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

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

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

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

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

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

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

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

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

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

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

Correspondence:

Dr. Panos Roussos

Icahn School of Medicine at Mount Sinai

Department of Psychiatry and Department of Genetics and Genomic Science and Institute for Multiscale Biology

One Gustave L. Levy Place,

New York, NY, 10029, USA

Panagiotis.roussos@mssm.edu

Dr. Claudia Giambartolomei

University of California, Los Angeles, Los Angeles

Department of Pathology and Laboratory Medicine,

Los Angeles, CA 90095, USA

claudia.giambartolomei@gmail.com

ABSTRACT

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

INTRODUCTION

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

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

88 mQTL data that were previously generated in human DLPFC tissue (N=121) to
89 investigate epigenetic variation in SCZ[11]. Integration of multiple phenotypes helps better
90 characterize the genes predisposing to complex diseases such as SCZ.
91
92 **MATHERIALS AND METHODS**
93 **Method Description**
94 We extended the model of Pickrell[3] and *coloc*[2] to analyze jointly multiple traits. For each
95 variant, we assume a simple linear regression model to relate the vector of phenotypes
96 $\vec{y}$ or a log-odds generalized linear model for the case-control dataset, and the vector of
97 genotypes $\vec{x}$. Under this model the expectation of the trait is:
98

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

99
100 We define a genomic region containing *Q* variants, for example a *cis* region around
101 expression or methylation probe. We are interested in a situation where summary
102 statistics (effect size estimates and standard errors) are available for all datasets in a
103 genomic region with *Q* variants.

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

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

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

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

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

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

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

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

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

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

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

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

For example,

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

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

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

153 where *n* is the number of traits considered, *h* is the configuration of interest out of the *H*
154 possible configurations and π is the prior probability according to how many traits the
155 SNP is associated with.
156 We set the prior probability that a SNP is the causal one for each trait, to be identical
157 ($\pi^{(1)} = \pi^{(2)} = \pi^{(3)}$) and refer to this as p1. We also set the prior probability that a SNP is
158 associated with two traits, to be identical ($\pi^{(1,2)} = \pi^{(1,3)} = \pi^{(2,3)}$) and refer to this as *p2*.
159 We refer to the prior probability that a SNP is causal for all traits as *p3*.
160 Finally, the posterior probability supporting configuration *h* among *H* possible
161 configurations, is computed:
162
163

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

164
165
166 **GWAS dataset**
167 Summary statistics for genome-wide SNP association with Schizophrenia were
168 obtained from the Psychiatric Genomics Consortium-Schizophrenia Workgroup (PGC-
169 SCZ) primary meta-analysis (35,476 cases and 46,839 controls) [9].
170
171 **Expression QTL (eQTL) analysis**
172 This analysis used RNA sequence data on individuals of European-ancestry (*N* =
173 467) from post-mortem DLPFC (Brodmann areas 9 and 46), and imputed genotypes
174 based on the Phase 1 reference panel from the 1,000 Genomes Project as previously
175 described[10]. MatrixEQTL[14] was used to fit an additive linear model between the

176 expression of 15,791 genes and imputed SNP genotypes within a 1 Mb window around
177 the transcription start site for each gene, including covariates for ancestry, diagnosis,
178 and known and hidden variables detected by surrogate variable analysis, as described
179 elsewhere[10]. Overall, the model identified 2,154,331 significant *cis*-eQTL, (i.e., SNP–
180 gene pairs within 1 Mb of a gene) at a false discovery rate (FDR) ≤ 5%, for 13,137
181 (80%) genes.
182
183 **Methylation QTL (mQTL) analysis**
184 DNA methylation of postmortem tissue homogenates of the dorsolateral
185 prefrontal cortex (DLPFC, Brodmann areas 9 and 46) from non-psychiatric adult
186 Caucasian control donors (age > 13, N=121) was measured using the Illumina
187 HumanMethylation450 ("450k") microarray (which measures CpG methylation across
188 473,058 probes covering 99% of RefSeq gene promoters). DNA for genotyping was
189 obtained from the cerebella of samples with either the Illumina Human Hap 650v3,1M
190 Duo V3, or Omni 5M BeadArrays and merged across the three platforms following
191 imputation to the 1000 Genomes Phase 3 reference panel as previously described[11].
192 The mQTL analyses was then conducted using the R package MatrixEQTL[14], fitting an
193 additive linear model up to 20kb distance between each SNP and CpG analyzed,
194 including covariates for ancestry and global epigenetic variation.
195
196 **Moloc Analysis**
197 Previous to running the analyses, the GWAS and eQTL datasets were filtered by
198 poorly imputed SNPs (kept only SNPs with Rsq > 0.3). The Major Histocompatibility

199 (MHC) region (chr 6: 25 Mb - 35 Mb) was excluded from all co-localization analyses due
200 to the extensive linkage disequilibrium. We applied a genic-centric approach, defined
201 *cis*-regions based on a 50kb upstream/downstream from the start/end of each gene,
202 since our goal is to link risk variants with changes in gene expression. We evaluated all
203 methylation probes overlapping the *cis*-region. The number of cis-regions/methylation
204 pairs is higher than the count of genes because, on average, there are more than one
205 methylation sites per gene. Common SNPs were evaluated in the colocalization
206 analysis for each gene, and each methylation probe, and GWAS. In total, 14,115 *cis*-
207 regions and 534,962 unique *cis*-regions/methylation probes were tested. Genomic
208 regions were analyzed only if 50 SNPs or greater were in common between all the
209 datasets. Across all of the analyses, a posterior probability equal to, or greater than,
210 80% for each configuration was considered evidence of colocalization.

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

216

217 **Simulations**

218 We simulated genotypes from sampling with replacement among haplotypes of
219 SNPs with a minor allele frequency of at least 5% found in the phased 1000 Genomes
220 Project within 49 genomic regions that have been associated with type 1 diabetes (T1D)

221 susceptibility loci (excluding the major histocompatibility complex (MHC) as previously
222 described[15]. These represent a range of region sizes and genomic topography that
223 reflect typical GWAS hits in a complex trait. For each trait, two, or three "causal
224 variants" were selected at random, and a Gaussian distributed quantitative trait for
225 which each causal variant SNP explains a specified proportion of the variance was
226 simulated. All analyses were conducted in R.

227 **RESULTS**

228 **Overview of the *moloc* method**

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

243

244 **Sample size requirements**

245 We explored the posterior probability under different sample sizes. **Figure S1**
246 illustrate the posterior probability distribution across all of the possible scenarios that
247 includes three traits: GWAS, eQTL and mQTL. With a GWAS sample size of 10,000

248 and eQTL and mQTL sample sizes of 300, the method provides reliable evidence to
249 detect a shared causal variant behind the GWAS and another trait (median posterior
250 probability of any hypothesis >50%). Although in this paper we analyze GWAS, eQTL
251 and mQTL, our method can be applied to any combinations of traits, including 2 GWAS
252 traits and an eQTL dataset. We explored the minimum sample size required when
253 analyzing two GWAS datasets (GWAS1, GWAS2) and one eQTL (**Figure S2**). The
254 method provides reliable evidence for all hypotheses when the two GWAS sample sizes
255 are 10,000 and eQTL sample size reaches 300.

256 It is instructive to observe where evidence for other hypotheses is distributed. **Figure**
257 **2A** illustrates the accuracy of our approach under different scenarios where two or three
258 causal variants are shared. For example, under simulations of one shared variant for
259 GWAS and eQTL and a second variant for mQTL (GE.M), on average 60% of the
260 evidence points to the simulated scenario, while 12% point to GE, 12% to G.E.M and
261 7.2% to GEM.
262 We examined whether the inclusion of a third trait increases power to detect
263 colocalization in comparison to running analysis with two traits on the same data. For
264 the colocalization of three traits, we consider any scenarios where there is evidence of
265 colocalization between the GWAS and eQTL datasets, i.e. GE, GE.M, GEM. We note
266 that using only eQTL data recovers fewer colocalizations with GWAS loci when there is
267 truly one single causal variant across the datasets (**Figure 2B**), providing additional
268 support for increasing power by adding mQTLs. In this study we focus on the
269 colocalization of GWAS with eQTL (GEM, GE.M or GE scenarios), due to smaller

270 sample size and limited power in the mQTL dataset.

271

272 **Choice of priors**

273 The algorithm requires the definition of prior probabilities at the SNP level for the
274 association with one (p1), two (p2), or three traits (p3). We set the priors to *p1* = 1 x 10^{-}
275 ^{4}, *p2* = 1 x 10^{-6}, *p3* = 1 x 10^{-7} based on simulations and exploratory analysis of genome-
276 wide enrichment of GWAS risk variants in eQTLs and mQTLs. We set the prior
277 probability that a variant is associated with one trait as 1 x 10^{-4} for GWAS, eQTL and
278 mQTL, assuming that each genetic variant is equally likely a priori to affect gene
279 expression or methylation or disease. This estimate has been suggested in the literature
280 for GWAS[16] and used in similar methods[6]. In **Figure S3**, we find eQTLs and mQTLs to
281 be similarly enriched in GWAS, justifying our choice of the same prior probability of
282 association across the two traits. These values are also suggested by a crude
283 approximation of *p2* and *p3* from the common genome-wide significant SNPs across the
284 three dataset.

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

291

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

293 We applied our method to SCZ GWAS using eQTLs derived from 467 CMC
294 samples and mQTL from 121 individuals. Our aim is to identify the genes important for
295 disease through colocalization of GWAS variants with changes in gene expression and
296 DNA methylation. We analyzed associations genome-wide, and report results both
297 across previously identified GWAS loci, and across potentially novel loci. While we
298 consider all 15 possible scenarios of colocalization, here we focus on gene discovery
299 due to higher power in our eQTL dataset, by considering the combined probabilities of
300 cases where the same variant is shared across all three traits GWAS, eQTLs and
301 mQTLs (GEM > 0.8) or scenarios where SCZ risk loci are shared with eQTL only (GE >
302 0.8 or GE.M > 0.8) (**Table 1**). We identified 1,173 cis-regions/methylation pairs with
303 posterior probability above 0.8 that are associated with all three traits (GEM), or eQTLs
304 alone (GE or GE.M). These biologically relevant scenarios affect overall 97 unique
305 genes. Fifty-six out of the 97 candidate genes influence SCZ, gene expression and
306 methylation (GEM>=0.8). One possible scenario is that the variants in these genes
307 could be influencing the risk of SCZ through methylation, although other potential
308 interpretations such as pleiotropy should be considered.

309

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

311 We examined whether moloc with 3 traits enhance power for GWAS and eQTL
312 colocalization compared to using 2 traits. Colocalization analysis of only GWAS and

313 eQTL traits identified 11 genes with GE >= 0.8 and two genes within a previously
314 associated SCZ LD block (*FURIN* and *PCCB*), which indicates a ~9 fold increase in the
315 genes discovered when we consider an additional trait. The 97 genes identified with a
316 high probability of influencing SCZ (GEM, GE, GE.M>=0.8) are listed in Table S3. The
317 89 additional genes that were found by adding methylation include genes such as
318 *AS3MT* that would have been missed by only GWAS and eQTL colocalization.
319

320 **Loci overlapping reported SCZ LD blocks**

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

336 provide illustrative examples of SCZ association with expression and regulatory DNA
337 region in the FURIN locus (**Figure 4** and **Figure S4**).
338
339 **Potentially novel SCZ loci**
340 We found 53 unique genes in below genome-wide significant regions (novel SCZ
341 associations). All genes were far from a SCZ LD block (more than 50kb, **Table S4**), and
342 contained SNPs with p-values for association with SCZ ranging from 10^{-4} to 10^{-9}. These
343 genes will likely be identified using just the GWAS signal if the sample size is increased.
344 *KCNN3* is among these genes which encodes an integral membrane protein that forms
345 a voltage-independent calcium-activated channel. It regulates neuronal excitability by
346 contributing to the slow component of synaptic afterhyperpolarization[17].
347
348 **Comparison with previous findings**
349 Our gene discovery analysis replicate several previous results[10,18–20] (**Table 2**).
350 One recent study[20] performed mQTL analysis on 1714 individuals from three
351 independent sample cohorts, and used colocalization between mQTLs and SCZ GWAS
352 to identify genomic regions associated with both schizophrenia and methylation. From
353 their analysis, 32 methylation probes have a posterior probability of colocalizing with
354 SCZ >=0.8. We analyzed 15 out of these methylation probes, and reproduced 7 for
355 colocalization of GWAS and methylation (combined GEM, GM, GM.E >=0.8,
356 cg00585072, cg02951883, cg08607108, cg08772003, cg19624444, cg26732615).
357 Hannon et al.[20] annotated these regions with 26 genes. Since we integrate eQTL
358 information, our analysis points to specific genes responsible for these associations.

359

360 **Association of gene expression with methylation**

361 DNA methylation is one the best studied epigenetic modifications**.** Methylation
362 can alter gene expression by disrupting transcription factor binding sites (with variable
363 consequences to expression depending on the TF), or by attracting methyl-binding
364 proteins that initiate chromatin compaction and gene silencing. Therefore methylation
365 can be associated with both increased or decreased gene expression[21,22]. Increased
366 CpG methylation in promoter regions is usually associated with silencing of gene
367 expression[23]. However, in genome-wide expression and methylation studies, the
368 correlation of methylation and gene expression is low or the pattern of association is
369 mixed, even for CpG methylation within promoter[22]. One challenge of examining DNA
370 methylation with expression is the uncertainty of linking the CpG site with a specific
371 gene, especially for CpG sites that are distal to any coding genes. To overcome this
372 challenge, we sought to explore direction of effects of methylation and expression, for
373 gene expression and DNA methylation that colocalize with posterior probability above
374 0.8 (GEM, EM, and G.EM scenarios) (**Table 1**). Overall, we tested 2,227 DNA
375 methylation and gene expression pairwise interactions and found a significant negative
376 correlation between the effect sizes of methylation and expression in the proximity of
377 the transcription start site (**Figure 5**).

DISCUSSION

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

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

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

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

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

CONFLICTS OF INTEREST

None reported.

SUPPLEMENTAL DATA

Supplemental Data include four figures and four tables.

CONSORTIA

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

ACKNOWLEDGMENTS

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

This work was supported by the National Institutes of Health (R01AG050986 Roussos and R01MH109677 Roussos), Brain Behavior Research Foundation (20540 Roussos), Alzheimer's Association (NIRG-340998 Roussos) and the Veterans Affairs (Merit grant BX002395 Roussos). Additionally, this work was supported in part through the

445 computational resources and staff expertise provided by Scientific Computing at the
446 Icahn School of Medicine at Mount Sinai.
447 Data were generated as part of the CommonMind Consortium supported by funding
448 from Takeda Pharmaceuticals Company Limited, F. Hoffman-La Roche Ltd and NIH
449 grants R01MH085542, R01MH093725, P50MH066392, P50MH080405,
450 R01MH097276, RO1-MH-075916, P50M096891, P50MH084053S1, R37MH057881
451 and R37MH057881S1, HHSN271201300031C, AG02219, AG05138 and MH06692.
452 Brain tissue for the study was obtained from the following brain bank collections: the
453 Mount Sinai NIH Brain and Tissue Repository, the University of Pennsylvania
454 Alzheimer's Disease Core Center, the University of Pittsburgh NeuroBioBank and Brain
455 and Tissue Repositories and the NIMH Human Brain Collection Core. CMC Leadership:
456 Pamela Sklar, Joseph Buxbaum (Icahn School of Medicine at Mount Sinai), Bernie
457 Devlin, David Lewis (University of Pittsburgh), Raquel Gur, Chang-Gyu Hahn (University
458 of Pennsylvania), Keisuke Hirai, Hiroyoshi Toyoshiba (Takeda Pharmaceuticals
459 Company Limited), Enrico Domenici, Laurent Essioux (F. Hoffman-La Roche Ltd), Lara
460 Mangravite, Mette Peters (Sage Bionetworks), Thomas Lehner, Barbara Lipska (NIMH).
461

462 **WEB RESOURCES**

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

FIGURE TITLES AND LEGENDS

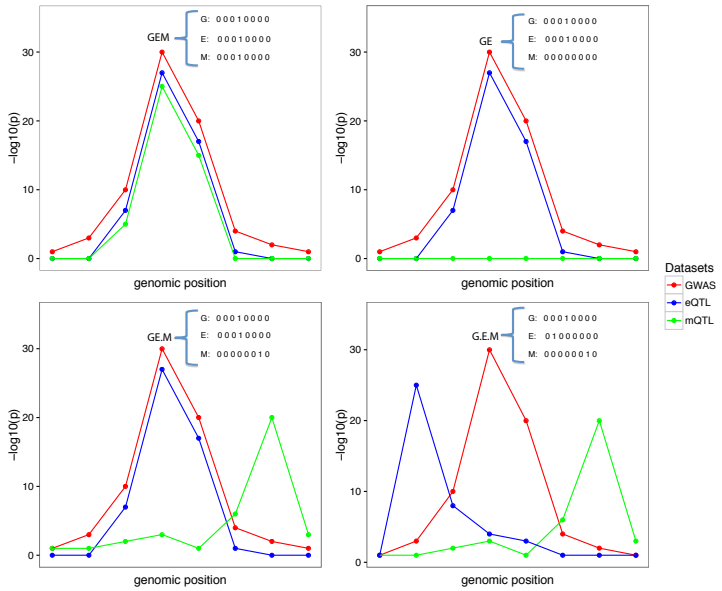
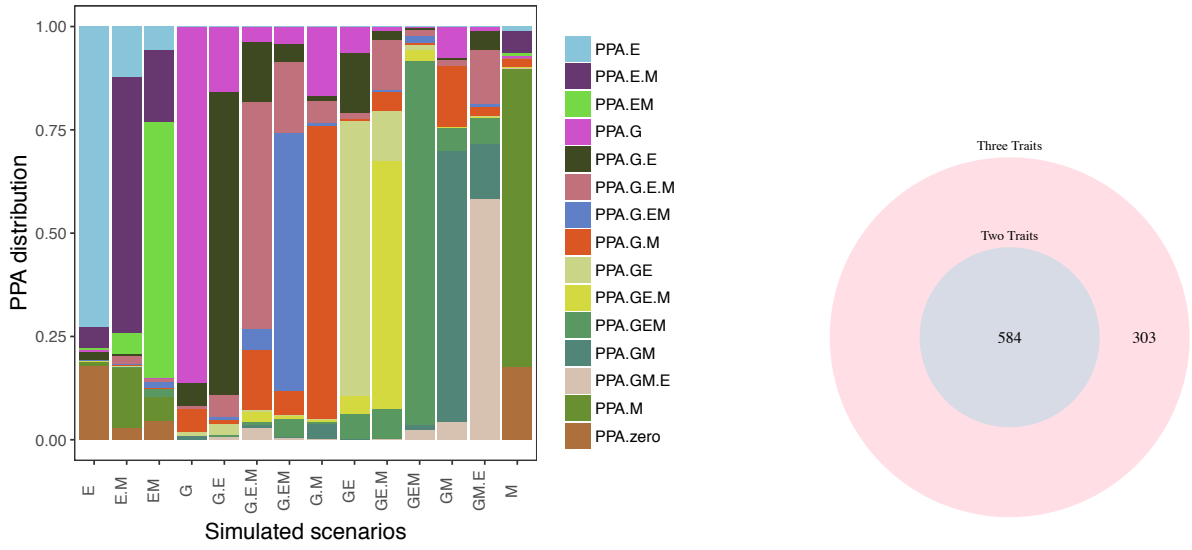


Figure 1. Graphical representation of four possible configurations at a locus with 8 SNPs in common across three traits. The traits are labeled as G, E, M representing GWAS (G), eQTL (E), and mQTL (M) datasets, respectively. Each plot represents one possible configuration, which is a possible combination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



481

482

483 **Figure 2. A.** Results from simulations under colocalization/non-colocalization scenarios
484 using a sample size of 10,000 individuals for GWAS trait (denoted as G), 300 for eQTL
485 trait (denoted as E), and 300 for mQTL trait (denoted as M). X-axis shows all 15
486 simulated scenarios, e.g. G.E.M, three different causal variants for each of the three
487 traits; G.EM, 2 different causal variants, one for G and one shared between E and M;
488 GE, 1 shared causal variant for G and E; GE.M, 2 different causal variants, one shared
489 between G and E and one for M; GEM, one causal variant shared between all the three
490 traits. The x-axis shows the distribution of posterior probabilities under the simulated
491 scenario. **B.** Venn diagram comparing number of colocalization of two traits (coloc PPA
492 >=80%) with three traits (moloc PPA GE + GE.M + GEM) in simulations with one causal
493 variant shared between all the three traits (GEM). Results include 887 out of 1,000
494 simulations passing 80% threshold for colocalization. The variance explained by the trait
495 is 0.01 for GWAS (1%), and 0.1 (10%) for the eQTL and mQTL.

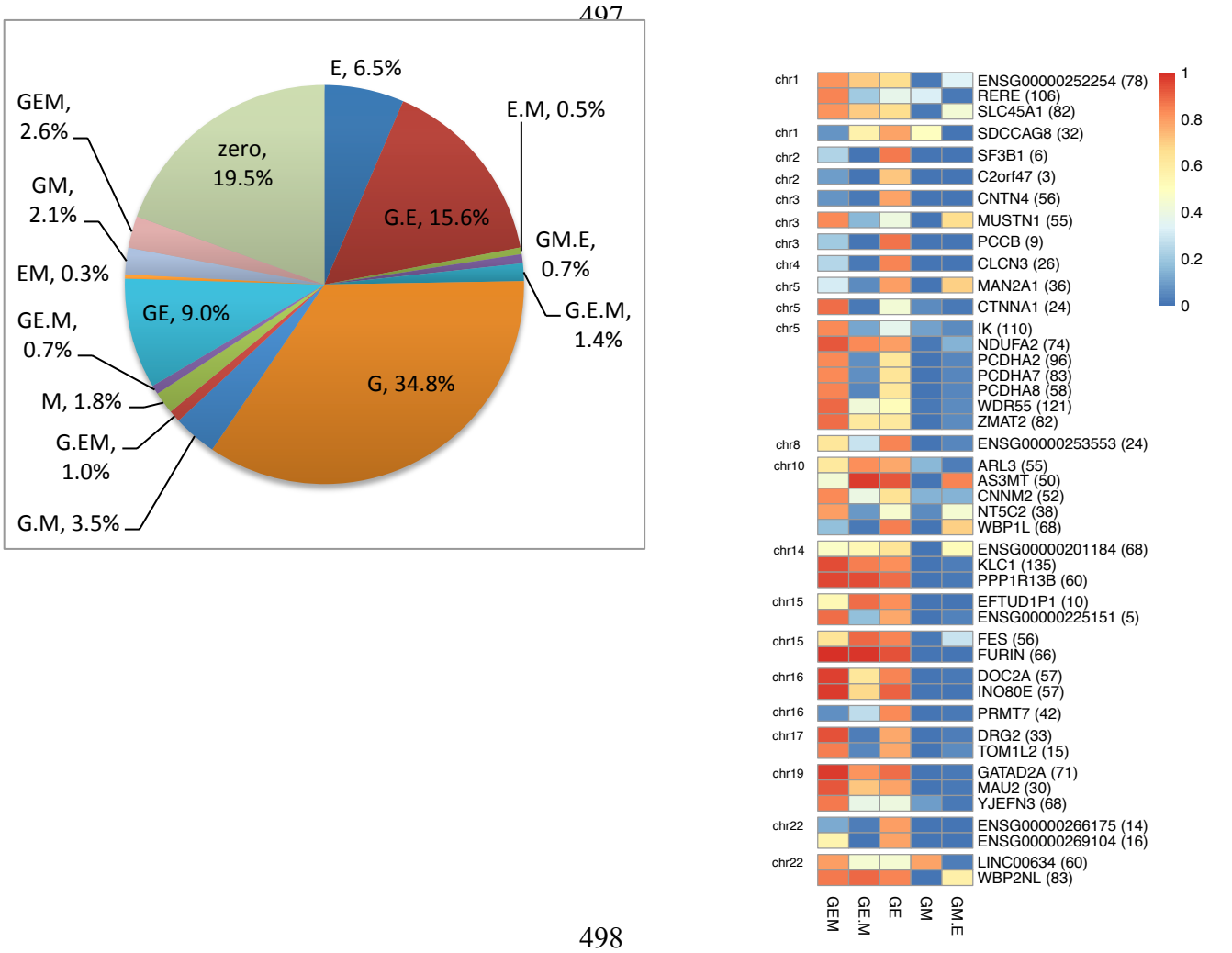
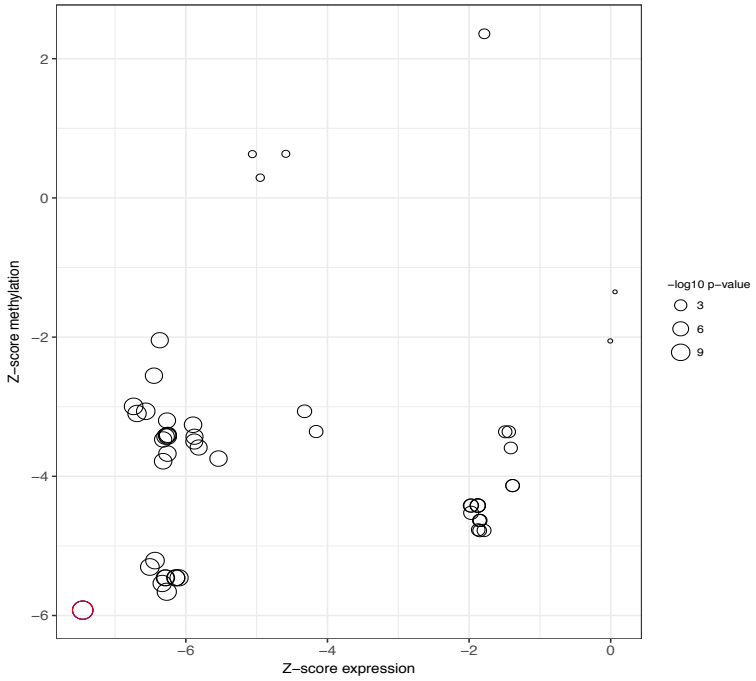


Figure 3. Summary of genes identified using three-trait colocalization within the SCZ LD blocks. **A**. Mean posterior probability for each hypotheses computed using the cis-regions overlapping the SCZ LD blocks. Sections of the pie chart represent the 15 scenarios representing the possible combination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

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



507

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

513

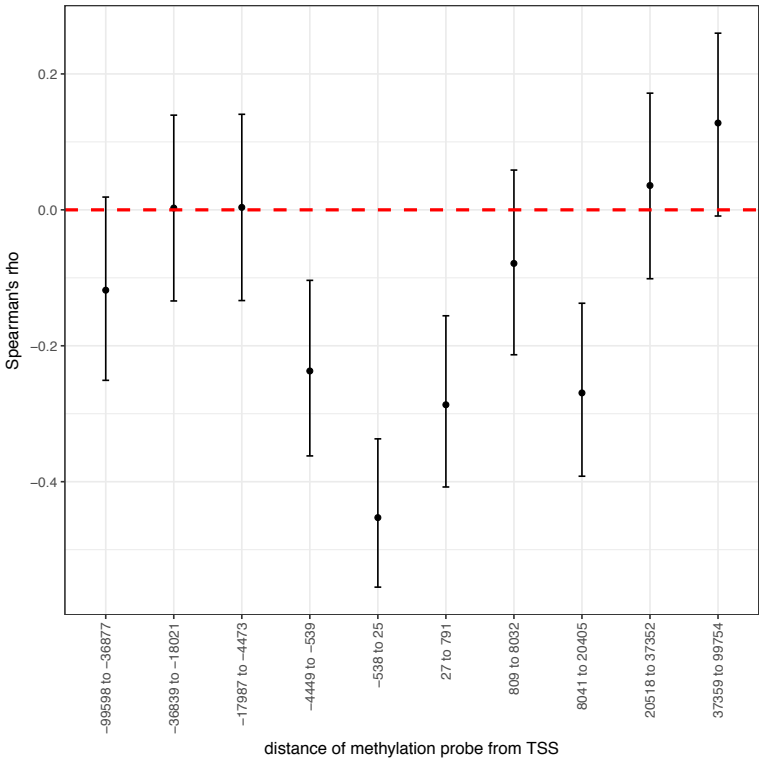


Figure 5. Spearman correlation of eQTL and mQTL effect estimates by distance from transcription start site of the gene. Intervals of methylation probe distance from TSS were estimated based on 10 equal size bins.

524

525 **TABLE TITLES AND LEGENDS**

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

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

528

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

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

531

532

533

534 **REFERENCES**

535 1. Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five Years of
536 GWAS Discovery. Am. J. Hum. Genet. *90*, 7–24.
537 2. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace,
538 C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic
539 association studies using summary statistics. PLoS Genet. *10*, e1004383.
540 3. Pickrell, J.K., Berisa, T., Liu, J.Z., Ségurel, L., Tung, J.Y., and Hinds, D.A. (2016).
541 Detection and interpretation of shared genetic influences on 42 human traits. Nat.
542 Genet. *48*, 709–717.
543 4. Pickrell, J.K. (2014). Joint analysis of functional genomic data and genome-wide
544 association studies of 18 human traits. Am. J. Hum. Genet. *94*, 559–573.
545 5. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W.J.H., Jansen, R., de
546 Geus, E.J.C., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for
547 large-scale transcriptome-wide association studies. Nat. Genet. *48*, 245–252.
548 6. Hormozdiari, F., van de Bunt, M., Segrè, A.V., Li, X., Joo, J.W.J., Bilow, M., Sul, J.H.,
549 Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of GWAS and
550 eQTL Signals Detects Target Genes. Am. J. Hum. Genet. *99*, 1245–1260.
551 7. Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A.L., Kraft,
552 P., and Pasaniuc, B. (2014). Integrating Functional Data to Prioritize Causal Variants in
553 Statistical Fine-Mapping Studies. PLoS Genet. *10*, e1004722.
554 8. Consortium, E.N.C.O.D.E.P., and Bernstein, B.E. et al. (2012). An integrated
555 encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.
556 9. of the Psychiatric Genomics Consortium, S.W.G. (2014). Biological insights from 108

557 schizophrenia-associated genetic loci. Nature *511*, 421–427.
558 10. Fromer, M., Roussos, P., Sieberts, S.K., Johnson, J.S., Kavanagh, D.H., Perumal,
559 T.M., Ruderfer, D.M., Oh, E.C., Topol, A., Shah, H.R., et al. (2016). Gene expression
560 elucidates functional impact of polygenic risk for schizophrenia. Nat. Neurosci. *19*,
561 1442–1453.
562 11. Jaffe, A.E., Gao, Y., Deep-Soboslay, A., Tao, R., Hyde, T.M., Weinberger, D.R., and
563 Kleinman, J.E. (2016). Mapping DNA methylation across development, genotype and
564 schizophrenia in the human frontal cortex. Nat. Neurosci. *19*, 40–47.
565 12. Wakefield, J. (2009). Bayes factors for genome-wide association studies:
566 comparison with P-values. Genet. Epidemiol. *33*, 79–86.
567 13. Ip, H., Jansen, R., Abdellaoui, A., Bartels, M., Boomsma, D.I., and Nivard, M.G.
568 (2017). Stratified Linkage Disequilibrium Score Regression reveals enrichment of eQTL
569 effects on complex traits is not tissue specific. bioRxiv.
570 14. Shabalin, A.A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix
571 operations. Bioinformatics *28*, 1353–1358.
572 15. Wallace, C. (2013). Statistical testing of shared genetic control for potentially related
573 traits. Genet. Epidemiol. *37*, 802–813.
574 16. Stephens, M., and Balding, D.J. (2009). Bayesian statistical methods for genetic
575 association studies. Nat. Rev. Genet. *10*, 681–690.
576 17. Deignan, J., Luján, R., Bond, C., Riegel, A., Watanabe, M., Williams, J.T., Maylie, J.,
577 and Adelman, J.P. (2012). SK2 and SK3 expression differentially affect firing frequency
578 and precision in dopamine neurons. Neuroscience *217*, 67–76.
579 18. Gusev, A., Mancuso, N., Finucane, H.K., Reshef, Y., Song, L., Safi, A., Oh, E., O

580 ’donovan, M.C., Katsanis, N., Crawford, G.E., et al. TITLE: Transcriptome-wide
581 association study of schizophrenia and chromatin activity yields mechanistic disease
582 insights.
583 19. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery,
584 G.W., Goddard, M.E., Wray, N.R., Visscher, P.M., et al. (2016). Integration of summary
585 data from GWAS and eQTL studies predicts complex trait gene targets. Nat. Genet. *48*,
586 481–487.
587 20. Hannon, E., Dempster, E., Viana, J., Burrage, J., Smith, A.R., Macdonald, R., St
588 Clair, D., Mustard, C., Breen, G., Therman, S., et al. (2016). An integrated genetic-
589 epigenetic analysis of schizophrenia: evidence for co-localization of genetic
590 associations and differential DNA methylation. Genome Biol. *17*, 176.
591 21. Tak, Y.G., and Farnham, P.J. (2015). Making sense of GWAS: using epigenomics
592 and genome engineering to understand the functional relevance of SNPs in non-coding
593 regions of the human genome. Epigenetics Chromatin *8*, 57.
594 22. Wagner, J.R., Busche, S., Ge, B., Kwan, T., Pastinen, T., and Blanchette, M.
595 (2014). The relationship between DNA methylation, genetic and expression inter-
596 individual variation in untransformed human fibroblasts. Genome Biol. *15*, R37.
597 23. Du, X., Han, L., Guo, A.-Y., and Zhao, Z. (2012). Features of methylation and gene
598 expression in the promoter-associated CpG islands using human methylome data.
599 Comp. Funct. Genomics *2012*, 598987.
600 24. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.-R.,
601 Duncan, L., Perry, J.R.B., Patterson, N., Robinson, E.B., et al. (2015). An atlas of
602 genetic correlations across human diseases and traits. Nat. Genet. *47*, 1236–1241.

603 25. Li, Y.I., Van De Geijn, B., Raj, A., Knowles, D.A., Petti, A.A., Golan, D., Gilad, Y.,
604 and Pritchard, J.K. RNA splicing is a primary link between genetic variation and
605 disease.
606 26. Wen, X., Pique-Regi, R., and Luca, F. (2017). Integrating molecular QTL data into
607 genome-wide genetic association analysis: Probabilistic assessment of enrichment and
608 colocalization. PLoS Genet. *13*,.
609 27. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery,
610 G.W., Goddard, M.E., Wray, N.R., Visscher, P.M., et al. (2016). Integration of summary
611 data from GWAS and eQTL studies predicts complex trait gene targets. Nat. Genet. *48*,
612 481–487.
613 28. Hauberg, M.E., Zhang, W., Giambartolomei, C., Franzén, O., Morris, D.L., Vyse,
614 T.J., Ruusalepp, A., Sklar, P., Schadt, E.E., Björkegren, J.L.M., et al. (2017). Large-
615 Scale Identification of Common Trait and Disease Variants Affecting Gene Expression.
616 Am. J. Hum. Genet. *100*, 885–894.
617