocalisation approach highlights six candidate causal genes in immune-mediated
828 diseases. Hum. Mol. Genet. *24*, 3305–3313.
829 29. Wen, X., Pique-Regi, R., and Luca, F. (2017). Integrating molecular QTL data into
830 genome-wide genetic association analysis: Probabilistic assessment of enrichment and
831 colocalization. PLoS Genet. *13*.
832

Supplemental Figures

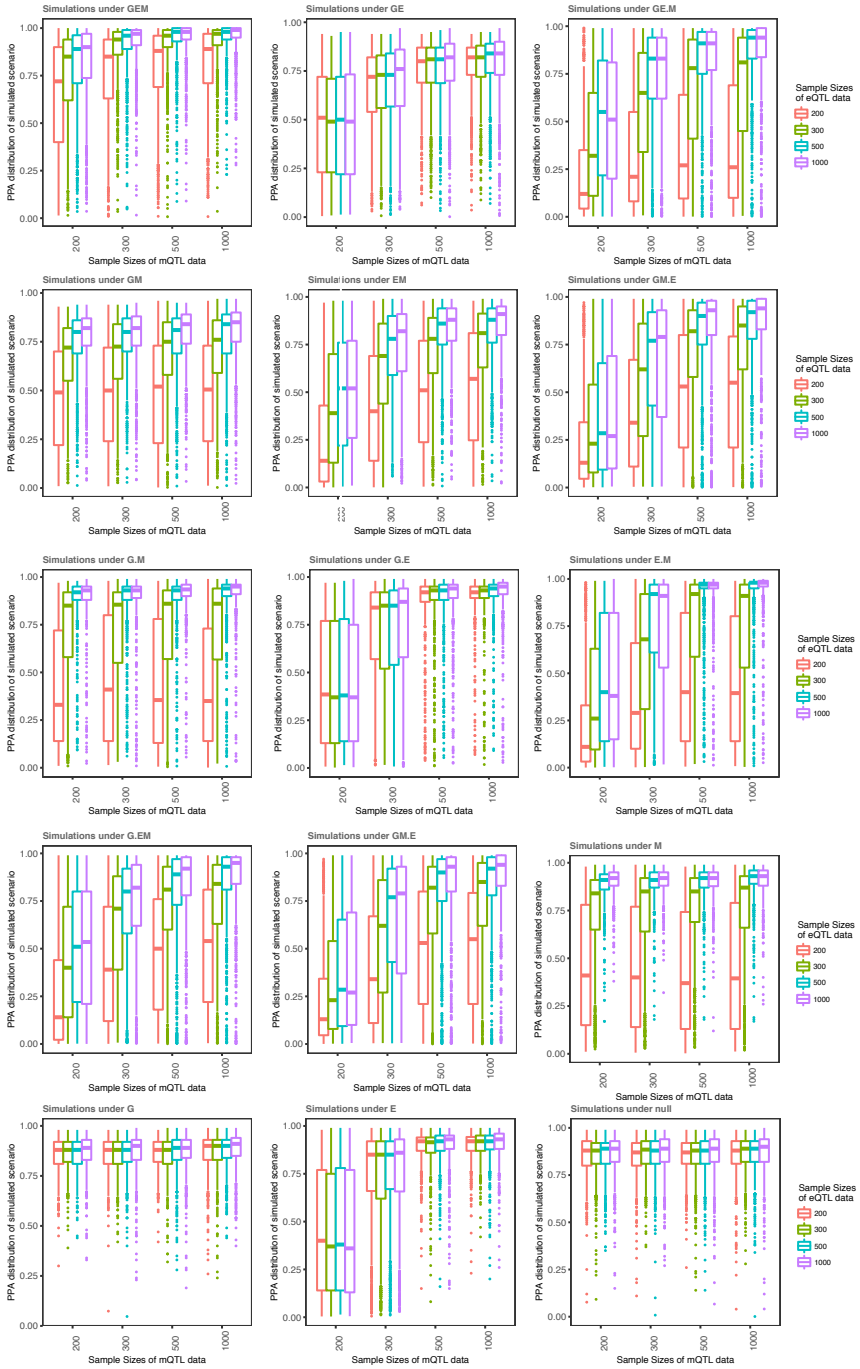


Figure S1. Simulations under different sample sizes for all scenarios in *moloc* of three traits (GWAS, eQTL, mQTL). All cases include 10,000 individuals in the

GWAS dataset. The variance explained by the trait was set to 0.01 for GWAS (1%), and to 0.1 (10%) for the eQTL and mQTL.

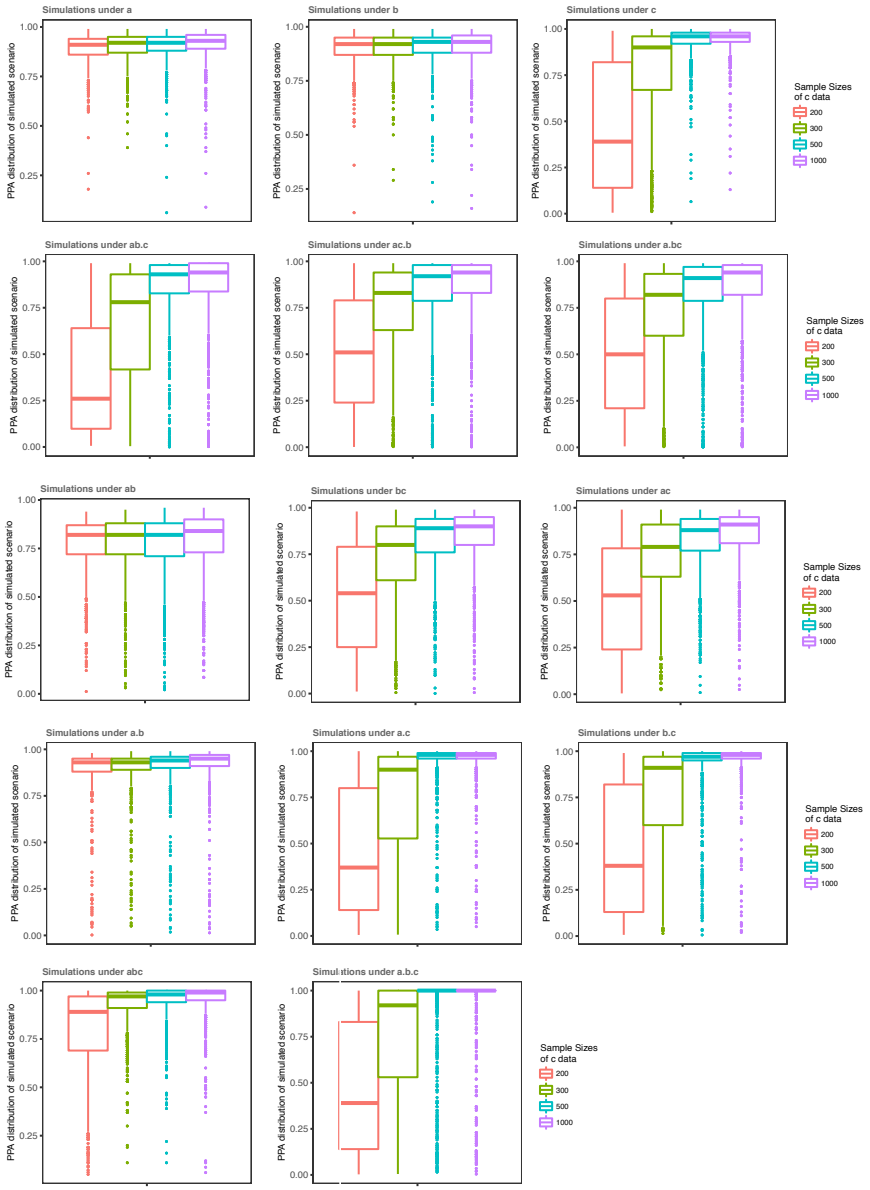


Figure S2. Simulations under different sample sizes for all scenarios in *moloc* of three traits ("a" is GWAS 1, "b" is GWAS 2, "c" is eQTL). All cases include 10,000 individuals in the biomarker dataset. The variance explained by the trait was set to 0.01 for GWAS 1 and GWAS 2, and to 1% (10%) for the eQTL.

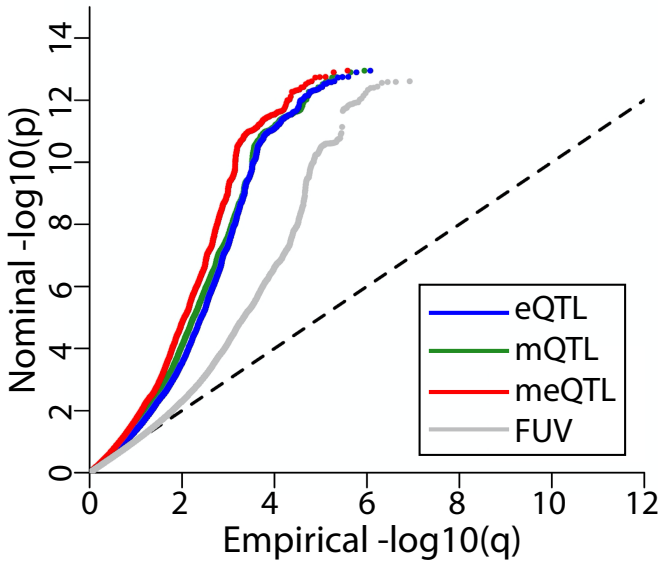


Figure S3. Stratified Q-Q plots for eQTL and mQTLs in the SCZ GWAS. Enrichment of each functional category (blue line: eQTL; green: mQTL; red: combined eQTL and mQTL) illustrate the estimated increase in terms of the proportion of p-values expected in each scenario and in the functionally unannotated variant (FUV) category (grey line). The estimated enrichment for eQTL and mQTL is similar.

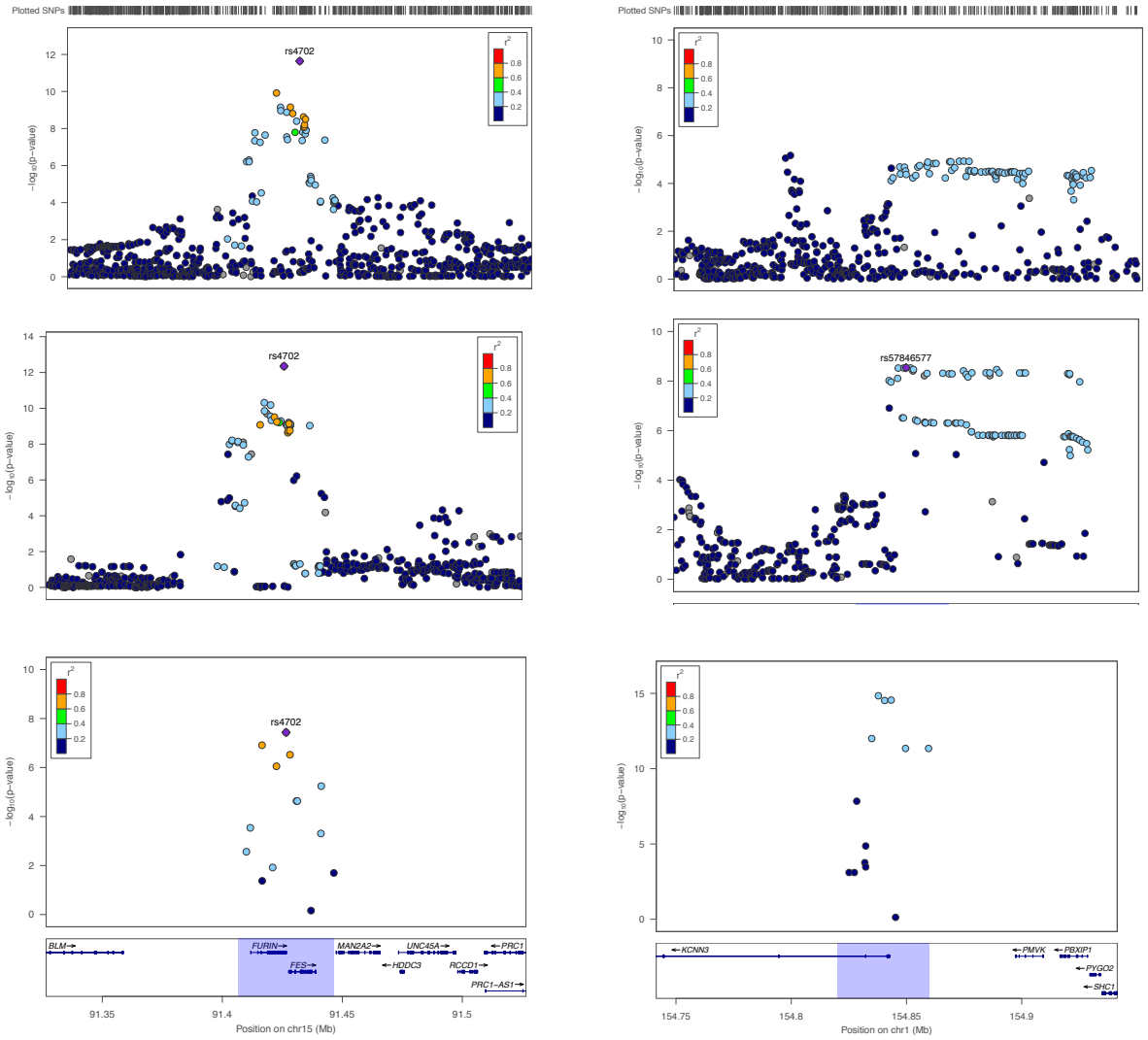


Figure S4. Illustration of two example of colocalization results with GWAS-eQTL-mQTL in the *FURIN* locus and cg24888049 and *KCNN3* and cg09359103. −log10(p) association p-values for GWAS (top panels), −log10(p) association p-values for eQTL (center panels), -log10(p) association p-values for mQTL (bottom panels) with location of genes at the FURIN locus, 200kb range. Highlighted is the region with the methylation probe (15:91426667 +/- 20kb for cg24888049 in the FURIN locus, and 1:154839909 +/- 20kb for cg09359103 in the *KCNN3 locus*).

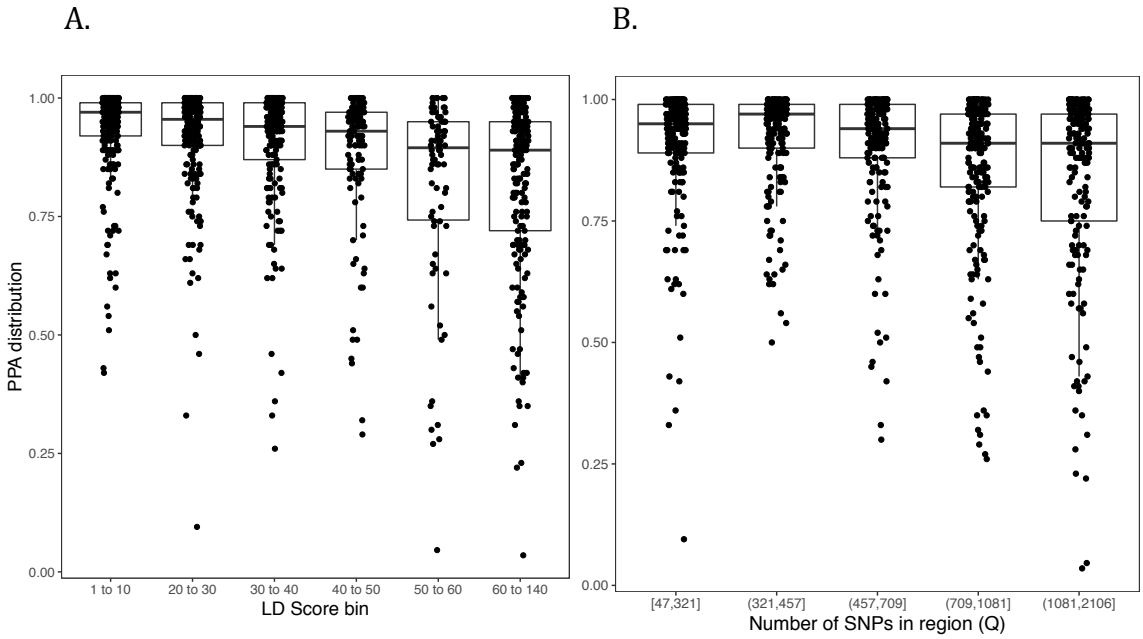


Figure S5. Simulation analysis with a shared causal variant between three studies using a sample size of 10,000 individuals for GWAS trait, 300 for eQTL trait, and 300 for mQTL trait. X-axis is the posterior probability of a shared causal causal variant for all three datasets (PPA.abc). **A.** Y-axis id the mean LD in the region. **B**. Y-axis is the number of SNPs Q in the region. Number of simulations 1,000.

Supplemental Tables

p1	p2	p3	Mean PPA Expectation	Mean Prior Expectation
1e-04	**1e-05**	**1e-06**	**0.78**	**0.93**
1e-04	**1e-05**	**1e-07**	**0.77**	**0.92**
1e-04	**1e-06**	**1e-07**	**0.64**	**0.60**
1e-04	**1e-06**	**1e-08**	**0.65**	**0.58**
1e-04	**1e-07**	**1e-08**	**0.48**	**0.13**
1e-04	**1e-05**	**1e-08**	**0.77**	**0.92**

Table S1. Effect of prior probability of co-localization for hypotheses of interest (GEM, GE, GE.M) on the prior and posterior support for colocalisation in regions

where there is evidence for association across the three dataests (PPA.GE + PPA.GE.M + PPA.GEM + PPA.G.E + PPA.G.EM + PPA.G.E.M). We repeated the moloc analysis with different combinations of priors p2 and p3 while keeping p1 fixed to $1x10^{-4}$. The posteriors with support for colocalization (PPA.coloc = PA.GE + PPA.GE.M + PPA.GEM) are contrasted to the ones with support against colocalization (PPA.nocoloc = PPA.G.M +PPA.G.EM+PPA.G.E.M) for our hypothesis of interest. For each analysis in which PPA.GE + PPA.GE.M + PPA.GEM + PPA.G.E + PPA.G.EM + PPA.G.E.M >= 0.8, we report the mean relative posterior support for colocalization, defined by PPA.coloc / (PPA.nocoloc + PPA.coloc), and the mean relative prior support, averaged across all regions. We find that the combination of priors p1=1e-04, p2=1e-06, p3=1e-07 has the lowest difference between the prior and posterior expectation.

Table S2. 39 genes falling within SCZ LD blocks with a high probability of influencing SCZ through expression or methylation (GEM, GE, GE.M>=0.8). "CHR", "START", "STOP" indicate the positions of the regions considered; "PGC2_LDclump_info" indicates position of the previously identified SCZ LD block; "PPA.combined" shows maximum posterior probability reached by a gene for evidence of association with GWAS and eQTL (PPA.GE + PPA.GE.M + PPA.GE.M); "numbMethyl" indicates the number of methylation probes overlapping this region.

Table S3. Potentially novel SCZ loci. 45 candidate genes outside of the SCZ LD blocks with a high probability of influencing SCZ (GEM, GE, GE.M>=0.8).

Table S4. False positive rates using 1000 and 500,000 simulations and our true sample sizes at a cut off of 80% (N simulations = 1000 and 500,000). We compute the false positives by counting the proportion of simulations under the null that passed a posterior probability supporting the "Detected Hypothesis" at a particular threshold (PPA>=threshold). PPA.comb is PPA.abc + PPA.ab + PPA.ab.c.

Table S5. False positive rates under the null of no association with any traits and different combination of sample sizes (N simulations = 1000). We compute the false positives by counting the proportion of simulations under the null that passed a posterior probability supporting the "Detected Hypothesis" at a particular threshold (PPA>=threshold). PPA.comb is PPA.abc + PPA.ab + PPA.ab.c.

Table S6. False positive rates under the null of no association with any traits and when the causal variant is masked (N simulations = 1000). We compute the false positives by counting the proportion of simulations under the null that passed a posterior probability supporting the "Detected Hypothesis" at a particular threshold (PPA>=threshold). PPA.comb is PPA.abc + PPA.ab + PPA.ab.c.

Table S7. Misclassification rates and power to detect the correct hypothesis (N simulations = 1000). We simulated each hypothesis of interest indicated under

the column "TRUE", and compute the number of misclassified calls under other 14 scenarios by counting the proportion of simulations that passed a posterior probability supporting a different hypothesis from the one simulated at a particular threshold (PPA>=cut). We estimated power to distinguish a particular hypothesis from the others by counting the proportion of simulations that did not reach a posterior probability supporting each scenario at a particular threshold (PPA(true)<cut). PPA.comb is PPA.abc + PPA.ab + PPA.ab.c.

Table S8. Misclassification rates and power to detect the correct hypothesis for combination of sample sizes (N simulations = 1000). We simulated each hypothesis of interest indicated under the column "TRUE", and compute the number of misclassified calls under other 14 scenarios by counting the proportion of simulations that passed a posterior probability supporting a different hypothesis from the one simulated at a particular threshold (PPA>=cut). We estimated power to distinguish a particular hypothesis from the others by counting the proportion of simulations that did not reach a posterior probability supporting each scenario at a particular threshold (PPA(true)<cut). PPA.comb is PPA.abc + PPA.ab + PPA.ab.c.

Table S9. Misclassification rates and power to detect the correct hypothesis for combination of sample sizes when the causal SNP is masked (N simulations = 1000). We simulated each hypothesis of interest indicated under the column "TRUE", and compute the number of misclassified calls under other 14 scenarios

by counting the proportion of simulations that passed a posterior probability supporting a different hypothesis from the one simulated at a particular threshold (PPA>=cut). We estimated power to distinguish a particular hypothesis from the others by counting the proportion of simulations that did not reach a posterior probability supporting each scenario at a particular threshold (PPA(true)<cut). PPA.comb is PPA.abc + PPA.ab + PPA.ab.c.