Bayesian Genetic Colocalization Test of Two Traits Using *coloc*

Danielle Rasooly,[1,4] Gina M. Peloso,[2] and Claudia Giambartolomei[3]

[1]Division of Aging, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts
[2]Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts
[3]Non-Coding RNAs and RNA-Based Therapeutics, Istituto Italiano di Tecnologia, Via Morego, Genova, Italy
[4]Corresponding author: *drasooly@bwh.harvard.edu*

Published in the Human Genetics section

Genetic colocalization is an approach for determining whether a genetic variant at a particular locus is shared across multiple phenotypes. Genome-wide association studies (GWAS) have successfully mapped genetic variants associated with thousands of complex traits and diseases. However, a large proportion of GWAS signals fall in non-coding regions of the genome, making functional interpretation a challenge. Colocalization relies on a Bayesian framework that can integrate summary statistics, for example those derived from GWAS and expression quantitative trait loci (eQTL) mapping, to assess whether two or more independent association signals at a region of interest are consistent with a shared causal variant. The results from a colocalization analysis may be used to evaluate putative causal relationships between omics-based molecular measurements and a complex disease, and can generate hypotheses that may be followed up by tailored experiments. In this article, we present an easy and straightforward protocol for conducting a Bayesian test for colocalization of two traits using the 'coloc' package in R with summary-level results derived from GWAS and eQTL studies. We also provide general guidelines that can assist in the interpretation of findings generated from colocalization analyses. © 2022 Wiley Periodicals LLC.

Basic Protocol: Performing a genetic colocalization analysis using the 'coloc' package in R and summary-level data
Support Protocol: Installing the 'coloc' R package

Keywords: bayesian • colocalization • eQTL • fine-mapping • genetic epidemiology • GWAS

How to cite this article:
Rasooly, D., Peloso, G. M., & Giambartolomei, C. (2022). Bayesian genetic colocalization test of two traits using *coloc*. *Current Protocols*, *2,* e627. doi: 10.1002/cpz1.627

INTRODUCTION

Genetic colocalization is a statistical approach that can assess shared genetic etiology across multiple traits. Profiling genetic variation has elucidated the polygenic architecture of complex traits and uncovered the genetic and molecular basis for disease. The genome-wide association study (GWAS) is a commonly used study design for identifying genetic variations that associate with a particular phenotype by systematically scanning genetic variants across the human genome (Uffelmann et al., 2021). Rather than

CURRENT PROTOCOLS
A Wiley Brand

Current Protocols e627, Volume 2
Published in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/cpz1.627

focusing on a few biological candidate genes, GWAS tests millions of genetic variants throughout the human genome for their association with a human trait or disease, without any prior predilection for a specific locus within the genome. This has led to a range of GWAS discoveries, enhancing our understanding of the underlying biology of diseases and genomic traits (Visscher et al., 2017). GWAS has generated insights on novel disease-causing genes and mechanisms, and has led to predictions about subsets of the population that are at higher risk based on genotype (Tam et al., 2019; Visscher, Brown, McCarthy, & Yang, 2012).

While GWAS has furthered our genetic understanding of complex diseases and traits, multiple technological and methodological limitations have made the interpretation of results of genetic studies challenging (Cano-Gamez & Trynka, 2020; Goldstein, 2009; Manolio, 2010). The majority of GWAS findings reside in non-coding regions of the genome, with limited ability to interpret these variants given an unknown functional understanding of these regions (Edwards, Beesley, French, & Dunning, 2013; Hrdlickova, de Almeida, Borek, & Withoff, 2014). Studies on the specific biological processes, such as the genes and cells that mediate a genetic association between a single nucleotide polymorphism (SNP) and a trait, are needed for further enhancing our interpretation of the effects of the genetic variants. For example, to identify potential causal genes of a trait, a straightforward approach is to overlap a GWAS with annotations of the closest gene; however, candidate genes are frequently not the genes located immediate to the GWAS signal (e.g., see GTEx Consortium et al., 2017). Integrating GWAS with studies on expression quantitative trait loci (eQTL), loci that can explain a proportion of the genetic variance of gene expression, has been an emerging direction for the analysis of genome function (Nica & Dermitzakis, 2013). Trait-associated variants are substantially overrepresented among eQTLs relative to expectation, highlighting the crucial role eQTLs play in the causal pathway between SNP and trait (Morley et al., 2004; Nicolae et al., 2010; Schadt et al., 2003). However, determining whether a GWAS signal is shared with an eQTL signal is challenging because linkage disequilibrium (LD), the nonrandom association of alleles at different loci, can obscure the true causal variant(s).

Colocalization analyses can determine whether the same causal variant is responsible for both eQTL and GWAS signals (Giambartolomei et al., 2014). Colocalization assesses shared variants from independent signals of association between pairs of traits (e.g., such as those from GWAS and eQTL studies) at a particular locus, where a locus can be defined as a region around the transcription start site (TSS) of a gene or any region of interest that may contain a significant signal of association, and under the assumption that the region has at most one causal variant per trait. Two traits that share a causal variant are referred to as 'colocalized', i.e., there is evidence for both traits sharing a causal mechanism. For an investigation using eQTL mapping and a GWAS of a complex disease, identifying the same SNP that affects both measurements may suggest that the expression of that gene influences disease pathogenesis. The colocalization method relies on a Bayesian framework (Giambartolomei et al., 2014) by integrating evidence over all possible configurations for the variants across a locus, where each configuration can be assigned a hypothesis:

- H_0: No association with either trait
- H_1: Association with trait 1, not with trait 2
- H_2: Association with trait 2, not with trait 1
- H_3: Association with trait 1 and trait 2, two independent SNPs
- H_4: Association with trait 1 and trait 2, one shared SNP

Evidence in support of H_4 suggests colocalization between the two traits, indicating that a shared causal variant is associated with both trait 1 and trait 2. The outputs include posterior probabilities that can be easily evaluated in favor of (or against) the colocalization hypothesis. For each hypothesis, the corresponding posterior probability is computed (termed 'PP.H0', 'PP.H1,' 'PP.H2,' 'PP.H3,' and 'PP.H4'), and the posterior probability of H_4 (termed 'PP.H4') determines whether the data support a single causal variant associated with both traits. This method can be implemented on both case-control and quantitative traits (Giambartolomei et al., 2014).

The 'coloc' package in R can be used to easily conduct this analysis and test for local genetic colocalization of two phenotypes using single SNP summary statistic results of association studies of each phenotype (Giambartolomei et al., 2014; Wallace, 2020, 2021). The 'coloc' package outputs the five posterior probabilities that can be used to interpret whether or not these two potentially related traits share a genetic causal variant in a genomic region of interest. This method is driven by information usually available from summary statistics (as opposed to individual-level data) to derive probabilities for association with each trait, and therefore is favorable from an implementation standpoint, considering frequent data-sharing restrictions. Additionally, the method allows for large systematic comparisons across multiple association studies (Giambartolomei et al., 2014), encouraging "phenome-wide" searches for trait connections that have not been discovered as of yet.

In this article, we present a protocol for conducting a genetic colocalization analysis for two traits using summary-level data and the package 'coloc' in R (Giambartolomei et al., 2014; Wallace, 2020, 2021). In the support protocol, we provide code for installing the package 'coloc'. We also discuss critical parameters to consider when performing a colocalization analysis, and provide general guidelines for the interpretation of findings.

STRATEGIC PLANNING

Colocalization can be performed when the genetic variants (frequently SNPs) have summary data available in both studies within a specific genomic region. Colocalization can be implemented with any trait such that the regional data contain SNP-level beta regression coefficient and its variance (or only *p*-value) from the association of genotypes with the trait. It is assumed that each dataset has been pre-filtered using common quality control (QC) steps and post-imputation QC to minimize potential false findings (Turner et al., 2011). For example, one might exclude variants with a minor allele frequency (MAF) less than 0.01 and variants with imputation score less than 0.3. Further, the data from both studies must densely cover (with imputed data) the selected genomic region. Additionally, it is important to match SNP identifiers (e.g., rsIDs) across the datasets and to preferably have greater than 50 SNPs in a region. Also, it is important to include all high-quality SNPs in the region independently of *p*-value significance without pruning or clumping.

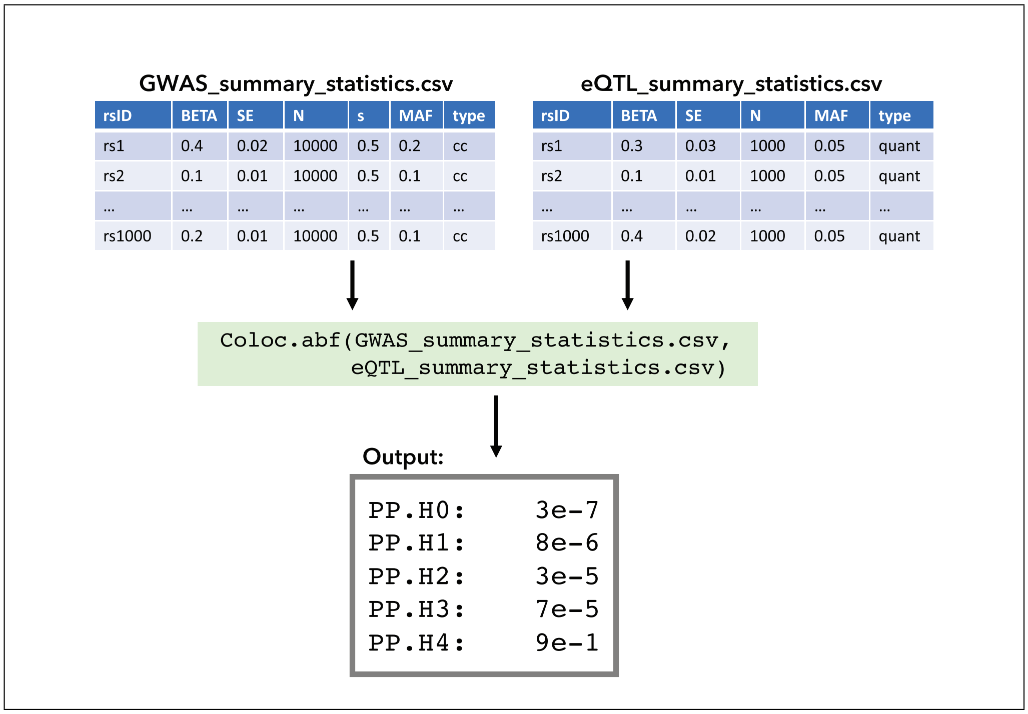
PERFORMING A GENETIC COLOCALIZATION ANALYSIS USING THE 'coloc' PACKAGE IN R AND SUMMARY-LEVEL DATA

BASIC PROTOCOL

In this protocol, we provide step-by-step guidelines and code for performing a genetic colocalization analysis of two traits using summary statistic data (see Fig. 1 for overview). The analysis shown demonstrates how to examine two phenotypes and test if they share a common genetic causal variant in a region of interest.

We thank the developer of the 'coloc' R package, Chris Wallace, for providing an easy-to-use resource and extensive documentation for conducting a genetic colocalization analysis (Giambartolomei et al., 2014; Wallace, 2020, 2021).

The full documentation for the 'coloc' R package is available at *https://cran.r-project.org/web/packages/coloc/index.html*.



GWAS_summary_statistics.csv

rsID	BETA	SE	N	s	MAF	type
rs1	0.4	0.02	10000	0.5	0.2	cc
rs2	0.1	0.01	10000	0.5	0.1	cc
…	…	…	…	…	…	…
rs1000	0.2	0.01	10000	0.5	0.1	cc

eQTL_summary_statistics.csv

rsID	BETA	SE	N	MAF	type
rs1	0.3	0.03	1000	0.05	quant
rs2	0.1	0.01	1000	0.05	quant
…	…	…	…	…	…
rs1000	0.4	0.02	1000	0.05	quant

```
Coloc.abf(GWAS_summary_statistics.csv,
          eQTL_summary_statistics.csv)
```

Output:

```
PP.H0:    3e-7
PP.H1:    8e-6
PP.H2:    3e-5
PP.H3:    7e-5
PP.H4:    9e-1
```

Figure 1 Overview of the coloc process. Summary data with association results from a GWAS and an eQTL study are analyzed, and the posterior probability for each causal variant configuration is calculated and displayed as output.

Necessary Resources

Hardware

A compute environment capable of running R/RStudio

Software

R version > 3.6, RStudio, ‘coloc’ R package, ‘remotes’ R package

Files

Genome-wide association study (GWAS) summary statistics and gene expression quantitative trait locus (eQTL) summary statistics for phenotypes of interest

Summary statistics data must be available for SNPs in the selected genomic region in both studies, and must have a dense coverage of the region of interest. More specifically, the summary statistics refer to the estimated effect sizes and their variance (standard error of the estimated effect sizes, squared) obtained from running associations between genetic variants in the region and the outcome using raw genotype data. The commonly used methods for association analysis are linear regression and logistic additive models when the outcome is a case-control association study. The data that are used as input by the ‘coloc’ package include the estimated effect size, its variance, the names of the genetic variants (e.g., rsID), and the type of outcome data (“quant” for quantitative data or “case-control”). The estimated effect size (beta) and its variance or standard error (SE) denote the regression coefficients between genotype and phenotype for each SNP. If beta and variance are unavailable in the summary statistics, they can be inferred from p-values, MAF, and sample size, and for case-control data, the fraction of samples that are cases.

1. Obtain summary statistics for the phenotypes of interest within a specified window. Additional information on the size of the window can be found in Critical Parameters. In this protocol, we will be conducting colocalization on two phenotypes: a disease GWAS and an eQTL study.

GWAS summary statistics can be obtained through publicly available resources, such as the NHGRI-EBI GWAS Catalog (Buniello et al., 2019).

2. Read in both summary statistic datasets into R (e.g., `read.csv` can read in a comma-separated value file; `read.table` can read in a tab-separated value file). For purposes of illustration, assume the summary statistics from the disease GWAS and eQTL study are in the files named `Disease_summarystats.csv` and `eQTL_summarystats.txt`, respectively.

```
gwas <- read.csv(file = "Disease_summarystats.csv", header = TRUE)
eqtl <- read.table(file = "eQTL_summarystats.txt", sep = "\t", header = TRUE)
```

3. Determine if the summary statistic data in both datasets have the following: effect size, variance, SNP position, type of data (quantitative or case-control/binary trait), and SNP identifiers (see next step). For binary traits, the effect size and the variance should be on the log odds scale (e.g., the effect size log odds ratio and its SE). If any of the variables are unavailable, the following steps may assist.

4. Determine an SNP identifier (SNPID). Any of the following would suffice:

```
chr:pos
rsIDs
chr:pos:allele1:allele2
```

Note that if using `chr:pos:allele1:allele2` as the SNPID, it may be challenging to match across datasets because alleles may be coded in different ways across different GWASs (e.g., indels can be coded for insertions/deletions (I/D) or with the alleles specified, depending on the GWAS).

Although harmonization is not required for running coloc, one could decide to use this step to harmonize allele IDs and to flip the betas for the same allele of reference in case downstream analysis may require it.

5. Remove any SNPs not measured in all constituent studies.

6. Ensure that the positions for both studies are on the same genomic build (e.g., both studies are in GRCh37). This only applies if the SNP positions are used as the SNP identifiers.

7. Check if the effect size and variance are available. If unavailable, provide *p*-values, MAF, and sample size for quantitative data. For case-control data, the fraction of samples that are cases must also be provided.

If the data have SE instead of variance, SE can be squared to obtain the variance.

```
data1$var <- (data1$SE)^2
```

8. For an association with a case-control trait (e.g., diseased vs. non-diseased), label the type as “cc.” For an association with a quantitative trait (e.g., blood pressure, weight, height), label as “quant.”

```
data1$type <- "cc"
```

or

```
data1$type <- "quant"
```

9. Calculate the fraction, s, of total samples (N_{total}) that are cases (N_{cases}):

$$s = \frac{N_{cases}}{N_{total}}$$

when `type ="quant"`, the input s is not needed (if provided, it will be ignored).

10. If MAF is unavailable for one of the studies, MAF of one study can be used for the other study because it is assumed that the LD of both traits is the same.

11. Check if any of the necessary data are missing or equal to zero. Identify any variables, including MAF, *p*-values (PVAL), beta, and SE, that may have missing data and any variables with data equal to zero:

```
names(which(colSums(is.na(gwas))>0)
subset(gwas, PVAL == 0)
```

If the datasets have missing data, data that are equal to zero, or nonsensical information (e.g., *p*-values or allele frequencies >1 or <0, SE = 0, or beta estimates and SE >10 or infinite), investigate the reason. Possible reasons include the numerical limits of R or an error in the data import. If the numerical limit is the issue for the *p*-value, and the user has beta and SE, it would be best to use these instead of the *p*-value.

If there are other issues due to wrong import of the data, these should be fixed before running coloc.

12. Construct the two datasets given the types of input data. For example, given sample size (N), *p*-values, and MAF:

```
data1 <- list(snp=gwas$rsID, N=gwas$N, pvalues=gwas$PVAL, type="quant", MAF =
   gwas$MAF)
```

Or, given beta and SE (most common case):

```
data2 <- list(snp=eqtl$rsID, beta = eqtl$BETA, varbeta = eqtl$SE^2,
   type="cc", N=eqtl$N, s = eqtl$s, MAF=eqtl$MAF))
```

Note that for `type="cc"`, the standard deviation of the trait (sdY) is internally computed using the information provided for s, the proportion of samples in the dataset that are cases. For `type="quant"`, it is encouraged to provide sdY in the 'coloc' command, if available. Note that, usually, quantitative traits have been standardized to have variance = 1 (so sdY can be set to 1). Otherwise, sdY will be computed internally by 'coloc' using variance, MAF, and N.

13. As a final check, make sure that both datasets have all the required information.

```
check_dataset(data1)
check_dataset(data2)
```

The output should be "NULL" if the data contain all the required information; otherwise an error will be output.

14. Ensure that the coverage in both datasets is dense.

```
plot_dataset(data1)
plot_dataset(data2)
```

15. Run the Bayes factor colocalization analysis. Calculate the posterior probabilities of different causal variant configurations.

```
res <- coloc.abf(dataset1 = data1, dataset2 = data2)
print(res)
```

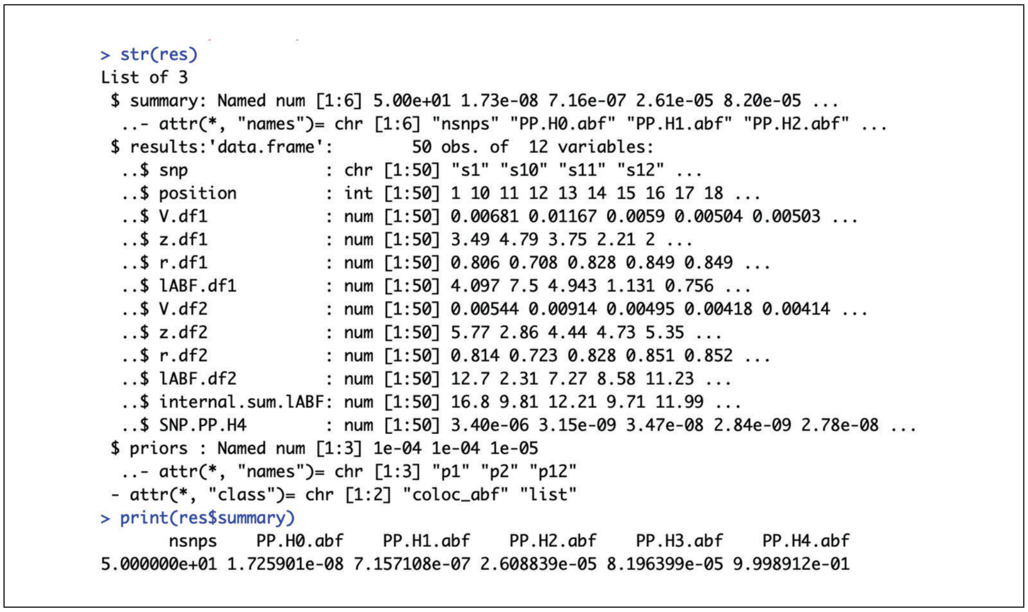
16. The five posterior probabilities (see Fig. 2) can be extracted using:

```
print(res$summary)
```

PP.H4 is the probability of a shared causal variant across the two traits.

17. *Optional:* To extract the most likely causal variants from the region and to obtain SNP-based ranking, including posterior probabilities for each SNP, subset for the posterior probabilities for each SNP found to be causal conditional on H4 being true:

```
subset(res$results, SNP.PP.H4>0.01)
```



```
> str(res)
List of 3
 $ summary: Named num [1:6] 5.00e+01 1.73e-08 7.16e-07 2.61e-05 8.20e-05 ...
  ..- attr(*, "names")= chr [1:6] "nsnps" "PP.H0.abf" "PP.H1.abf" "PP.H2.abf" ...
 $ results:'data.frame':        50 obs. of  12 variables:
  ..$ snp              : chr [1:50] "s1" "s10" "s11" "s12" ...
  ..$ position         : int [1:50] 1 10 11 12 13 14 15 16 17 18 ...
  ..$ V.df1            : num [1:50] 0.00681 0.01167 0.0059 0.00504 0.00503 ...
  ..$ z.df1            : num [1:50] 3.49 4.79 3.75 2.21 2 ...
  ..$ r.df1            : num [1:50] 0.806 0.708 0.828 0.849 0.849 ...
  ..$ lABF.df1         : num [1:50] 4.097 7.5 4.943 1.131 0.756 ...
  ..$ V.df2            : num [1:50] 0.00544 0.00914 0.00495 0.00418 0.00414 ...
  ..$ z.df2            : num [1:50] 5.77 2.86 4.44 4.73 5.35 ...
  ..$ r.df2            : num [1:50] 0.814 0.723 0.828 0.851 0.852 ...
  ..$ lABF.df2         : num [1:50] 12.7 2.31 7.27 8.58 11.23 ...
  ..$ internal.sum.lABF: num [1:50] 16.8 9.81 12.21 9.71 11.99 ...
  ..$ SNP.PP.H4        : num [1:50] 3.40e-06 3.15e-09 3.47e-08 2.84e-09 2.78e-08 ...
 $ priors : Named num [1:3] 1e-04 1e-04 1e-05
  ..- attr(*, "names")= chr [1:3] "p1" "p2" "p12"
 - attr(*, "class")= chr [1:2] "coloc_abf" "list"
> print(res$summary)
       nsnps    PP.H0.abf    PP.H1.abf    PP.H2.abf    PP.H3.abf    PP.H4.abf
5.000000e+01 1.725901e-08 7.157108e-07 2.608839e-05 8.196399e-05 9.998912e-01
```

Figure 2 The outputs of coloc are presented. The first command displays the structure of the R object 'res' in a list with three elements: (1) "summary" returns the number of SNPs analyzed and the posterior probabilities, (2) "results" returns the log Approximate Bayes Factors, intermediate calculations, and per-SNP posterior probability (SNP.PP.H4) of the SNP being causal for the shared signal if there is strong evidence for H4, and (3) "priors" returns the priors used for p1, p2, and p12 (see Critical Parameters for more details). The second command displays a shortened summary of the output, printing the first element of the list called "summary," which is typically the main output of interest.

INSTALLING THE 'coloc' R PACKAGE

SUPPORT PROTOCOL

This Support Protocol describes how to install the necessary packages and dependencies for performing a genetic colocalization analysis using the 'coloc' R package.

Necessary Resources

Hardware

A compute environment capable of running R/RStudio

Software

R version > 3.6, RStudio, 'coloc' R package, 'remotes' R package

Files

None

1. Install the 'remotes' R package.

```
if(!require("remotes"))
install.packages("remotes")
```

2. Load the 'remotes' R package.

```
library(remotes)
```

3. Install the 'coloc' R package.

```
install_github("chr1swallace/coloc",build_vignettes=TRUE)
```

Functions within the 'coloc' R package are described in the R documentation. Please refer to the 'coloc' R documentation for a complete list of the parameters, options, and additional details regarding each function within the program. Further details of particular functions can be accessed through the help page, which can be accessed by prepending a question mark to the function name within the R interface.

4. Load the ‘coloc’ R package.

```
library(coloc)
```

GUIDELINES FOR UNDERSTANDING RESULTS

The ‘coloc’ package outputs posterior probabilities for each hypothesis. The posterior probability for hypothesis H4 (PP.H4) is the probability of a shared causal variant across the two traits. A high PP.H4 suggests that a single variant affects both traits 1 and 2. On the other hand, a high posterior probability for hypothesis H3 (PP.H3) suggests that two independent causal SNPs are associated with traits 1 and 2. While studies use different thresholds to evaluate evidence for a shared causal variant, suggestive evidence for colocalization can be defined by a threshold of $PP.H4 \geq 0.5$, and strong evidence can be defined by a more stringent threshold of $PP.H4 \geq 0.8$.

The posterior probabilities should be interpreted with caution (Giambartolomei et al., 2014). A low PP.H3 and a low PP.H4 may not necessarily indicate evidence against colocalization—they may arise simply due to limited power in one or both of the studies. This can be identified by high values of PP.H0 (indicative of no signal in either of the two datasets), PP.H1 (suggesting a signal only in the first dataset but not in the second dataset), and/or PP.H2 (suggesting a signal only in the second but not in the first dataset) (Giambartolomei et al., 2014). An accumulation of signal in PP.H3 and/or PP.H4 suggests clear evidence, and therefore enough power, for colocalization to occur in both datasets; however, evidence can point to either the same or a different causal variant between the two datasets.

While a high PP.H4 is indicative of strong evidence for the same variant affecting both traits, this does not imply causality, i.e., it cannot be concluded from this analysis that the two traits are causally related. Further, this method relies on the assumption that there is at most a single causal variant per locus. In a scenario of loci with multiple causal variants, this assumption may influence the results and bias against colocalization.

The probabilities derived from colocalization can also be combined and adapted to the information needed. One approach is to perform a genome-wide colocalization analysis by splitting the genome into regions, which reflects a “hypothesis-free” search for colocalization. In this scenario, the probabilities of PP.H3 and/or PP.H4 can guide the regions being tested across the genome with sufficient power to identify a signal in both datasets. In this context, a threshold of PP.H3 + PP.H4 > 0.99 has been suggested as evidence of overlapping signals, and PP.H4/PP.H3 > 5 has been suggested as convincing evidence of colocalization (Guo et al., 2015). A high PP.H3 can indicate either the presence of different causal variants across the two traits or the presence of more than one causal variant in the region. Recent approaches have led to the development of the Sum of Single Effects (SuSiE) regression framework, which can be used with ‘coloc’ for fine-mapping genetic signals by evaluating evidence for association at multiple causal variants simultaneously (Wallace, 2021).

COMMENTARY

Background Information

Genome-wide association studies (GWAS) have mapped the polygenic architecture of thousands of human traits and diseases, uncovering the genetic basis of thousands of complex common or multifactorial diseases. Despite the large number of trait-SNP associations discovered from GWAS, it is challenging to translate findings into the clinic until we have an understanding of how the identified SNPs contribute to disease (Cookson, Liang, Abecasis, Moffatt, & Lathrop, 2009). eQTL studies associate genetic variants with gene expression levels measured in thousands of individuals. Integrating GWAS with eQTL data on the association of variants with expression levels of nearby genes can help discriminate the gene that influences the GWAS signal, since eQTLs test for association with a specific gene expression. In general, integrating GWAS with molecular QTLs can help delineate the genetic elements that mediate disease susceptibility (Hawkins, Hon, & Ren, 2010), strengthening our understanding of the genetic basis for the trait (Nicolae et al., 2010).

One of the critical factors that can affect the interpretation of GWAS findings is linkage disequilibrium (LD)—the correlation between nearby genomic variants that can result in neighboring alleles occurring together more often than by random chance (Slatkin, 2008). LD can confound the causal association between a genetic variant and a disease; an observed association may arise from a true causal effect of the variant or from the variant in LD with the causal variant. A challenge in the interpretation of GWAS findings is that over 90% of risk variants revealed by GWAS fall into non-coding regions, including intronic, promoter, and enhancer regions, micro-RNAs, and long non-coding RNAs, or may overlap with open-chromatin regions regulating transcriptional activity (Edwards et al., 2013; Hrdlickova et al., 2014). Prior studies have found that GWAS signals may be more likely to be eQTL compared to SNPs chosen at random (Dubois et al., 2010; Nicolae et al., 2010). However, because of LD, it is challenging to discriminate whether the signal identified in eQTL studies points to the same causal variant as that identified in a GWAS. Determining if a single SNP is associated with two traits of interest is not sufficient to indicate colocalization because the associations could be driven by two distinct causal SNPs in LD. Further, visual methods for evaluating colocalization, such as subjectively examining for overlaps of association signals between the two datasets, are difficult because of the LD pattern and the abundance of eQTLs in the genome, which make it likely that some overlaps are coincidental, rather than driven by the same functional variants (Dixon et al., 2007). An early approach for distinguishing colocalization from shared causal variants relies on a method for scoring the SNPs that influence the two traits using the residuals of the association discovered with standard linear regression, conditioned on the most associated SNP (Nica et al., 2010). While the scoring scheme corrects for the local correlation structure due to LD, it does not provide a formal test of a null hypothesis for colocalization (Giambartolomei et al., 2014). Discriminating the causal variant from other variants that are in LD with such a causal variant has led to the development of statistical fine-mapping approaches for prioritizing variants while accounting for LD structure (LaPierre et al., 2021; Schaid, Chen, & Larson, 2018). However, these methods require information on LD, preferably computed from in-sample, which is often not easily accessible.

Another method for colocalization entails examining whether the coefficients from regressions of each trait against two or more SNPs are proportional, which should be true if the traits share causal variants (Plagnol, Smyth, Todd, & Clayton, 2009; Wallace et al., 2012). However, this method requires individual-level genotype data, which are much more challenging to access compared to summary statistics and can be more difficult to process due to the computational burden. With increasingly available GWAS summary statistics, methods that rely on summary statistics are more favorable. The colocalization methodology presented in this protocol addresses these limitations by leveraging summary statistics data and utilizing a Bayesian framework that tests multiple hypotheses for distinguishing no causal variants for either trait or distinct causal variants for each of the traits (or both) (Giambartolomei et al., 2014). Colocalization discoveries have provided insight into genes and molecular mechanisms underlying complex diseases such as atherosclerosis (Franceschini et al., 2018) and cancer (Giambartolomei et al., 2021).

Colocalization is increasingly used with Mendelian randomization (MR) studies for evaluating causal relationships for complex traits and disease (Zuber et al., 2022). MR is a method that assesses the causal effect of an exposure on an outcome using measured variation in genes (Smith & Ebrahim, 2003). While colocalization determines whether two potentially related traits are affected by the same causal variant, evidence from an MR analysis suggests that an exposure has a causal effect on the outcome. These approaches can be complementary. For example, colocalization can be performed as a follow-up to a *cis*-MR investigation for evaluating the validity of the MR assumptions. A review and comparison of these two approaches can be found in Zuber et al. (2022). Additionally, step-by-step guidelines for performing a univariate two-sample MR analysis using the R package 'TwoSampleMR' and GWAS summary statistics can be found in Rasooly & Patel (2019); the corresponding guidelines for a multivariable two-sample MR analysis using the 'MVMR' R package can be found in Rasooly & Peloso (2021).

Critical Parameters

There are a number of critical parameters that must be considered in a colocalization study.

First, similar to a GWAS where the statistical power to detect causal variants increases with larger sample size, a colocalization study also benefits from larger sample sizes due to having more power to identify evidence for (or against) colocalization. As shown in prior simulation analyses using eQTL and biomarker datasets with varied parameters for sample size and proportion of variance explained by the shared genetic variant, the causal variant needs to explain nearly 2% of the variance of the biomarker in a sample size of 2000 individuals to achieve reliable evidence in support of colocalization (Giambartolomei et al., 2014). However, with larger sample sizes, smaller effects can show evidence in support of colocalization.

Second, the density of the region (i.e., more genetic variants matching across datasets) is also important for generating evidence for (or against) colocalization.

Third, a frequently asked question is how big of a region should be used for colocalization? Because this method relies on the assumption that the region has at most one causal variant per trait, the region should include all the variants in LD with the lead SNP of interest. The region cannot be too large (for example, a region greater than 3 Mb will likely contain more than one signal). One could take approximately independent LD blocks of 1 megabase (Mb) (Loh et al., 2015) or 5000 SNPs (Pickrell, 2014). Usually, the region contains between 200 kb to 3 Mb and between 1000 and 10,000 SNPs.

Fourth, the selection of the region (the base pair-defined window), and the associations used, depends on the goal of the study. Some examples of region definitions are:

- the region is defined by the TSS of a gene; for example, as a scan around the TSS of each gene in the genome, if the goal is to identify genes that have the most evidence in *cis* of sharing a signal with the GWAS;
- a region on either side of the lead GWAS SNP, using the *cis*-eQTLs of genes nearby, if the aim is to understand molecular traits sharing a signal in this region (similar to a fine-mapping goal);
- a region on either side of the lead GWAS SNP, and using the *trans*-eQTLs of genes located anywhere in the genome: in this case, the suggestion is to test genes only if there is previous evidence of shared genetic mechanism; otherwise the prior would need to be very high (Giambartolomei et al., 2014).

Fifth, as a Bayesian method, colocalization requires specification of three prior probabilities:

- p_1, the prior probability that a SNP in the region is causally associated to trait 1,
- p_2, the prior probability that a SNP in the region is causally associated to trait 2, and
- p_{12}, the prior probability that a SNP in the region is causally associated to both traits.

The default for p_1 and p_2 is 1×10^{-4} and the default for p_{12} is 1×10^{-5}. The default prior probabilities can be adjusted based on the available data and judgment for evaluating the prior beliefs. For a more stringent threshold, for example, p_{12} can be set to1×10^{-6}. A detailed guide on the selection of prior probabilities and relevant sensitivity analyses can be found in Wallace (2020).

Acknowledgments

The authors are very grateful to Chris Wallace for developing the 'coloc' R package and for providing an easy-to-use resource and extensive documentation for implementing a genetic colocalization analysis. The 'coloc' package is available for download at *https://cran.r-project.org/web/packages/coloc/index.html* (Giambartolomei et al., 2014; Wallace, 2020, 2021).

D.R. receives funding from the Department of Veterans Affairs Award No. 36C24E18D0052. C.G. receives funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 754490-MINDED project.

Author Contributions

Danielle Rasooly: methodology, writing original draft, writing review and editing; **Claudia Giambartolomei**: methodology, writing original draft, writing review and editing; **Gina Peloso**: methodology, writing original draft, writing review and editing.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Literature Cited

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., …Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(D1), D1005–D1012. doi: 10.1093/nar/gky1120

Cano-Gamez, E., & Trynka, G. (2020). From GWAS to function: Using functional genomics to identify the mechanisms underlying complex diseases. *Frontiers in Genetics*, *11*, 424. doi: 10.3389/fgene.2020.00424

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., & Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nature Reviews Genetics*, *10*(3), 184–194. doi: 10.1038/nrg2537

Dixon, A. L., Liang, L., Moffatt, M. F., Chen, W., Heath, S., Wong, K. C. C., …Cookson, W. O. C. (2007). A genome-wide association study of global gene expression. *Nature Genetics*, *39*(10), 1202–1207. doi: 10.1038/ng2109

Dubois, P. C. A., Trynka, G., Franke, L., Hunt, K. A., Romanos, J., Curtotti, A., …van Heel, D. A. (2010). Multiple common variants for celiac disease influencing immune gene expression. *Nature Genetics*, *42*(4), 295–302. doi: 10.1038/ng.543

Edwards, S. L., Beesley, J., French, J. D., & Dunning, A. M. (2013). Beyond GWASs: Illuminating the dark road from association to function. *American Journal of Human Genetics*, *93*(5), 779–797. doi: 10.1016/j.ajhg.2013.10.012

Franceschini, N., Giambartolomei, C., de Vries, P. S., Finan, C., Bis, J. C., Huntley, R. P., …O'Donnell, C. J. (2018). GWAS and colocalization analyses implicate carotid intima-media thickness and carotid plaque loci in cardiovascular outcomes. *Nature Communications*, *9*(1), 5141. doi: 10.1038/s41467-018-07340-5

Giambartolomei, C., Seo, J.-H., Schwarz, T., Freund, M. K., Johnson, R. D., Spisak, S., …Freedman, M. L. (2021). H3K27ac HiChIP in prostate cell lines identifies risk genes for prostate cancer susceptibility. *American Journal of Human Genetics*, *108*(12), 2284–2300. doi: 10.1016/j.ajhg.2021.11.007

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., & Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics*, *10*(5), e1004383. doi: 10.1371/journal.pgen.1004383

Goldstein, D. B. (2009). Common genetic variation and human traits. *New England Journal of Medicine*, *360*(17), 1696–1698. doi: 10.1056/NEJMp0806284

GTEx Consortium. (2017). Genetic effects on gene expression across human tissues. *Nature*, *550*(7675), 204–213. doi: 10.1038/nature24277

Guo, H., Fortune, M. D., Burren, O. S., Schofield, E., Todd, J. A., & Wallace, C. (2015). Integration of disease association and eQTL data using a Bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases. *Human Molecular Genetics*, *24*(12), 3305–3313. doi: 10.1093/hmg/ddv077

Hawkins, R. D., Hon, G. C., & Ren, B. (2010). Next-generation genomics: An integrative approach. *Nature Reviews Genetics*, *11*(7), 476–486. doi: 10.1038/nrg2795

Hrdlickova, B., de Almeida, R. C., Borek, Z., & Withoff, S. (2014). Genetic variation in the noncoding genome: Involvement of micro-RNAs and long non-coding RNAs in disease. *Biochimica et Biophysica Acta*, *1842*(10), 1910–1922. doi: 10.1016/j.bbadis.2014.03.011

LaPierre, N., Taraszka, K., Huang, H., He, R., Hormozdiari, F., & Eskin, E. (2021). Identifying causal variants by fine mapping across multiple studies. *PLoS Genetics*, *17*(9), e1009733. doi: 10.1371/journal.pgen.1009733

Loh, P.-R., Bhatia, G., Gusev, A., Finucane, H. K., Bulik-Sullivan, B. K., Pollack, S. J., …Price, A. L. (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nature Genetics*, *47*(12), 1385–1392. doi: 10.1038/ng.3431

Manolio, T. A. (2010). Genomewide association studies and assessment of the risk of disease. *The New England Journal of Medicine*, *363*(2), 166–176. doi: 10.1056/NEJMra0905980

Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S., & Cheung, V. G. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature*, *430*(7001), 743–747. doi: 10.1038/nature02797

Nica, A. C., & Dermitzakis, E. T. (2013). Expression quantitative trait loci: Present and future. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences*, *368*(1620), 20120362. doi: 10.1098/rstb.2012.0362

Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I., & Dermitzakis, E. T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genetics*, *6*(4), e1000895. doi: 10.1371/journal.pgen.1000895

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., & Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS. *PLoS Genetics*, *6*(4), e1000888. doi: 10.1371/journal.pgen.1000888

Pickrell, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *American Journal of Human Genetics*, *94*(4), 559–573. doi: 10.1016/j.ajhg.2014.03.004

Plagnol, V., Smyth, D. J., Todd, J. A., & Clayton, D. G. (2009). Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13. *Biostatistics*, *10*(2), 327–334. doi: 10.1093/biostatistics/kxn039

Rasooly, D., & Patel, C. J. (2019). Conducting a reproducible mendelian randomization analysis using the R analytic statistical environment. *Current Protocols in Human Genetics*, *101*(1), e82. doi: 10.1002/cphg.82

Rasooly, D., & Peloso, G. M. (2021). Two-sample multivariable mendelian randomization analysis using R. *Current Protocols*, *1*(12), e335. doi: 10.1002/cpz1.335

Schadt, E. E., Monks, S. A., Drake, T. A., Lusis, A. J., Che, N., Colinayo, V., …Friend, S. H. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature*, *422*(6929), 297–302. doi: 10.1038/nature01434

Schaid, D. J., Chen, W., & Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews. Genetics*, *19*(8), 491–504. doi: 10.1038/s41576-018-0016-z

Slatkin, M. (2008). Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Reviews. Genetics*, *9*(6), 477–485. doi: 10.1038/nrg2361

Smith, G. D., & Ebrahim, S. (2003). “Mendelian randomization”: Can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, *32*(1), 1–22. doi: 10.1093/ije/dyg070

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews. Genetics*, *20*(8), 467–484. doi: 10.1038/s41576-019-0127-1

Turner, S., Armstrong, L. L., Bradford, Y., Carlson, C. S., Crawford, D. C., Crenshaw, A. T., …Ritchie, M. D. (2011). Quality control procedures for genome-wide association studies. *Current Protocols in Human Genetics*, *68*, 1.19.1–1.19.18. doi: 10.1002/0471142905.hg0119s68

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., …Lappalainen, T. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), 1–21. doi: 10.1038/s43586-021-00056-9

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five Years of GWAS Discovery. *American Journal of Human Genetics*, *90*(1), 7. doi: 10.1016/j.ajhg.2011.11.029

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 years of GWAS discovery: Biology, function, and translation. *American Journal of Human Genetics*, *101*(1), 5–22. doi: 10.1016/j.ajhg.2017.06.005

Wallace, C. (2020). Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLoS Genetics*, *16*(4), e1008720. doi: 10.1371/journal.pgen.1008720

Wallace, C. (2021). A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genetics*, *17*(9), e1009440. doi: 10.1371/journal.pgen.1009440

Wallace, C., Rotival, M., Cooper, J. D., Rice, C. M., Yang, J. H. M., McNeill, M., …Blankenberg, S. (2012). Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. *Human Molecular Genetics*, *21*(12), 2815–2824. doi: 10.1093/hmg/dds098

Zuber, V., Grinberg, N. F., Gill, D., Manipur, I., Slob, E. A. W., Patel, A., …Burgess, S. (2022). Combining evidence from Mendelian randomization and colocalization: Review and comparison of approaches. *American Journal of Human Genetics*, *109*(5), 767–782. doi: 10.1016/j.ajhg.2022.04.001

Internet Resources

https://chr1swallace.github.io/coloc/index.html

Github link to the ‘coloc’ package developed by Chris Wallace and related documentation.

https://www.ebi.ac.uk/gwas/

The NHGRI-EBI GWAS catalog provides a resource for viewing, querying, and downloading summary statistics spanning hundreds of thousands of associations from thousands of publications (Buniello et al., 2019).