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 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

Correspondence:
Dr. Panos Roussos
Icahn School of Medicine at Mount Sinai, Institute for Multiscale Biology
Department of Psychiatry and Department of Genetics and Genomic Science and
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.

AUTHOR SUMMARY

In this paper, we propose **m**ultiple-trait-c**oloc** (***moloc***), a statistical method that can use single variant summary statistics from multiple studies to test for colocalization. Using ***moloc***, we integrate risk variants identified through genome-wide association studies (GWAS) in schizophrenia with quantitative trait loci that affect gene expression (eQTL) and DNA methylation (mQTL). Our method links non-coding risk variants with transcripts to identify possible biological mechanisms that affect the *cis* regulation of

67 gene expression. Overall, we find that integrating mQTL, eQTL and GWAS increases
68 discovery power as compared to standard pairwise colocalization. Our approach is
69 publically available as an R package from https://github.com/clagiamba/moloc.
70

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

94 largest GWAS by the Psychiatric Genomics Consortium[10], which reported association
95 for 108 independent genomic loci. eQTL data were derived from the CommonMind
96 Consortium[2], which generated the largest eQTL dataset in the dorsolateral prefrontal
97 cortex (DLPFC) from SCZ cases and control subjects (N=467). Finally, we leveraged
98 mQTL data that were previously generated in human DLPFC tissue (N=121) to
99 investigate epigenetic variation in SCZ[6]. Integration of multiple phenotypes helps better
100 characterize the genes predisposing to complex diseases such as SCZ.

RESULTS

Overview of the Method

We introduce an approach (***moloc***) to detect colocalization among any number of traits in a region in the genome. The input of the model is the set of summary statistics derived from three (or more) traits measured in distinct datasets of unrelated individuals. As a simplifying example, consider colocalization of three traits (e.g., GWAS, eQTL and mQTL); under a maximum of a single causal variant per trait, there can be up to three causal variants and 15 possible scenarios summarizing how the variants are shared among the traits (see Methods). For simplicity, we denote GWAS as G, eQTL as E, and mQTL as M. To illustrate our notation, consider a region with 8 SNP. Four examples of configurations are shown 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). Our approach computes the posterior probability of any of the 15 configurations by weighting the likelihood of the data *D* given a configuration *S*, $P(D|S)$, by the prior probability of a configuration, $P(S)$ (see Methods). Although our approach estimates a posterior for each of the 15 configurations, in general, we are interested in the scenarios supporting a shared causal variant for two and three traits.

Sample Size Requirements

In a first set of simulations we explored false positive rates (**Tables S1, S2, S3**) and 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. False positive rates are below 0.05 even if a threshold of 0.3 for posteriors is used, and where the causal variant is masked (**Tables S1**, **S2**, **S3**). We then explored misclassification rates (**Tables S4, S5, S6**). When samples are 10,000 for GWAS and greater than 300 for eQTL and mQTL, misclassification rates for detecting our hypotheses of interests at 80% threshold are all below 0.05 (**Table S4**). Where the causal variant is masked, sample sizes also need to reach 10,000 for GWAS and greater than 300 for eQTL and mQTL, for misclassification rates to be below 0.05 (**Table S5**). Given the small sample size for the mQTL data, the method has trouble detecting a different causal variant for the mQTL dataset (**Table S6**). For example, evidence pointing to two different causal variants between GWAS and eQTL could be generated by the presence of three causal variants in reality, but the causal variant for

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

147 It is instructive to observe where evidence for other hypotheses is distributed. **Figure 2**
148 illustrates the accuracy of our approach under different scenarios where two or three
149 causal variants are shared. For example, under simulations of one shared variant for
150 GWAS and eQTL and a second variant for mQTL (GE.M), on average 60% of the
151 evidence points to the simulated scenario, while 12% point to GE, 12% to G.E.M and
152 7.2% to GEM.
153

154 **Choice of priors**

155 The method requires the definition of prior probabilities at the SNP level for the
156 association with one (p1), two (p2), or three traits (p3). We set the prior probability that a
157 variant is associated with one trait as 1×10^{-4} for GWAS, eQTL and mQTL, assuming
158 that each genetic variant is equally likely a priori to affect gene expression or
159 methylation or disease. This estimate has been suggested in the literature for GWAS[12]
160 and used in similar methods[13]. We set the priors $p2 = 1 \times 10^{-6}$, $p3 = 1 \times 10^{-7}$ based on
161 sensitivity and exploratory analysis of genome-wide enrichment of GWAS risk variants
162 in eQTLs and mQTLs. In **Figure S3**, we find eQTLs and mQTLs to be similarly enriched
163 in GWAS, justifying our choice of the same prior probability of association across the
164 two traits. These values are also suggested by a crude approximation of *p2* and *p3* from
165 the common genome-wide significant SNPs across the three datasets.

166 We performed sensitivity analyses using different priors. Specifically, we fixed *p1* to
167 $1x10^{-4}$ and tested a range of priors for *p2* and *p3* from $1x10^{-5}$ to $1x10^{-8}$, with increasing
168 difference between *p1*, *p2* and *p3*. We used a form of internal empirical calibration to
169 compare our prior and posterior expectations. We considered a range of values for *p2*
170 and *p3* while keeping *p1* fixed, and find that the posterior expectation of colocalization
171 most closely resembled the prior expectation under our choice of priors (**Table S7**). We
172 note that our R package implementation allows users to specify a different set of priors.
173

174 **Co-localization of eQTL, mQTL and risk for Schizophrenia**

175 We applied our method to SCZ GWAS using eQTLs derived from 467 samples
176 and mQTL from 121 individuals. Our aim is to identify the genes important for disease
177 through colocalization of GWAS variants with changes in gene expression and DNA
178 methylation. We analyzed associations genome-wide, and report results both across
179 previously identified GWAS loci, and across potentially novel loci. While we consider all
180 15 possible scenarios of colocalization, here we focus on gene discovery due to higher
181 power in our eQTL dataset, by considering the combined probabilities of cases where
182 the same variant is shared across all three traits GWAS, eQTLs and mQTLs (GEM >
183 0.8) or scenarios where SCZ risk loci are shared with eQTL only (GE > 0.8 or GE.M >
184 0.8) (**Table 1**). We identified 1,053 cis-regions/methylation pairs with posterior
185 probability above 0.8 that are associated with all three traits (GEM), or eQTLs alone
186 (GE or GE.M). These biologically relevant scenarios affect overall 84 unique genes and
187 include 39 genes that fall within the previously identified SCZ LD blocks (**Table S8**) and

188 45 potentially novel genes outside of these regions (**Table S9**). Fifty-two out of the 84
189 candidate genes influence SCZ, gene expression and methylation (GEM>=0.8). One
190 possible scenario is that the variants in these genes could be influencing the risk of SCZ
191 through methylation, although other potential interpretations such as pleiotropy should
192 be considered.
193
194 **Addition of a third trait increases gene discovery**
195 We examined whether moloc with 3 traits enhance power for GWAS and eQTL
196 colocalization compared to using 2 traits. In simulations to compare coloc and moloc
197 under one causal variant and our true sample sizes for all three datasets, we observe a
198 fold increase of 1.5 for gene discovery using moloc versus coloc. Moloc with three traits
199 recovers all the genes discovered using *coloc* with eQTL and mQTL, and additional
200 genes from the inclusion of the third layer. In our real data, colocalization analysis of
201 only GWAS and eQTL traits identified 45 genes with a posterior probability, PP4 in
202 *coloc*, of >= 0.8. The 39 additional genes that were found by adding methylation include
203 genes such as *CALN1,* a neuronal transcript associated with abnormalities in
204 sensorimotor gating in humans[14], that would have been missed by only GWAS and
205 eQTL colocalization.
206
207 **Loci overlapping reported SCZ LD blocks**
208 Psychiatric Genomics Consortium (PGC) identified 108 independent loci and
209 annotated LD blocks around these, 104 of which are within non-HLA, autosomal regions
210 of the genome[10]. In **Table 1** we report the number of identified gene-methylation pairs

211 and unique genes under each scenario that overlap the SCZ-associated LD blocks. Out
212 of the 78 SCZ-associated LD blocks we examined in our analysis, we found
213 colocalizations in 20 of them with an average gene density per block of 2.4 unique
214 genes. **Figure 3A** illustrates the average distribution of the posteriors across these
215 regions. Cumulatively, 12% of the evidence points to shared variation with an eQTL
216 (GE, GE.M and GEM). The majority of the evidence within these regions (64%) did not
217 reach support for shared variation across the three traits, with 20% not reaching
218 evidence for association with any traits, and 44% with only one of the three traits (36%
219 with GWAS; 6% with eQTL, 2% with mQTLs). The lack of evidence in these regions
220 could be addressed with greater sample sizes. **Figure 3B** shows the evidence for
221 colocalization of GWAS with eQTL or mQTL across the 39 candidate genes. We
222 provide illustrative examples of SCZ association with expression and DNA methylation
223 in the *FURIN* locus (**Figure 4** and **Figure S4A**).
224
225 **Potentially novel SCZ loci**
226 We found 45 unique genes that have a high posterior for SCZ and eQTL, but fall
227 in regions not previously identified to be associated with SCZ (at p-value of 5×10^{-8}).
228 All genes were far from a SCZ LD block (more than 150kb, **Table S9**), and contained
229 SNPs with p-values for association with SCZ ranging from 10^{-4} to 10^{-8}. These genes will
230 likely be identified using just the GWAS signal if the sample size is increased. *KCNN3* is
231 among these genes which encodes an integral membrane protein that forms a voltage-
232 independent calcium-activated channel. It regulates neuronal excitability by contributing

233 to the slow component of synaptic afterhyperpolarization[15]. A plot of the associations
234 with the three datasets within this locus is shown in **Figure S4B.**
235
236 **Comparison with previous findings**
237 We compare our gene discovery results to previous studies that assess GWAS-
238 eQTL[2–5] or GWAS-mQTL[5,7–9] colocalization using the same or similar datasets (**Table**
239 **2**). A substantial proportion of genes detected in our study (range 44%-85%) was
240 validated with four studies [2–5] that used eQTL and GWAS integration to prioritize genes
241 important for schizophrenia. Several studies have also linked methylation data with
242 schizophrenia [7–9]. Two recent studies[7,9] used blood mQTL data from 639 samples and
243 identified colocalization of SCZ loci with 32 and 200 methylation probes by applying
244 COLOC and SMR, respectively. A proportion of SCZ-mQTL colocalization was validated
245 in our study (COLOC: 46%; SMR: 18%, **Table 2**). Overlap between our analysis in brain
246 and these analyses point to shared mechanisms in blood and brain. Another study[8]
247 used mQTL data from 166 fetal brain samples and identified 297 methylation probes
248 important for schizophrenia. We analyzed 184 of those and found evidence for 13
249 probes. We note that our methylation data did not included fetal brain samples. Finally,
250 a recent study[5] identified 44 genes involved in schizophrenia through TWAS, followed
251 by integration with chromatin data in blood that resulted in 11 genes associated with
252 GWAS, eQTL and epigenome QTL. We analyzed 8 out of the 11 associations and
253 confirmed 6 of these genes that, in our study, influence SCZ through eQTL and mQTL.
254
255 **Association of gene expression with methylation**

256 We examined the association of DNA methylation and gene expression as a
257 function of distance from transcription start site. We explored direction of effects of
258 methylation and expression, for gene expression and DNA methylation that colocalize
259 (PPA.GEM + PPA.EM + PPA.G.EM >= 0.8). This approach has the advantage of linking
260 changes in methylation with specific transcripts, avoiding the issues of arbitrary
261 annotating CpG methylation sites to the nearby genes. Overall, we tested 1,947 DNA
262 methylation and gene expression pairwise interactions and found a significant negative
263 correlation between the effect sizes of methylation and expression in the proximity of
264 the transcription start site (**Figure 5**, p-value:<2.2e-16). **Table S10** provides a list of
265 methylation and gene expression pairwise interactions in human brain tissue for DNA
266 methylation probes that are proximal to the transcription start site (20kb upstream to 2kb
267 downstream of transcription start site).

268 **DISCUSSION**

269 In this paper, we propose a statistical method for integrating genetic data from
270 molecular quantitative trait loci (QTL) mapping into genome-wide genetic association
271 analysis of complex traits. The proposed approach requires only summary-level
272 statistics and provides evidence of colocalization of their association signals. To our
273 knowledge, a method integrating more than two traits is lacking. In contrast to other
274 methods that attempt to estimate the true genetic correlation between traits such as LD
275 score regression[16] and TWAS[17], *moloc* focuses on genes that are detectable from the
276 datasets at hand. Thus, if the studies are underpowered, most of the evidence will lie in
277 the null scenarios. We note that our model is the same as *gwas-pw* in Pickrell et al.[18]
278 under specific settings. Precisely, *gwas-pw* averages over Bayes factors computed with
279 W = 0.01, W = 0.1, and W = 0.5 (Methods). We provide this as an option that can be
280 specified by the user. Additionally, *gwas-pw* estimates the prior parameters genome-
281 wide using a maximization procedure. However, we note that, unlike *gwas-pw* that
282 focuses on genome-wide estimation across pairs of traits, our approach focuses on one
283 locus at a time with multiple traits.

284 Our method is complementary to methods quantifying local genetic covariance
285 across traits. The aim of this method is to identify cases where the same causal variant
286 is shared between the traits, and we do not capture cases where traits are driven by a
287 different causal SNP. We argue that it is valuable to identify cases where the same
288 signal is influencing multiple traits, for example when studying application to drug
289 development and possible side effects of drugs to non-target biomarkers. Other

290 methods such as genetic covariance[19] can be used to identify genes shared across
291 traits even where the causal variants differ.

292 We expose one possible application of this approach in SCZ. In this application,
293 we focus on scenarios involving eQTLs and GWAS, alone or in combination with
294 mQTLs. Other scenarios are also biologically important. For example, colocalization of
295 GWAS and mQTL excluding eQTLs (GM.E scenario) could unveil important methylation
296 mechanisms affecting disease but not directly influencing gene expression in *cis*. We
297 report these and other scenarios in our web resource and encourage further
298 examination of these cases in future analyses. The GEM scenario provides evidence
299 that SCZ risk association is mediated through changes in DNA methylation and gene
300 expression. While our method does not detect causal relationships among the
301 associated traits, i.e. whether risk allele leads to changes in gene expression through
302 methylation changes or vice versa, there is evidence supporting the notion that risk
303 alleles might affect transcription factor binding and epigenome regulation that drives
304 downstream alterations in gene expression[20,21].

305 Our method helps to link a specific gene to a methylation probe. DNA
306 methylation is one the best studied epigenetic modifications. Methylation can alter gene
307 expression by disrupting transcription factor binding sites (with variable consequences
308 to expression depending on the transcription factor), or by attracting methyl-binding
309 proteins that initiate chromatin compaction and gene silencing. Therefore methylation
310 can be associated with both increased or decreased gene expression[20,22]. Increased
311 CpG methylation in promoter regions is usually associated with silencing of gene
312 expression[23]. However, in genome-wide expression and methylation studies, the

313 correlation of methylation and gene expression is low or the pattern of association is
314 mixed, even for CpG methylation within promoter regions[22]. We found a negative
315 correlation between the effect sizes of methylation and expression in the proximity of
316 the transcription start site. However, we note that the current assay (Illumina 450K)
317 cannot discriminate DNA methylation from hydroxymethylation sites. DNA methylation is
318 generally associated with the inhibition of gene expression, while DNA
319 hydroxymethylation, the oxidized form of methylation, has been associated with
320 enhancers and increased gene expression.

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

328 We assign a prior probability that a SNP is associated with one trait (1×10^{-4}), to
329 two (1×10^{-6}), and to three traits (1×10^{-7}). We find support for our choice of priors in the
330 data using two methods. The first uses stratified QQ plots (**Figure S3**). We find that
331 eQTL enrichment in GWAS has a similar enrichment to mQTL in GWAS. The second is
332 a form of empirical calibration as in Guo et al.[24]. We find that the prior and posterior
333 expectations of colocalisation matched more closely under our choice of priors (**Figure**
334 **S6**). However the choices for prior beliefs for each hypothesis are always arguable. One
335 could estimate priors for the different combinations of datasets. Pickrell et al.[18] proposed

336 estimation of enrichment parameters from genome-wide results maximizing *a posteriori*
337 estimates for two traits. For multiple traits, another possibility is using deterministic
338 approximation of posteriors[25]. We leave these explorations to future research.

339 The results of this analysis apply to a particular brain region (dorsolateral
340 prefrontal cortex), age period (adulthood), and ancestry (Caucasians only). We justify
341 using functional data from DLPFC to examine the colocalization with schizophrenia
342 variants as it is the most proximal to disease etiopathogenesis tissue [4,5]. In addition, due
343 to the tissue specificity of gene expression regulatory mechanisms, we matched eQTL
344 and mQTL data on the same brain region and age period. The genotypes and gene
345 expression data reported by the CommonMind Consortium (CMC) comprises the largest
346 existing postmortem brain genomic resource from individuals with psychiatric disorders
347 as well as unaffected controls. The mQTL dataset used here is the largest dataset in the
348 adult brain measured in the brain region and age period considered. Although this was
349 out of the scope of this paper, we encourage application of similar methods to functional
350 data derived from different tissues and age periods to elucidate the spatiotemporal
351 regulatory mechanisms of gene expression that are involved in the etiopathogenesis of
352 schizophrenia.

353 The method makes use of LD to weigh out the posteriors for the different models,
354 therefore the statistics (priors and posteriors of configurations) will depend on the
355 pattern of association (LD) and the number of SNPs in the region (Q). Specifically,
356 increased LD in a region with one causal variant decreases our chance to detect a
357 shared causal effect (**Figure S5A**). This is because in loci where the extent of LD is
358 strong (a large number of variants with the same strength of association), there is

359 uncertainty on whether we are dealing with a single variant causal for both traits, or two
360 independent variants in the same LD block. Similarly, the chance of the same SNP
361 being causal for all three traits decreases as Q increases (**Figure S5B**).

362 The Bayes Factors under the models of colocalization (for the case of a single
363 variant affecting more than one trait) are computed by multiplying the individual WABFs
364 together, and this assumes independence of the effect sizes for the traits.
365 Independence is appropriate in the case of analyzing different phenotypes, or eQTLs in
366 different tissues. On the other hand, this assumption may be inappropriate in other
367 cases, for example if we were looking at colocalization of a variant for the same trait in
368 two different studies, or the same eQTLs across different tissue type, or the same trait
369 in two different studies. A more general BF computation which includes cases where the
370 effects are correlated has been proposed in [26] and could be used to extend our
371 approach to these situations. Additionally, correlation among the SNP effect sizes may
372 result if the studies contain overlapping individuals. In this case we may overestimate
373 the evidence in favor of colocalization. We could adjust SNP Bayes factors to account
374 for this overlap as in [27]. The datasets we used in the current analysis do not contain
375 overlapping individuals, however we provide the code to use this approach.

376 We note that this approach can be extended to more than three traits. Since the
377 calculations are analytical and no recursive method is used, computation time for a
378 region with 1000 SNPs is less than one second. However, time increases exponentially
379 as number of traits increases. For four traits it is about 3 seconds, for five traits it is
380 greater than 22 minutes. Overall, owing to the increasing availability of summary
381 statistics from multiple datasets, the systematic application of this approach can provide

382 clues into the molecular mechanisms underlying GWAS signals and how regulatory
383 variants influence complex diseases.

384

385 **MATERIALS AND METHODS**

386 **Ethics statement**

387 This paper re-analyzes previously published datasets. All samples and patient
388 data were handled in accordance with the policies and procedures of the participating
389 organizations.

390 **GWAS dataset**

391 Summary statistics for genome-wide SNP association with Schizophrenia were
392 obtained from the Psychiatric Genomics Consortium-Schizophrenia Workgroup (PGC-
393 SCZ) primary meta-analysis (35,476 cases and 46,839 controls)[10].

394

395 **Expression QTL (eQTL) analysis**

396 For the eQTL associations, we used pre-QC'ed expression and imputed
397 genotypes, on case-control of European-ancestry (N=467) collected by the
398 CommonMind Consortium. Briefly, gene expression was assayed data from post-
399 mortem DLPFC (Brodmann areas 9 and 46) with RNA-seq and mapped to human
400 Ensembl gene reference (v70) using TopHat version 2.0.9 and Bowtie version 2.1.0.
401 Genes with less than 1 CPM (counts per million) in at least 50% of the samples were
402 discarded and RNA-seq data for a total of 16,423 Ensembl genes were considered for
403 downstream analyses. After RNA sample QC, and retaining only genetically-identified
404 European-ancestry individuals, a total of 467 samples were used for downstream
405 analyses. These comprised 209 SCZ cases, 52 AFF (Bipolar, Major depressive
406 disorder, or Mood disorder, unspecified) cases, and 206 controls. The expression data

407 was voom-adjusted[28] for both known covariates (RIN, library batch effect, institution,
408 diagnosis, post-mortem interval, age and sex) and hidden confounds identified based
409 on surrogate variable analysis.

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

416 The R package MatrixEQTL [29] was used to fit an additive linear model between
417 the expression of 16,423 genes and SNP dosage data at 6.4 million markers within a 1
418 Mb window around the transcription start site for each gene, including covariates for
419 diagnosis, five ancestry covariates and known and hidden variables detected by
420 surrogate variable analysis.

421 Following eQTL identification, summary statistics on 72,441,005 SNPs and 15,791
422 autosomal genes were retained for downstream analyses. Overall, the model identified
423 2,154,331 significant cis-eQTL, (i.e., SNP–gene pairs within 1 Mb of a gene) at a false
424 discovery rate (FDR) ≤ 5%, for 13,137 (80%) genes.

425

426 **Methylation QTL (mQTL) analysis**

427 For the mQTL dataset, we used pre-QC'd DNA methylation data and imputed
428 genotypes on non-psychiatric adult Caucasian control donors (age > 13, N=121). We

429 briefly summarize here the procedures to process the raw data. More details on data
430 processing, normalization, and QC can be found in [6].
431 DNA methylation was assessed on postmortem tissue homogenates of the dorsolateral
432 prefrontal cortex (DLPFC, Brodmann areas 9 and 46) on 534 unique subjects using the
433 Illumina HumanMethylation450 ("450k") microarray (which measures CpG methylation
434 across 473,058 probes covering 99% of RefSeq gene promoters). The average intensity
435 was used to check for low quality samples, intensities from the sex chromosomes were
436 used to drop samples that had predicted sex which was different from its recorded
437 value, and across-sample quantile normalization was used. DNA for genotyping was
438 obtained from the cerebella of samples with either the Illumina Human Hap 650v3, 1M
439 Duo V3, or Omni 5M BeadArrays. Genotypes were called, QC'd, phased and imputed to
440 the 1000 Genomes Phase 3 variant set, separately by genotyping platform. Imputed
441 genotypes were then merged across the three platforms, and SNPs with MAF > 5% and
442 imputation quality above 90% were retained across the samples. Multidimensional
443 scaling (MDS) components were obtained for quantitative measures of ancestry.
444 For the mQTL association analysis, we only considered 121 adult Caucasian
445 individuals (age > 13). The analysis was conducted using the R package MatrixEQTL [29],
446 fitting an additive linear model up to 20kb distance between each SNP and CpG
447 analyzed, including as covariates the first five MDS components from the genetic data
448 and the first 8 PCs (based on the 100,000 most inter-individual variable probes for
449 computational efficiency). Summary statistics on 47,675,913 SNPs and 473,058
450 Illumina 450k probes were used for downstream analyses. Overall, the model identified

451 3,601,681 significant cis-mQTL, (i.e., SNP–probe pairs within 20kb distance of a probe)
452 at a false discovery rate (FDR) ≤ 5%, for 127,237 (27%) probes.
453
454 **Method Description**
455 Our approach is a multi-trait extension of the two-trait model described in
456 ***coloc*[11].** We define a genomic region containing **Q** variants, for example a *cis* region
457 around expression or methylation probe. We are interested in a situation where
458 summary statistics (effect size estimates and standard errors) are available for all
459 datasets in the genomic region. We first derive our model using three traits, then
460 generalize to any number of traits. If we consider three traits, there are 15 possible
461 scenarios summarizing how the variants are shared among the traits. Each hypothesis
462 can be represented by a set of index sets according to which of the traits each SNP is
463 associated with: H_0- no association to any trait, H_1- association only for trait {1}, H_2-
464 association only for trait {2}, H_3- association only for trait {3}, H_4- association for traits {1
465 and 2}, H_5- association for traits {2 and 3}, H_6- association for traits {1 and 3}, H_7-
466 association for trait 1 and 2, but different causal variants for {1} and {2}, H_8- association
467 for traits 2 and 3, but different causal variants for {2} and {3}, H_9- association for traits 1
468 and 3, but different causal variants for {1} and {3}, H_{10}- association for traits 1, 2 and 3,
469 but different causal variants for {1} and {2,3}, H_{11}- association for traits 1, 2 and 3, but
470 different causal variants for {3} and {1,2}, H_{12}- association for traits 1, 2 and 3, but
471 different causal variants for {2} and {1,3}, H_{13}- association for traits 1, 2 and 3, but
472 different causal variants for {1}, for {2} and for {3}, and H_{14} SNP is associated with all 3
473 traits {1,2,3}.

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 can reformulate the posterior probability for each hypothesis as a ratio by dividing each by the baseline likelihood supporting the first model of no association with any trait H_0. The probability of the data for hypothesis h is then the sum over all configurations Sh, which are consistent with the given hypothesis:

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

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, computed from:

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

485 Therefore, in our application, the method outputs 15 posterior probabilities. We are most
486 interested in the scenarios supporting a shared causal variant for two and three traits.
487 We make three important assumptions in *moloc*, the same that are made in our
488 previous *coloc* methodology. Firstly, that the causal variant is included in the set of *Q*
489 common variants, either directly typed or well imputed. If the causal SNP is not present,
490 the power to detect a common variant will be reduced depending on the linkage
491 disequilibrium (LD) between other SNPs included in the model and the causal SNP.
492 Secondly, we assume at most one causal variant is present for each trait. In the
493 presence of multiple causal variants per trait, this method is not able to identify
494 colocalization between additional association signals independent from the primary one.
495 Thirdly, as we do not explicitly model LD between SNPs, we assume the samples are
496 drawn from the same ethnic population and therefore have identical allele frequencies
497 and patterns of LD.
498
499 **Bayes Factor of a SNP with one trait**

We start by computing a Bayes Factor for each SNP and each 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[30] (WABF), only the variance and effect estimates from regression analysis are needed, as shown below and previously described[11,18]:

$$WABF_i^j = \frac{1}{\sqrt{1-r}} \times exp\left[-\frac{Z_i^{2(j)}}{2} \times r\right] \quad (3)$$

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 [27]. 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]:

$$V_i = Var(\hat{\beta}_i) = \frac{1}{Ns_i(1-s_i) \times 2f_i(1-f_i)} \qquad (4)$$

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:

$$BF_{i,\mathbf{s}} = \prod_{j \in s} BF_i^j \qquad (5)$$

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[26]. Two key assumptions are necessary for the following computations. Firstly, that the traits are measured in unrelated individuals, and secondly

513 that the effect sizes for the two traits are independent. We will consider when these
514 assumptions are appropriate in Discussion.

515 **Prior probabilities that SNP i associates with traits indexed in s**

516 The prior probability that SNP i associates with all traits indexed in a set in our
517 three trait model is: π_{φ}SNP i associates with no trait, with one trait, pairs or traits or all
518 traits $\pi_{\{1,2,3\}}$ such that they sum to 1: $\pi_{\varphi} + \pi_{\{1\}} + \pi_{\{2\}} + \pi_{\{3\}} + \pi_{\{1,2\}} + \pi_{\{1,3\}} + \pi_{\{2,3\}} +$
519 $\pi_{\{1,2,3\}} = 1$.

520 **Simplified model under the assumption that all SNPs have the same prior**

521 The probability of the data under each hypothesis can be computed by summing
522 the probability of all the causal configurations consistent with a particular hypothesis,
523 weighted by the prior probabilities (Equation 1).
524 In our model, *P(S)* is the prior probability under any one of the 15 hypotheses. We can
525 define these priors from the prior probability that a SNP i associates with traits indexed
526 in π_s (section above). Additionally, since the prior probability *P(S)* of any one
527 configuration in the different sets do not vary across SNPs that belongs to the same set
528 S_h, we can multiply the likelihoods by one common prior supporting the different
529 hypothesis[11]. In this framework, we can see that *P(S)* depends on a ratio of π_s and on
530 Q, the number of SNPs in the region (Text S1).
531 Therefore, across a set of j traits {1,2,3...,}, we compute the probability of the data
532 supporting hypothesis h, where one SNP is associated with j traits as:

$$\frac{P(H_h|D)}{P(H_0|D)} = \prod_{s \in h} \pi_s \sum_{i=1}^{Q} BF_{i,s} \qquad (6)$$

where $BF_{i,s}$ are the Bayes factor of a SNP across traits indexed in s (Equation 5), π are the prior probabilities that SNP i is the causal SNP under a specific model.

The probability of the data where there are more than one independent associations among the j traits (i.e. |h| > 1) can be derived from the pre-computed probability of the data where there is one association among the j traits (i.e. |h| = 1, Text S1):

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

where t is the union of the index set in h.

Prior Probabilities of each hypothesis

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.

Moloc Analysis

The GWAS, eQTL, mQTL datasets were filtered by minor allele frequency greater than 5% and had individually been filtered by imputation quality (Methods).. 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-

551 centric approach, defined *cis*-regions based on a 50kb upstream/downstream from the
552 start/end of each gene, since our goal is to link risk variants with changes in gene
553 expression. We evaluated all methylation probes overlapping the *cis*-region. The
554 number of cis-regions/methylation pairs is higher than the count of genes because, on
555 average, there are more than one methylation sites per gene. Common SNPs were
556 evaluated in the colocalization analysis for each gene, and each methylation probe, and
557 GWAS. In total, 12,003 *cis*-regions and 481,995 unique *cis*-regions/methylation probes
558 were tested. Genomic regions were analyzed only if greater than 50 SNPs were in
559 common between all the datasets. Across all of the analyses, a posterior probability
560 equal to, or greater than, 80% for each configuration was considered evidence of
561 colocalization.

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

567 **Simulations**

568 We simulated genotypes from sampling with replacement among haplotypes of
569 SNPs with a minor allele frequency of at least 5% found in the phased 1000 Genomes
570 Project within 49 genomic regions that have been associated with type 1 diabetes (T1D)
571 susceptibility loci (excluding the major histocompatibility complex (MHC) as previously
572 described[31]. These represent a range of region sizes and genomic topography that

573 reflect typical GWAS hits in a complex trait. For each trait, two, or three "causal
574 variants" were selected at random. We have simulated continuous traits, and assume
575 that causal effects follow a multivariate Gaussian distribution, with each causal variant
576 explaining 0.01 variance of the trait in the GWAS data, and 0.1 in the eQTL and mQTL
577 datasets. Note that colocalisation testing may be applied equally to quantitative data
578 (using linear regression), and to case control data (using logistic regression). In our
579 application, one of the datasets is a case-control and we are directly using estimates,
580 and associated standard errors, from fitting a log-odds generalised linear model. This is
581 equivalent to applying *moloc* on the latent liability scores, which is a quantitative
582 phenotype. For the null scenario, the causal variants explain zero variance of the traits.
583 Each scenario was simulated 1,000 times. To quantify false positive rates on a large
584 number of tests, we simulated the null 500,000 times. We simulated the 15 possible
585 scenarios with different sharing patterns between the GWAS, eQTL, and mQTL
586 datasets. We used sample sizes of 82,315, 467, and 121 individuals to reflect our true
587 sample sizes. We also used different combinations of sample sizes to explore power to
588 detect the correct hypothesis.

589 We estimated the number of false positives within each simulated scenario, by
590 counting the proportion of simulations under the null that passed a posterior probability
591 supporting each of the 14 hypothesis at a particular threshold (PPA>=threshold). We
592 also report the false positives using the sum of the posteriors (PPA.ab + PPA.ab.c +
593 PPA.abc). The false positive rate is the number of false positives over 1,000
594 simulations. We repeated this procedure using 500,000 simulations under our true
595 sample sizes.

596 We next sought to compare the misclassification rates, and power to detect the
597 correct hypothesis. To compute the number of misclassified calls within each simulated
598 scenario, we counted the proportion of simulations that passed a posterior probability
599 supporting a different hypothesis from the one simulated at a particular threshold
600 (PPA>=threshold). We estimated power to distinguish a particular hypothesis from the
601 others by counting the proportion of correctly identified simulations at a particular
602 threshold (PPA(true)>=threshold). Since in most cases the causal variant will not be
603 included in the panel, we repeated simulations after removing the causal variant.
604 To explore the effect of linkage disequilibrium (LD) on estimated posterior
605 probability, we first computed an LD score for each SNP in the region, defined as the
606 sum of the squared correlation between a SNP and all the SNPs in the region. To
607 assess the degree of LD at a locus we took the average of these scores.
608 All analyses were conducted in R.

609 **Web Resources**

610 We developed a web site to visualize the colocalization results of SCZ GWAS,
611 eQTL, mQTLs under all possible scenarios (**icahn.mssm.edu/moloc**). The browser
612 allows searches by gene, methylation probe, and scenario of interest. The *moloc*
613 method is available as an R package from https://github.com/clagiamba/moloc.
614

615 **ACKNOWLEDGMENTS**

616 We would like to thank Chris Wallace at the Department of Medicine and MRC
617 Biostatistics Unit, Cambridge Biomedical Campus, University of Cambridge, Cambridge,
618 UK.
619 This work was supported by the National Institutes of Health (R01AG050986 Roussos
620 and R01MH109677 Roussos), Brain Behavior Research Foundation (20540 Roussos),
621 Alzheimer's Association (NIRG-340998 Roussos) and the Veterans Affairs (Merit grant
622 BX002395 Roussos). Additionally, this work was supported in part through the
623 computational resources and staff expertise provided by Scientific Computing at the
624 Icahn School of Medicine at Mount Sinai.
625 Data were generated as part of the CommonMind Consortium supported by funding
626 from Takeda Pharmaceuticals Company Limited, F. Hoffman-La Roche Ltd and NIH
627 grants R01MH085542, R01MH093725, P50MH066392, P50MH080405,
628 R01MH097276, RO1-MH-075916, P50M096891, P50MH084053S1, R37MH057881
629 and R37MH057881S1, HHSN271201300031C, AG02219, AG05138 and MH06692.
630 Brain tissue for the study was obtained from the following brain bank collections: the
631 Mount Sinai NIH Brain and Tissue Repository, the University of Pennsylvania
632 Alzheimer’s Disease Core Center, the University of Pittsburgh NeuroBioBank and Brain
633 and Tissue Repositories and the NIMH Human Brain Collection Core. CMC Leadership:
634 Pamela Sklar, Joseph Buxbaum (Icahn School of Medicine at Mount Sinai), Bernie
635 Devlin, David Lewis (University of Pittsburgh), Raquel Gur, Chang-Gyu Hahn (University
636 of Pennsylvania), Keisuke Hirai, Hiroyoshi Toyoshiba (Takeda Pharmaceuticals
637 Company Limited), Enrico Domenici, Laurent Essioux (F. Hoffman-La Roche Ltd), Lara
638 Mangravite, Mette Peters (Sage Bionetworks), Thomas Lehner, Barbara Lipska (NIMH).

639 **REFERENCES**

640 1. Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five Years of
641 GWAS Discovery. Am. J. Hum. Genet. *90*, 7–24.
642 2. Fromer, M., Roussos, P., Sieberts, S.K., Johnson, J.S., Kavanagh, D.H., Perumal,
643 T.M., Ruderfer, D.M., Oh, E.C., Topol, A., Shah, H.R., et al. (2016). Gene expression
644 elucidates functional impact of polygenic risk for schizophrenia. Nat. Neurosci. *19*,
645 1442–1453.
646 3. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery,
647 G.W., Goddard, M.E., Wray, N.R., Visscher, P.M., et al. (2016). Integration of summary
648 data from GWAS and eQTL studies predicts complex trait gene targets. Nat. Genet. *48*,
649 481–487.
650 4. Hauberg, M.E., Zhang, W., Giambartolomei, C., Franzén, O., Morris, D.L., Vyse, T.J.,
651 Ruusalepp, A., Sklar, P., Schadt, E.E., Björkegren, J.L.M., et al. (2017). Large-Scale
652 Identification of Common Trait and Disease Variants Affecting Gene Expression. Am. J.
653 Hum. Genet. *100*, 885–894.
654 5. Gusev, A., Mancuso, N., Finucane, H.K., Reshef, Y., Song, L., Safi, A., Oh, E., O
655 ’donovan, M.C., Katsanis, N., Crawford, G.E., et al. TITLE: Transcriptome-wide
656 association study of schizophrenia and chromatin activity yields mechanistic disease
657 insights.
658 6. Jaffe, A.E., Gao, Y., Deep-Soboslay, A., Tao, R., Hyde, T.M., Weinberger, D.R., and
659 Kleinman, J.E. (2016). Mapping DNA methylation across development, genotype and
660 schizophrenia in the human frontal cortex. Nat. Neurosci. *19*, 40–47.
661 7. Hannon, E., Dempster, E., Viana, J., Burrage, J., Smith, A.R., Macdonald, R., St

662 Clair, D., Mustard, C., Breen, G., Therman, S., et al. (2016). An integrated genetic-
663 epigenetic analysis of schizophrenia: evidence for co-localization of genetic
664 associations and differential DNA methylation. Genome Biol. *17*, 176.
665 8. Hannon, E., Spiers, H., Viana, J., Pidsley, R., Burrage, J., Murphy, T.M., Troakes, C.,
666 Turecki, G., O'Donovan, M.C., Schalkwyk, L.C., et al. (2016). Methylation QTLs in the
667 developing brain and their enrichment in schizophrenia risk loci. Nat. Neurosci. *19*, 48–
668 54.
669 9. Hannon, E., Weedon, M., Bray, N., O'Donovan, M., and Mill, J. (2017). Pleiotropic
670 Effects of Trait-Associated Genetic Variation on DNA Methylation: Utility for Refining
671 GWAS Loci. Am. J. Hum. Genet. *100*, 954–959.
672 10. of the Psychiatric Genomics Consortium, S.W.G. (2014). Biological insights from
673 108 schizophrenia-associated genetic loci. Nature *511*, 421–427.
674 11. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D.,
675 Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of
676 genetic association studies using summary statistics. PLoS Genet. *10*, e1004383.
677 12. Stephens, M., and Balding, D.J. (2009). Bayesian statistical methods for genetic
678 association studies. Nat. Rev. Genet. *10*, 681–690.
679 13. Hormozdiari, F., van de Bunt, M., Segrè, A.V., Li, X., Joo, J.W.J., Bilow, M., Sul,
680 J.H., Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of GWAS
681 and eQTL Signals Detects Target Genes. Am. J. Hum. Genet. *99*, 1245–1260.
682 14. Roussos, P., Giakoumaki, S.G., Zouraraki, C., Fullard, J.F., Karagiorga, V.-E.,
683 Tsapakis, E.-M., Petraki, Z., Siever, L.J., Lencz, T., Malhotra, A., et al. (2016). The
684 Relationship of Common Risk Variants and Polygenic Risk for Schizophrenia to

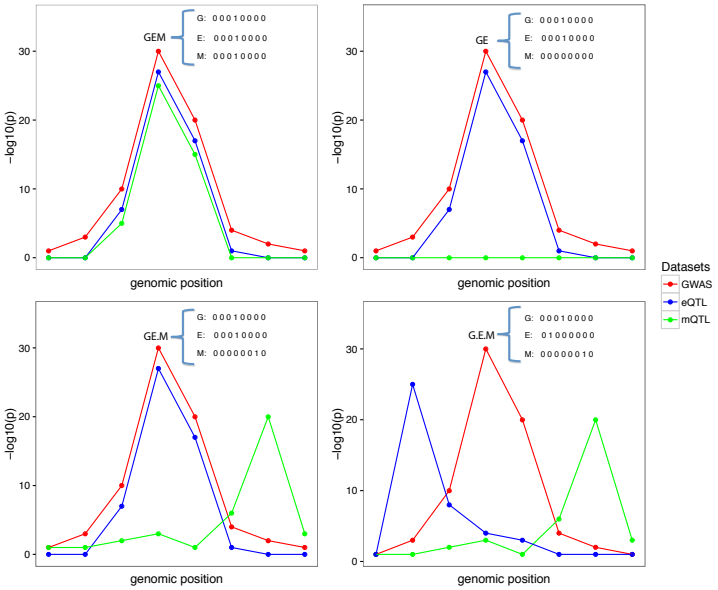
685 Sensorimotor Gating. Biol. Psychiatry *79*, 988–996.
686 15. Deignan, J., Luján, R., Bond, C., Riegel, A., Watanabe, M., Williams, J.T., Maylie, J.,
687 and Adelman, J.P. (2012). SK2 and SK3 expression differentially affect firing frequency
688 and precision in dopamine neurons. Neuroscience *217*, 67–76.
689 16. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.-R.,
690 Duncan, L., Perry, J.R.B., Patterson, N., Robinson, E.B., et al. (2015). An atlas of
691 genetic correlations across human diseases and traits. Nat. Genet. *47*, 1236–1241.
692 17. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W.J.H., Jansen, R., de
693 Geus, E.J.C., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for
694 large-scale transcriptome-wide association studies. Nat. Genet. *48*, 245–252.
695 18. Pickrell, J.K., Berisa, T., Liu, J.Z., Ségurel, L., Tung, J.Y., and Hinds, D.A. (2016).
696 Detection and interpretation of shared genetic influences on 42 human traits. Nat.
697 Genet. *48*, 709–717.
698 19. Shi, H., Mancuso, N., Spendlove, S., and Pasaniuc, B. (2016). Local genetic
699 correlation gives insights into the shared genetic architecture of complex traits. Doi.org
700 92668.
701 20. Tak, Y.G., and Farnham, P.J. (2015). Making sense of GWAS: using epigenomics
702 and genome engineering to understand the functional relevance of SNPs in non-coding
703 regions of the human genome. Epigenetics Chromatin *8*, 57.
704 21. Li, Y.I., Van De Geijn, B., Raj, A., Knowles, D.A., Petti, A.A., Golan, D., Gilad, Y.,
705 and Pritchard, J.K. RNA splicing is a primary link between genetic variation and
706 disease.
707 22. Wagner, J.R., Busche, S., Ge, B., Kwan, T., Pastinen, T., and Blanchette, M.

708 (2014). The relationship between DNA methylation, genetic and expression inter-
709 individual variation in untransformed human fibroblasts. Genome Biol. *15*, R37.
710 23. Du, X., Han, L., Guo, A.-Y., and Zhao, Z. (2012). Features of methylation and gene
711 expression in the promoter-associated CpG islands using human methylome data.
712 Comp. Funct. Genomics *2012*, 598987.
713 24. Guo, H., Fortune, M.D., Burren, O.S., Schofield, E., Todd, J. a, and Wallace, C.
714 (2015). Integration of disease association and eQTL data using a Bayesian
715 colocalisation approach highlights six candidate causal genes in immune-mediated
716 diseases. Hum. Mol. Genet. *24*, 3305–3313.
717 25. Wen, X., Pique-Regi, R., and Luca, F. (2017). Integrating molecular QTL data into
718 genome-wide genetic association analysis: Probabilistic assessment of enrichment and
719 colocalization. PLoS Genet. *13*,.
720 26. Wen, X., and Stephens, M. (2011). Bayesian methods for genetic association
721 analysis with heterogeneous subgroups: From meta-analyses to gene-environment
722 interactions.
723 27. Pickrell, J.K. (2014). Joint analysis of functional genomic data and genome-wide
724 association studies of 18 human traits. Am. J. Hum. Genet. *94*, 559–573.
725 28. Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: precision weights
726 unlock linear model analysis tools for RNA-seq read counts. Genome Biol. *15*, R29.
727 29. Shabalin, A.A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix
728 operations. Bioinformatics *28*, 1353–1358.
729 30. Wakefield, J. (2009). Bayes factors for genome-wide association studies:
730 comparison with P-values. Genet. Epidemiol. *33*, 79–86.

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

733

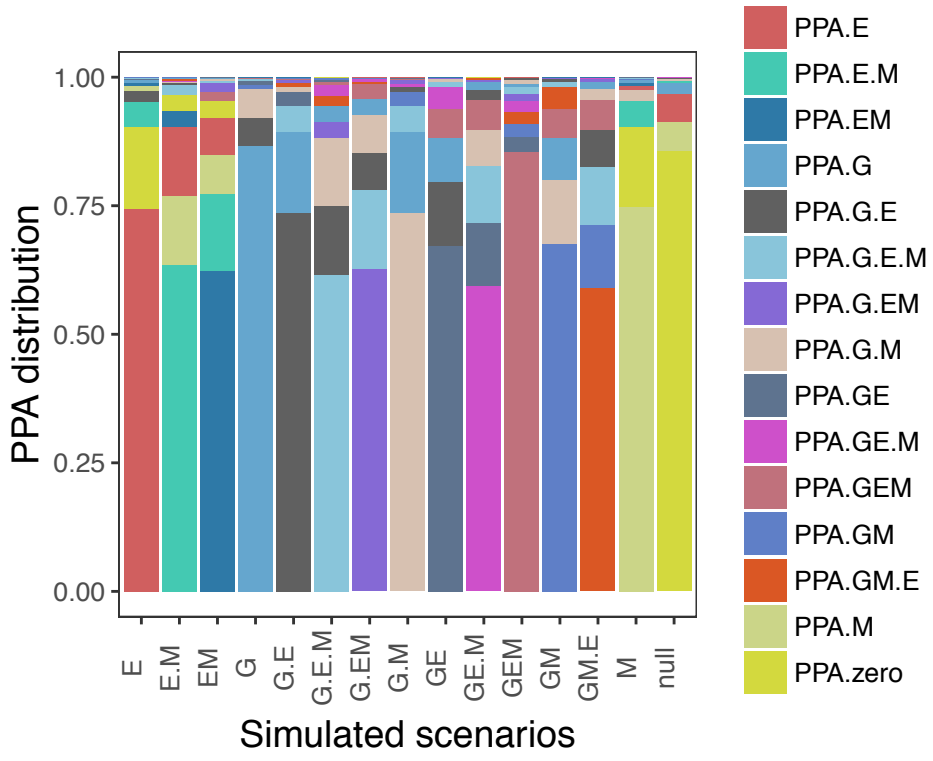
734 **FIGURES**



735

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

745 and E, and a different causal variant for M; Right plot bottom panel (G.E.M. scenario):
746 represents the case of three distinct causal variants behind each of the datasets
747 considered.



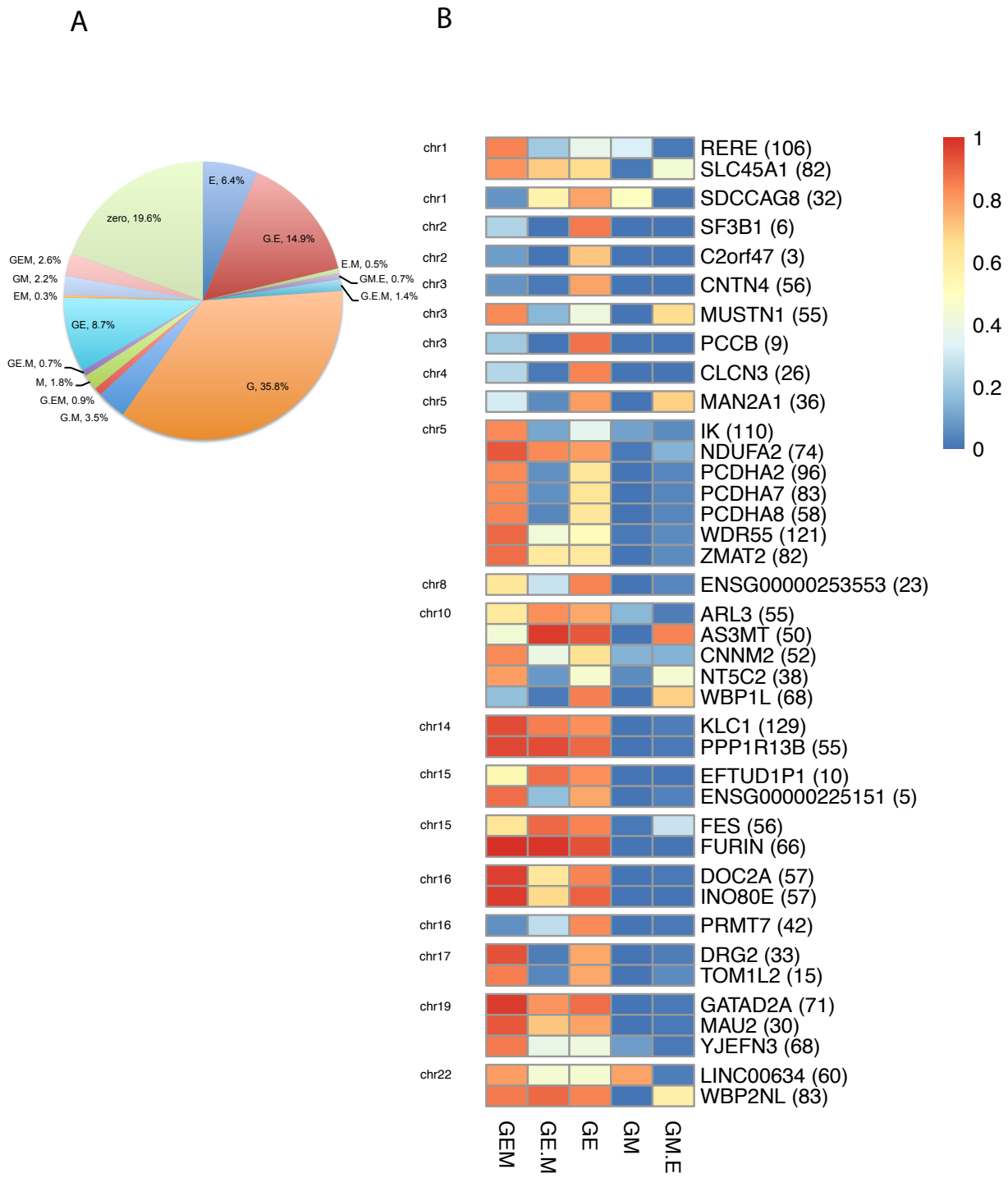
748

749 **Figure 2.** Results from simulations under colocalization/non-colocalization scenarios
750 using a sample size of 10,000 individuals for GWAS trait (denoted as G), 300 for eQTL
751 trait (denoted as E), and 300 for mQTL trait (denoted as M). X-axis shows all 15
752 simulated scenarios, e.g. G.E.M, three different causal variants for each of the three
753 traits; G.EM, 2 different causal variants, one for G and one shared between E and M;
754 GE, 1 shared causal variant for G and E; GE.M, 2 different causal variants, one shared
755 between G and E and one for M; GEM, one causal variant shared between all the three
756 traits. Y-axis shows the distribution of posterior probabilities under the simulated

757 scenario. The height of the bar represents the mean of the PPA for each configuration

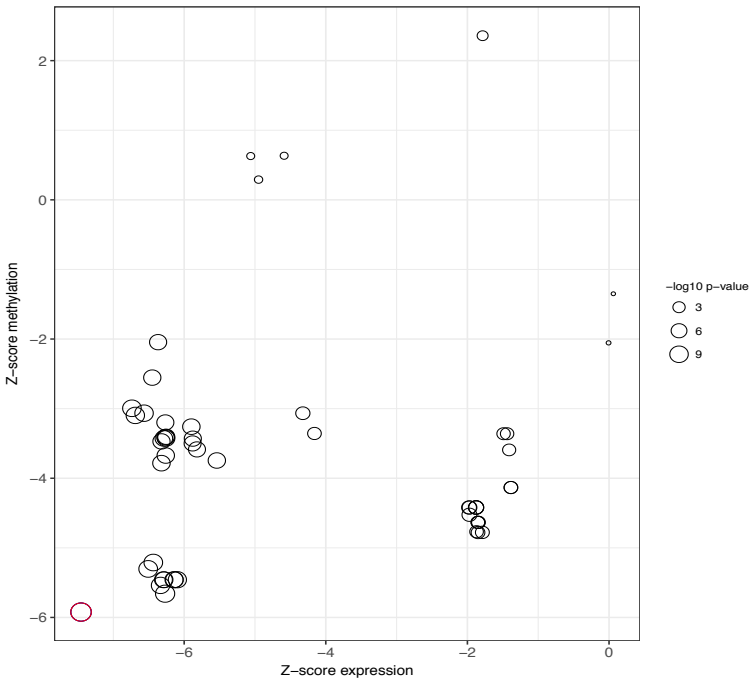
758 across simulations.

759


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

760

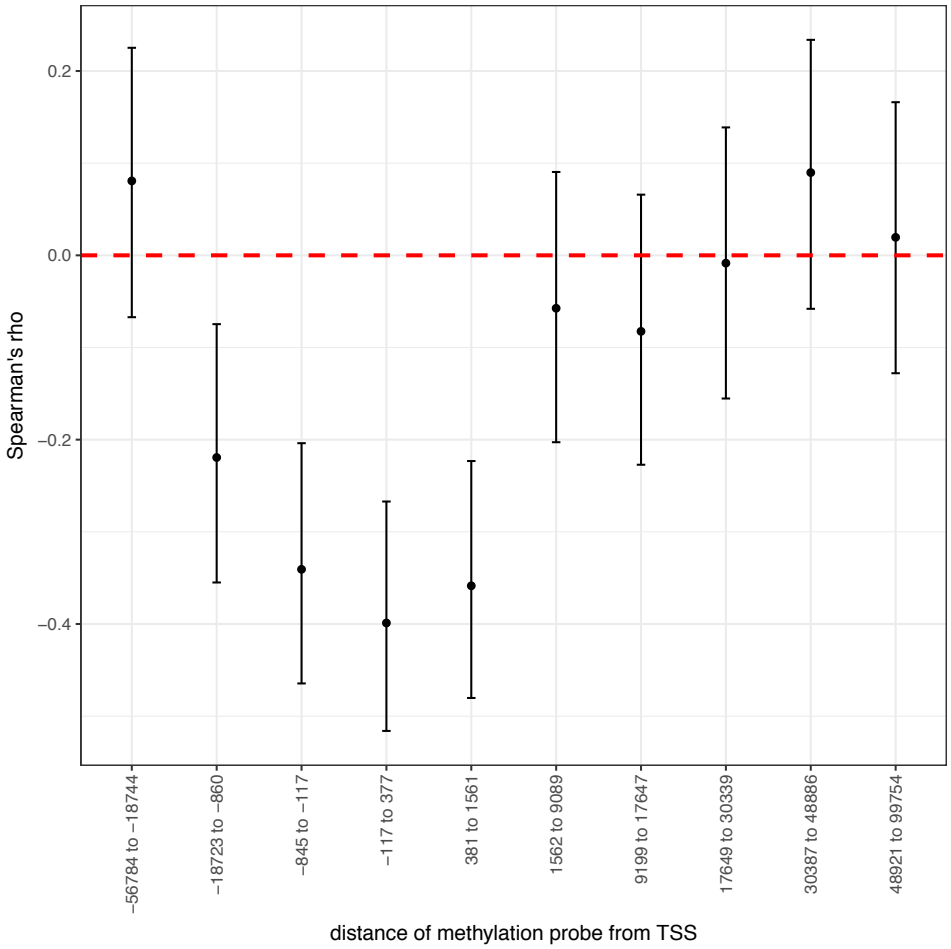
761 **Figure 3.** Summary of genes identified using three-trait colocalization within the SCZ LD
762 blocks. **A**. Mean posterior probability for each hypotheses computed using the cis-
763 regions overlapping the SCZ LD blocks. Sections of the pie chart represent the 15
764 scenarios representing the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



769

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

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



776

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

780

781

782

783 **TABLES**

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

785 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

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

787 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*

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

791

792 **SUPPORTING INFORMATION CAPTIONS**

793 **Figure S1**. Simulations under different sample sizes for all scenarios in *moloc* of three
794 traits (GWAS, eQTL, mQTL). All cases include 10,000 individuals in the GWAS dataset.
795 The variance explained by the trait was set to 0.01 for GWAS (1%), and to 0.1 (10%) for
796 the eQTL and mQTL.

797 **Figure S2.** Simulations under different sample sizes for all scenarios in *moloc* of three
798 traits ("a" is GWAS 1, "b" is GWAS 2, "c" is eQTL). All cases include 10,000 individuals
799 in the biomarker dataset. The variance explained by the trait was set to 0.01 for GWAS
800 1 and GWAS 2, and to 1% (10%) for the eQTL.

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

806 **Figure S4.** Illustration of two example of colocalization results with GWAS-eQTL-mQTL
807 in the *FURIN* locus and cg24888049 and *KCNN3* and cg09359103. −log10(p)
808 association p-values for GWAS (top panels), −log10(p) association p-values for eQTL

809 (center panels), -log10(p) association p-values for mQTL (bottom panels) with location
810 of genes at the FURIN locus, 200kb range. Highlighted is the region with the
811 methylation probe (15:91426667 +/- 20kb for cg24888049 in the FURIN locus, and
812 1:154839909 +/- 20kb for cg09359103 in the *KCNN3 locus*).

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

818

819 **Table S1**. False positive rates under the null of no association with any traits and
820 different combination of sample sizes (N simulations = 1000). We compute the false
821 positives by counting the proportion of simulations under the null that passed a posterior
822 probability supporting the "Detected Hypothesis" at a particular threshold
823 (PPA>=threshold). PPA.comb is PPA.abc + PPA.ab + PPA.ab.c.

824

825 **Table S2**. False positive rates under the null of no association with any traits and when
826 the causal variant is masked (N simulations = 1000). We compute the false positives by
827 counting the proportion of simulations under the null that passed a posterior probability
828 supporting the "Detected Hypothesis" at a particular threshold (PPA>=threshold).
829 PPA.comb is PPA.abc + PPA.ab + PPA.ab.c.

830

831 **Table S3**. False positive rates using 1000 and 500,000 simulations and our true sample
832 sizes at a cut off of 80% (N simulations = 1000 and 500,000). We compute the false
833 positives by counting the proportion of simulations under the null that passed a posterior
834 probability supporting the "Detected Hypothesis" at a particular threshold
835 (PPA>=threshold). PPA.comb is PPA.abc + PPA.ab + PPA.ab.c.
836
837 **Table S4**. Misclassification rates and power to detect the correct hypothesis for
838 combination of sample sizes (N simulations = 1000). We simulated each hypothesis of
839 interest indicated under the column "TRUE", and compute the number of misclassified
840 calls under other 14 scenarios by counting the proportion of simulations that passed a
841 posterior probability supporting a different hypothesis from the one simulated at a
842 particular threshold (PPA>=cut). We estimated power to distinguish a particular
843 hypothesis from the others by counting the number of correctly identified cases out of
844 1000 simulations. PPA.comb is PPA.abc + PPA.ab + PPA.ab.c.
845
846 **Table S5**. Misclassification rates and power to detect the correct hypothesis for
847 combination of sample sizes when the causal SNP is masked (N simulations = 1000).
848 The column labels are the same as Table S4.
849
850 **Table S6**. Misclassification rates and power to detect the correct hypothesis (N
851 simulations = 1000) under the sample sizes of 82,315 for GWAS, 467 for eQTL, and
852 121 for mQTL. The column labels are the same as Table S4.
853

854 **Table S7.** Effect of prior probability of co-localization for hypotheses of interest (GEM,
855 GE, GE.M) on the prior and posterior support for colocalisation in regions where there is
856 evidence for association across the three dataests (PPA.GE + PPA.GE.M + PPA.GEM
857 + PPA.G.E + PPA.G.EM + PPA.G.E.M). We repeated the moloc analysis with different
858 combinations of priors p2 and p3 while keeping p1 fixed to $1x10^{-4}$. The posteriors with
859 support for colocalization (PPA.coloc = PA.GE + PPA.GE.M + PPA.GEM) are
860 contrasted to the ones with support against colocalization (PPA.nocoloc = PPA.G.M
861 +PPA.G.EM+PPA.G.E.M) for our hypothesis of interest. For each analysis in which
862 PPA.GE + PPA.GE.M + PPA.GEM + PPA.G.E + PPA.G.EM + PPA.G.E.M >= 0.8, we
863 report the mean relative posterior support for colocalization, defined by PPA.coloc /
864 (PPA.nocoloc + PPA.coloc), and the mean relative prior support, averaged across all
865 regions. We find that the combination of priors p1=1e-04, p2=1e-06, p3=1e-07 has the
866 lowest difference between the prior and posterior expectation.

867

868 **Table S8**. 39 genes falling within SCZ LD blocks with a high probability of influencing
869 SCZ through expression or methylation (GEM, GE, GE.M>=0.8). "CHR", "START",
870 "STOP" indicate the positions of the regions considered; "PGC2_LDclump_info"
871 indicates position of the previously identified SCZ LD block; "PPA.combined" shows
872 maximum posterior probability reached by a gene for evidence of association with
873 GWAS and eQTL (PPA.GE + PPA.GE.M + PPA.GE.M); "numbMethyl" indicates the
874 number of methylation probes overlapping this region.

875

876 **Table S9**. Potentially novel SCZ loci. 45 candidate genes outside of the SCZ LD blocks
877 with a high probability of influencing SCZ (GEM, GE, GE.M>=0.8).

878

879 **Table S10.** Methylation-gene expression pairs that colocalize (PPA.GEM + PPA.EM +
880 PPA.G.EM >= 0.8) for DNA methylation probes that are proximal to the transcription
881 start site and enhancer sequences (20kb upstream to 2kb downstream of transcription
882 start site). “PPA.m” is PPA.GEM + PPA.EM + PPA.G.EM.

883

884 **Text S1**. Expanded methods, derivations and analyses.

1 Models

For three traits, the possible combinations of causal configurations (binary vectors with at most 1 entry equal to 1) can be assigned to one of the following 15 hypotheses listed below. Each hypothesis can be represented by a set of index set containing the traits each causal SNP associates with.

- $H_0 \longmapsto \{\phi\}$: No association in any trait
- $H_1 \longmapsto \{\{1\}\}$: One SNP is associated with trait 1 (a)
- $H_2 \longmapsto \{\{2\}\}$: One SNP is associated with trait 2 (b)
- $H_3 \longmapsto \{\{3\}\}$: One SNP is associated with trait 3 (c)
- $H_4 \longmapsto \{\{1,2\}\}$: One SNP is associated with trait 1 and 2 (ab)
- $H_5 \longmapsto \{\{2,3\}\}$: One SNP is associated with trait 2 and 3 (bc)
- $H_6 \longmapsto \{\{1,3\}\}$: One SNP is associated with trait 1 and 3 (ac)
- $H_7 \longmapsto \{\{1\},\{2\}\}$: One SNP is associated with trait 1, and a different SNP is associated with trait 2 (a.b)
- $H_8 \longmapsto \{\{2\},\{3\}\}$: One SNP is associated with trait 2, and a different SNP is associated with trait 3 (b.c)
- $H_9 \longmapsto \{\{1\},\{3\}\}$: One SNP is associated with trait 1, and a different SNP is associated with trait 3 (a.c)
- $H_{10} \longmapsto \{\{1\},\{2,3\}\}$: One SNP is associated with trait 1, and a different SNP is associated with trait 2 and 3 (a.bc)
- $H_{11} \longmapsto \{\{3\},\{1,2\}\}$: One SNP is associated with trait 3, and a different SNP is associated with trait 1 and 2 (c.ab)
- $H_{12} \longmapsto \{\{2\},\{1,3\}\}$: One SNP is associated with trait 2, and a different SNP is associated with trait 1 and 3 (b.ac)
- $H_{13} \longmapsto \{\{1\},\{2\},\{3\}\}$: One SNP is associated with trait 1, a different SNP is associated with trait 2, and a different SNP is associated with trait 3 (a.b.c)
- $H_{14} \longmapsto \{\{1,2,3\}\}$: One SNP is associated with trait 1 and 2 and 3 (abc)

To illustrate our notation, consider a region with 4 SNPs, then the set of combinations of causal configurations of each trait consistent with $H_4 = \{\{1,2\}\}$ (one SNP associates with trait 1 and 2) is the following set of combinations of binary vectors

$$\left\{ \begin{pmatrix} 1000 \\ 1000 \\ 0000 \end{pmatrix}, \begin{pmatrix} 0100 \\ 0100 \\ 0000 \end{pmatrix}, \begin{pmatrix} 0010 \\ 0010 \\ 0000 \end{pmatrix}, \begin{pmatrix} 0001 \\ 0001 \\ 0000 \end{pmatrix} \right\}. \tag{1}$$

This in turn can be represented by a set of tuples, indicating which SNP associates with which trait,

$$\{\{(1,1),(1,2)\},\{(2,1),(2,2)\},\{(3,1),(3,2)\},\{(4,1),(4,2)\}\}. \tag{2}$$

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

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

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

Next, we aim to assess the support for each hypothesis. We can compute the posterior probabilities given the data for each of these hypothesis by summing the probability of all the causal configurations consistent with the hypothesis. We find the probability of the data under each hypothesis using the following equation,

$$\frac{P(H_h|D)}{P(H_0|D)} = \sum_{S \in S_h} \frac{P(D \mid S)}{P(D \mid S_0)} \times \frac{P(S)}{P(S_0)}, \tag{4}$$

where the first ratio in this equation is a Bayes Factor for each configuration, and the second ratio is the prior odds of a configuration compared with the baseline configuration S_0.

We can now assess the evidence supporting each hypothesis. To compute the likelihood of the data and the prior probabilities of each hypothesis, we first compute a BF for each SNP and each trait. To connect these approximate Bayes factors to the alternative models for the SNP, we find the likelihood of the data and the prior probabilities π_s under the model that SNP i associates with traits indexed in **s**.

1.1 Bayes Factor of a SNP with traits indexed in s

First, a Bayes Factor for each SNP and each trait 1 and 2 and 3 is computed from summary statistics using the Asymptotic Bayes Factor derivation (Wakefield [2009], equation 3 in main text). This measures the relative support for a model in which the SNP is associated with the trait compared to the null model of no association.

Second, to assess the evidence for a SNP with j traits, let **s** be the set of trait indices for which SNP i is associated with. For example, $s = \{\phi\}$ means SNP i is associated with no trait; and $s = \{\{1, 2\}\}$ means SNP i is associated with trait 1 and 2. In this case, the BF for association of a SNP i with the two traits is:

$$BF_{i,\{1,2\}} = \sum_{i=1}^{Q} WABF_i^1 WABF_i^2 \tag{5}$$

In general, the strength of association of a SNP i with all traits indexed in **s** is:

$$BF_{i,\mathbf{s}} = \prod_{j \in s} BF_i^j \tag{6}$$

1.2 Prior Probabilities that SNP i associates with traits in s

In our three-trait model, the prior probability that SNP i associates with all traits indexed in a set is:

- π_ϕ: SNP i associates with no trait
- $\pi_{\{1\}}$: SNP i associates with trait 1

- $\pi_{\{2\}}$: SNP i associates with trait 2
- $\pi_{\{3\}}$: SNP i associates with trait 3
- $\pi_{\{1,2\}}$: SNP i associates with trait 1 and 2
- $\pi_{\{1,3\}}$: SNP i associates with trait 1 and 3
- $\pi_{\{2,3\}}$: SNP i associates with trait 2 and 3
- $\pi_{\{1,2,3\}}$: SNP i associates with trait 1 and 2 and 3

These are mutually exclusive, so we have:

$$\pi_\phi + \pi_{\{1\}} + \pi_{\{2\}} + \pi_{\{3\}} + \pi_{\{1,2\}} + \pi_{\{1,3\}} + \pi_{\{2,3\}} + \pi_{\{1,2,3\}} = 1 \tag{7}$$

1.3 Prior probabilities of each hypothesis

$$P(H_h \mid D) \propto \sum_{S \in S_h} P(D \mid S)P(S) \propto P(S \mid S \in S_h) \times \sum_{S \in S_h} P(D \mid S) \tag{8}$$

where *P(S)* is the prior probability of a configuration, $P(D \mid S)$ is the probability of the observed data D given a configuration S, and the sum is over all configurations S which are consistent with a given hypothesis H_h, where h=(1 ... 14). Since each binary vector that belongs to the same set S_h has the same prior probability, we can multiply likelihoods by one common prior probability for each hypothesis.

The prior probability of any one configuration in the different sets is:

- if $S \in S_0$, then $P(S) = \pi_\phi^Q$
- if $S \in S_1$, then $P(S) = Q \times \pi_0^{Q-1}\pi_{\{1\}}$
- if $S \in S_2$, then $P(S) = Q \times \pi_0^{Q-1}\pi_{\{2\}}$
- if $S \in S_3$, then $P(S) = Q \times \pi_0^{Q-1}\pi_{\{3\}}$
- if $S \in S_4$, then $P(S) = Q \times \pi_0^{Q-1}\pi_{\{1,2\}}$
- if $S \in S_5$, then $P(S) = Q \times \pi_0^{Q-1}\pi_{\{2,3\}}$
- if $S \in S_6$, then $P(S) = Q \times \pi_0^{Q-1}\pi_{\{1,3\}}$
- if $S \in S_7$, then $P(S) = Q(Q-1) \times \pi_0^{Q-2}\pi_{\{1\}}\pi_{\{2\}}$
- if $S \in S_8$, then $P(S) = Q(Q-1) \times \pi_0^{Q-2}\pi_{\{1\}}\pi_{\{3\}}$
- if $S \in S_9$, then $P(S) = Q(Q-1) \times \pi_0^{Q-2}\pi_{\{2\}}\pi_{\{2\}}$
- if $S \in S_{10}$, then $P(S) = Q(Q-1) \times \pi_0^{Q-2}\pi_{\{1\}}\pi_{\{2,3\}}$
- if $S \in S_{11}$, then $P(S) = Q(Q-1) \times \pi_0^{Q-2}\pi_{\{3\}}\pi_{\{1,2\}}$
- if $S \in S_{12}$, then $P(S) = Q(Q-1) \times \pi_0^{Q-2}\pi_{\{2\}}\pi_{\{1,3\}}$
- if $S \in S_{13}$, then $P(S) = Q(Q-1)(Q-2) \times \pi_0^{Q-3}\pi_{\{1\}}\pi_{\{2\}}\pi_{\{3\}}$
- if $S \in S_{14}$, then $P(S) = Q \times \pi_0^{Q-1}\pi_{\{1,2,3\}}$

1.4 Probability of the data under each hypothesis (the 3-trait special case)

Putting it all together. For three traits, the probability of the data under each hypothesis are:

- $\frac{P(H_0|D)}{P(H_0|D)} = 1$
- $\frac{P(H_1|D)}{P(H_0|D)} = \pi_{\{1\}} \times \sum_{i=1}^{Q} BF_{i,\{1\}}$
- $\frac{P(H_2|D)}{P(H_0|D)} = \pi_{\{2\}} \times \sum_{i=1}^{Q} BF_{i,\{2\}}$
- $\frac{P(H_3|D)}{P(H_0|D)} = \pi_{\{3\}} \times \sum_{i=1}^{Q} BF_{i,\{3\}}$
- $\frac{P(H_4|D)}{P(H_0|D)} = \pi_{\{1,2\}} \times \sum_{i=1}^{Q} BF_{i,\{1,2\}}$
- $\frac{P(H_5|D)}{P(H_0|D)} = \pi_{\{2,3\}} \times \sum_{i=1}^{Q} BF_{i,\{2,3\}}$
- $\frac{P(H_6|D)}{P(H_0|D)} = \pi_{\{1,3\}} \times \sum_{i=1}^{Q} BF_{i,\{1,3\}}$
- $\frac{P(H_7|D)}{P(H_0|D)} = \pi_{\{1\}} \times \pi_{\{2\}} \times \sum_{i=1}^{Q} \sum_{j=1}^{Q} BF_{i,\{1\}} BF_{j,\{2\}} I[i \neq j]$
- $\frac{P(H_8|D)}{P(H_0|D)} = \pi_{\{2\}} \times \pi_{\{3\}} \times \sum_{i=1}^{Q} \sum_{j=1}^{Q} BF_{i,\{2\}} BF_{j,\{3\}} I[i \neq j]$
- $\frac{P(H_9|D)}{P(H_0|D)} = \pi_{\{1\}} \times \pi_{\{3\}} \times \sum_{i=1}^{Q} \sum_{j=1}^{Q} BF_{i,\{1\}} BF_{j,\{3\}} I[i \neq j]$
- $\frac{P(H_{10}|D)}{P(H_0|D)} = \pi_{\{1\}} \times \pi_{\{2,3\}} \times \sum_{i=1}^{Q} \sum_{j=1}^{Q} BF_{i,\{1\}} BF_{j,\{2\}} BF_{j,\{3\}} I[i \neq j]$
- $\frac{P(H_{11}|D)}{P(H_0|D)} = \pi_{\{3\}} \times \pi_{\{1,2\}} \times \sum_{i=1}^{Q} \sum_{j=1}^{Q} BF_{i,\{3\}} BF_{j,\{1\}} BF_{j,\{2\}} I[i \neq j]$
- $\frac{P(H_{12}|D)}{P(H_0|D)} = \pi_{\{2\}} \times \pi_{\{1,3\}} \times \sum_{i=1}^{Q} \sum_{j=1}^{Q} BF_{i,\{2\}} BF_{j,\{1\}} BF_{j,\{3\}} I[i \neq j]$
- $\frac{P(H_{13}|D)}{P(H_0|D)} = \pi_{\{1\}} \times \pi_{\{2\}} \times \pi_{\{3\}} \times \sum_{i=1}^{Q} \sum_{j=1}^{Q} \sum_{k=1}^{Q} BF_{i,\{1\}} BF_{j,\{2\}} BF_{k,\{3\}} I[i \neq j \neq k]$
- $\frac{P(H_{14}|D)}{P(H_0|D)} = \pi_{\{1\}} \times \pi_{\{2\}} \times \pi_{\{3\}} \times \sum_{i=1}^{Q} BF_{i,\{1\}} BF_{i,\{2\}} BF_{i,\{3\}}$

where $BF_{i,\mathbf{s}}$ are computed from Wakefield's Asymptotic Bayes Factors for each SNP and each trait 1, 2 and 3 as outlined in eq. 3, $I[i \neq j \neq k]$ is an indicator that evaluates to 1 if SNP i and j and k are different and 0 otherwise, π are the prior probabilities that SNP i is the causal SNP under a specific model.

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

1.5 Probability of the data under hypotheses involving more than one independent associations

The probability of the data under scenarios of more than one causal variants [7 . . . 13] can be derived from the probability of the data under one causal variant [1 . . . 6, 14]. This generalization speeds up computation significantly.

- $\frac{P(H_7|D)}{P(H_0|D)} = \pi_{\{1\}} \times \pi_{\{2\}} \times \sum_{i=1}^{Q} BF_{i,\{1\}} \sum_{j=1}^{Q} BF_{j,\{2\}} - \left[\frac{\pi_{\{1\}} \times \pi_{\{2\}}}{\pi_{\{1,2\}}} \times \frac{P(H_4|D)}{P(H_0|D)} \right]$
- $\frac{P(H_8|D)}{P(H_0|D)} = \pi_{\{2\}} \times \pi_{\{3\}} \times \sum_{i=1}^{Q} BF_{i,\{2\}} \sum_{j=1}^{Q} BF_{j,\{3\}} - \left[\frac{\pi_{\{2\}} \times \pi_{\{3\}}}{\pi_{\{2,3\}}} \times \frac{P(H_5|D)}{P(H_0|D)} \right]$
- $\frac{P(H_9|D)}{P(H_0|D)} = \pi_{\{1\}} \times \pi_{\{3\}} \times \sum_{i=1}^{Q} BF_{i,\{1\}} \sum_{j=1}^{Q} BF_{j,\{3\}} - \left[\frac{\pi_{\{1\}} \times \pi_{\{3\}}}{\pi_{\{1,3\}}} \times \frac{P(H_6|D)}{P(H_0|D)} \right]$
- $\frac{P(H_{10}|D)}{P(H_0|D)} = \pi_{\{1\}} \times \pi_{\{2,3\}} \times \sum_{i=1}^{Q} BF_{i,\{1\}} \sum_{j=1}^{Q} BF_{j,\{2,3\}} - \left[\frac{\pi_{\{1\}} \times \pi_{\{2,3\}}}{\pi_{\{1,2,3\}}} \times \frac{P(H_{14}|D)}{P(H_0|D)} \right]$
- $\frac{P(H_{11}|D)}{P(H_0|D)} = \pi_{\{3\}} \times \pi_{\{1,2\}} \times \sum_{i=1}^{Q} BF_{i,\{3\}} \sum_{j=1}^{Q} BF_{j,\{1,2\}} - \left[\frac{\pi_{\{3\}} \times \pi_{\{1,2\}}}{\pi_{\{1,2,3\}}} \times \frac{P(H_{14}|D)}{P(H_0|D)} \right]$
- $\frac{P(H_{12}|D)}{P(H_0|D)} = \pi_{\{2\}} \times \pi_{\{1,3\}} \times \sum_{i=1}^{Q} BF_{i,\{2\}} \sum_{j=1}^{Q} BF_{j,\{1,3\}} - \left[\frac{\pi_{\{2\}} \times \pi_{\{1,3\}}}{\pi_{\{1,2,3\}}} \times \frac{P(H_{14}|D)}{P(H_0|D)} \right]$

In general, across a set of j traits $\{1, 2, 3..., j\}$, for a hypothesis h, where there are more than one independent associations among the traits (i.e. $|h| > 1$), the probability of the data supporting h can be written as

$$\frac{P(H_h|D)}{P(H_0|D)} = \prod_{s \in h} \pi_s \sum_{i=1}^{Q} BF_{i,s} - \frac{\prod_{s \in h} \pi_s}{\pi_t} \sum_{i=1}^{Q} \pi_t BF_{i,t}, \tag{9}$$

where 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, $j = 3$, $h = \{ \{1, 2\}, \{3\} \}$ and:

$$\frac{P(H_h|D)}{P(H_0|D)} = \pi_{\{1,2\}} \pi_{\{3\}} \sum_{i=1}^{Q} BF_{i\{1,2\}} \sum_{j=1}^{Q} BF_{j\{3\}} - \frac{\pi_{\{1,2\}} \pi_{\{3\}}}{\pi_{\{1,2,3\}}} \sum_{i=1}^{Q} \pi_{\{1,2,3\}} BF_{i\{1,2,3\}} \tag{10}$$

Since we pre-compute all the single terms in the above equation, the computations are greatly simplified.

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:

$$\frac{P(H_h|D)}{P(H_0|D)} = \pi_{\{1,2\}} \pi_{\{4,5\}} \sum_{i=1}^{Q} BF_{i\{1,2\}} \sum_{j=1}^{Q} BF_{j\{4,5\}} - \frac{\pi_{\{1,2\}} \pi_{\{4,5\}}}{\pi_{\{1,2,4,5\}}} \sum_{i=1}^{Q} \pi_{\{1,2,4,5\}} BF_{i\{1,2,4,5\}} \tag{11}$$

References

Claudia Giambartolomei, Damjan Vukcevic, Eric E Schadt, Lude Franke, Aroon D Hingorani, Chris Wallace, and Vincent Plagnol. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS genetics*, 10:e1004383, May 2014. ISSN 1553-7404. doi: 10.1371/journal.pgen.1004383.

Jon Wakefield. Bayes factors for genome-wide association studies: comparison with p-values. *Genetic epidemiology*, 33:79–86, January 2009. ISSN 0741-0395. doi: 10.1002/gepi.20359.

Xiaoquan Wen and Matthew Stephens. Bayesian methods for genetic association analysis with heterogeneous subgroups: from meta-analyses to gene-environment interactions. *The annals of applied statistics*, 8(1):176, 2014.

Supplementary Figures

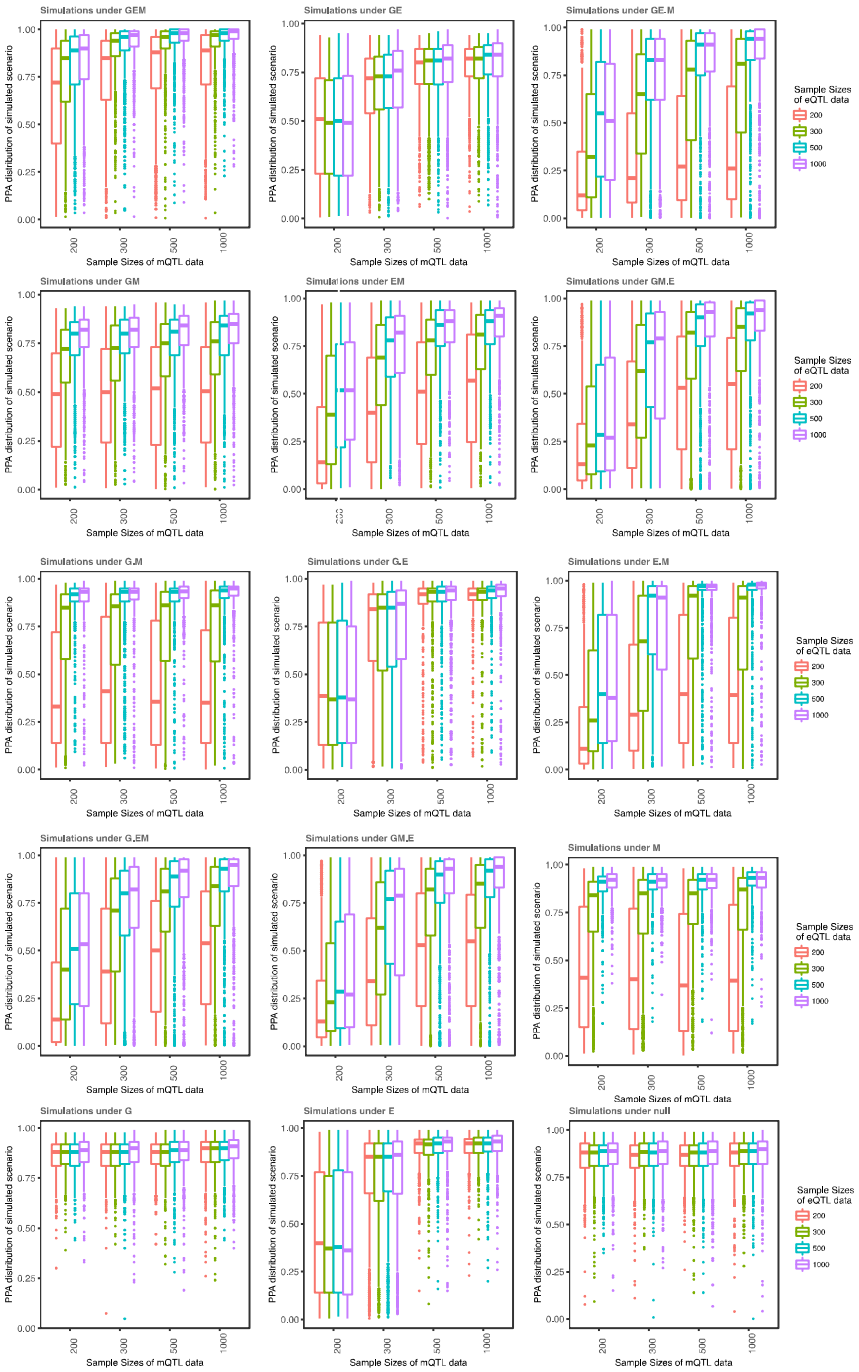


Figure S1. Simulations under different sample sizes for all scenarios in *moloc* of three traits (GWAS, eQTL, mQTL). All cases include 10,000 individuals in the

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.

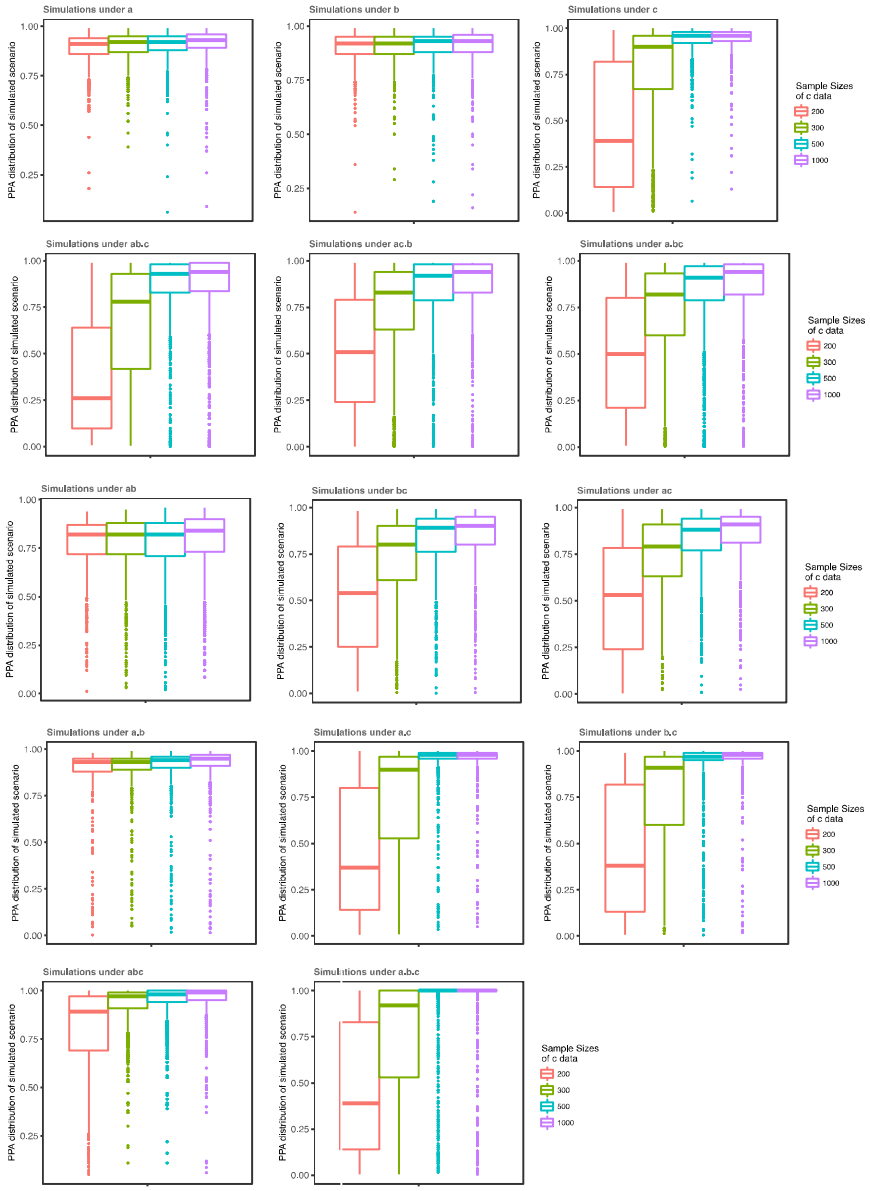


Figure S2. Simulations under different sample sizes for all scenarios in *moloc* of three traits ("a" is GWAS 1, "b" is GWAS 2, "c" is eQTL). All cases include 10,000 individuals in the biomarker dataset. The variance explained by the trait was set to 0.01 for GWAS 1 and GWAS 2, and to 1% (10%) for the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

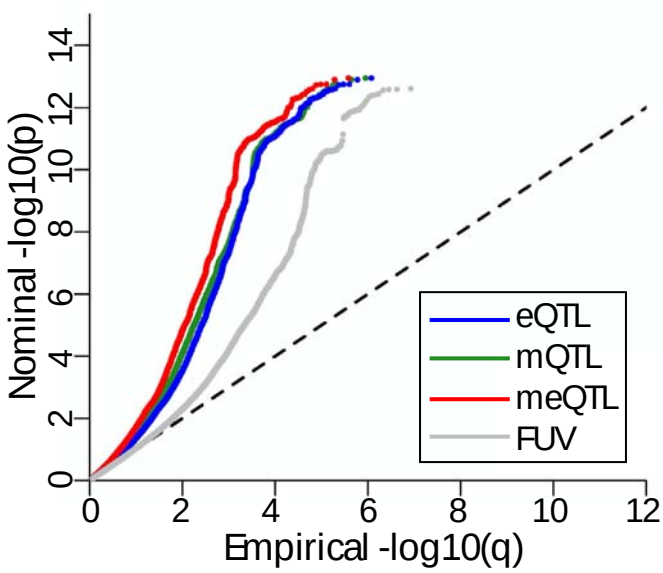


Figure S3. Stratified Q-Q plots for eQTL and mQTLs in the SCZ GWAS. Enrichment of each functional category (blue line: eQTL; green: mQTL; red: combined eQTL and mQTL) illustrate the estimated increase in terms of the proportion of p-values expected in each scenario and in the functionally unannotated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

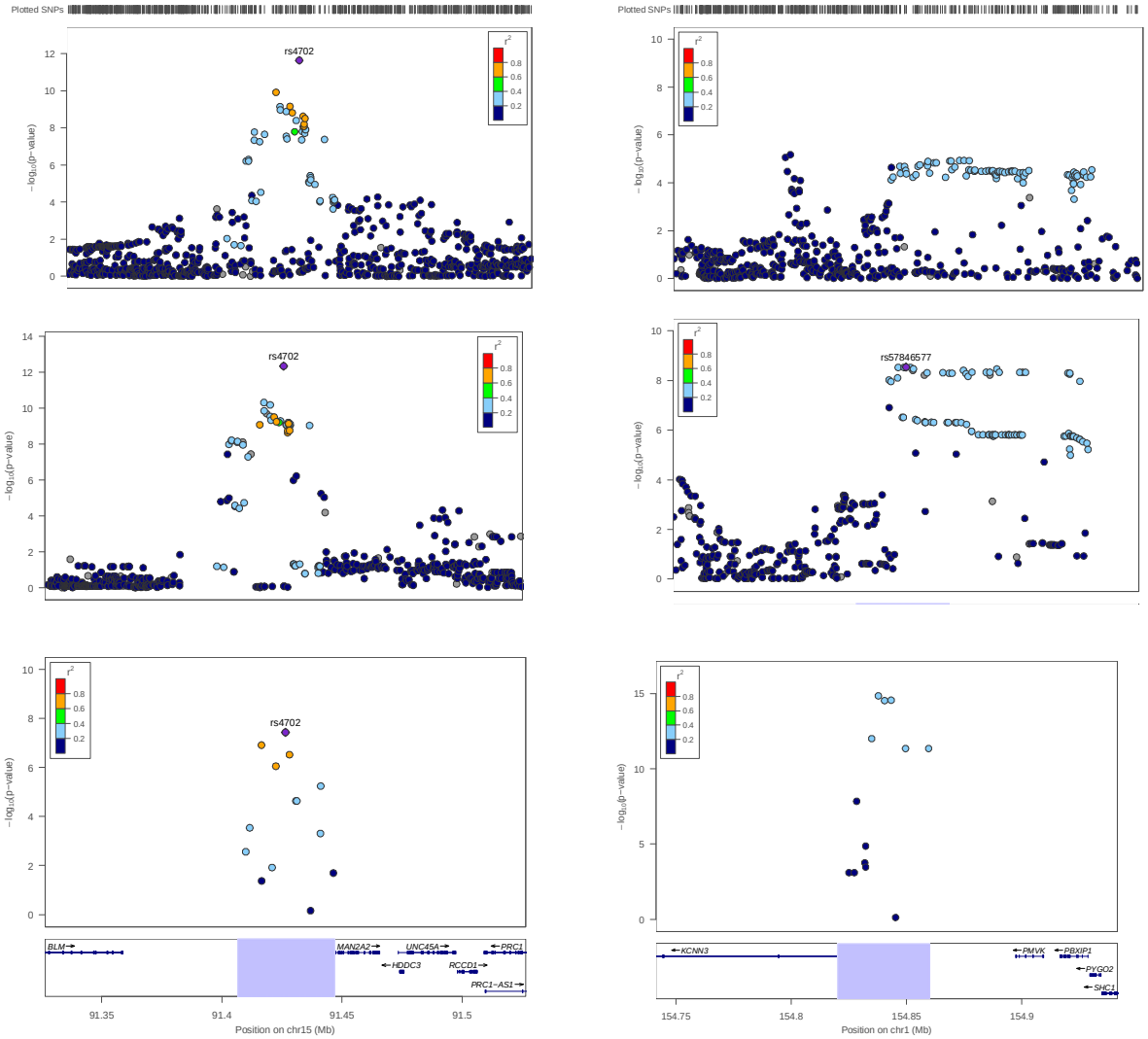


Figure S4. Illustration of two example of colocalization results with GWAS-eQTL-mQTL in the *FURIN* locus and cg24888049 and *KCNN3* and cg09359103. −log10(p) association p-values for GWAS (top panels), −log10(p) association p-values for eQTL (center panels), -log10(p) association p-values for mQTL (bottom 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*).

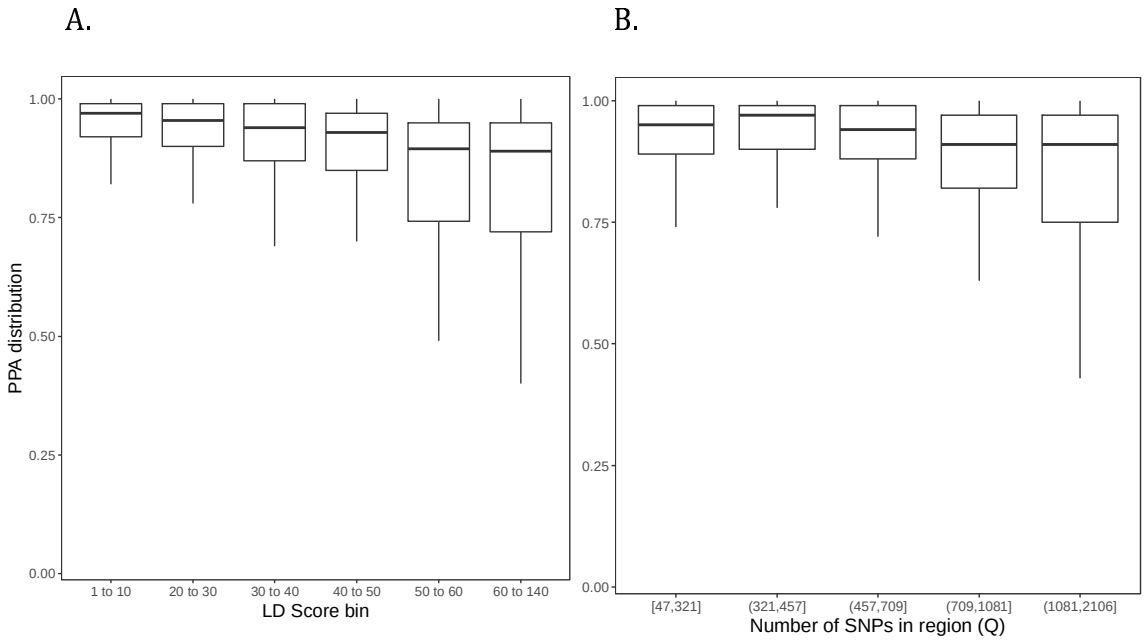


Figure S5. Simulation analysis with a shared causal variant between three studies using a sample size of 10,000 individuals for GWAS trait, 300 for eQTL trait, and 300 for mQTL trait. Y-axis is the posterior probability of a shared causal causal variant for all three datasets (PPA.abc). **A.** X-axis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