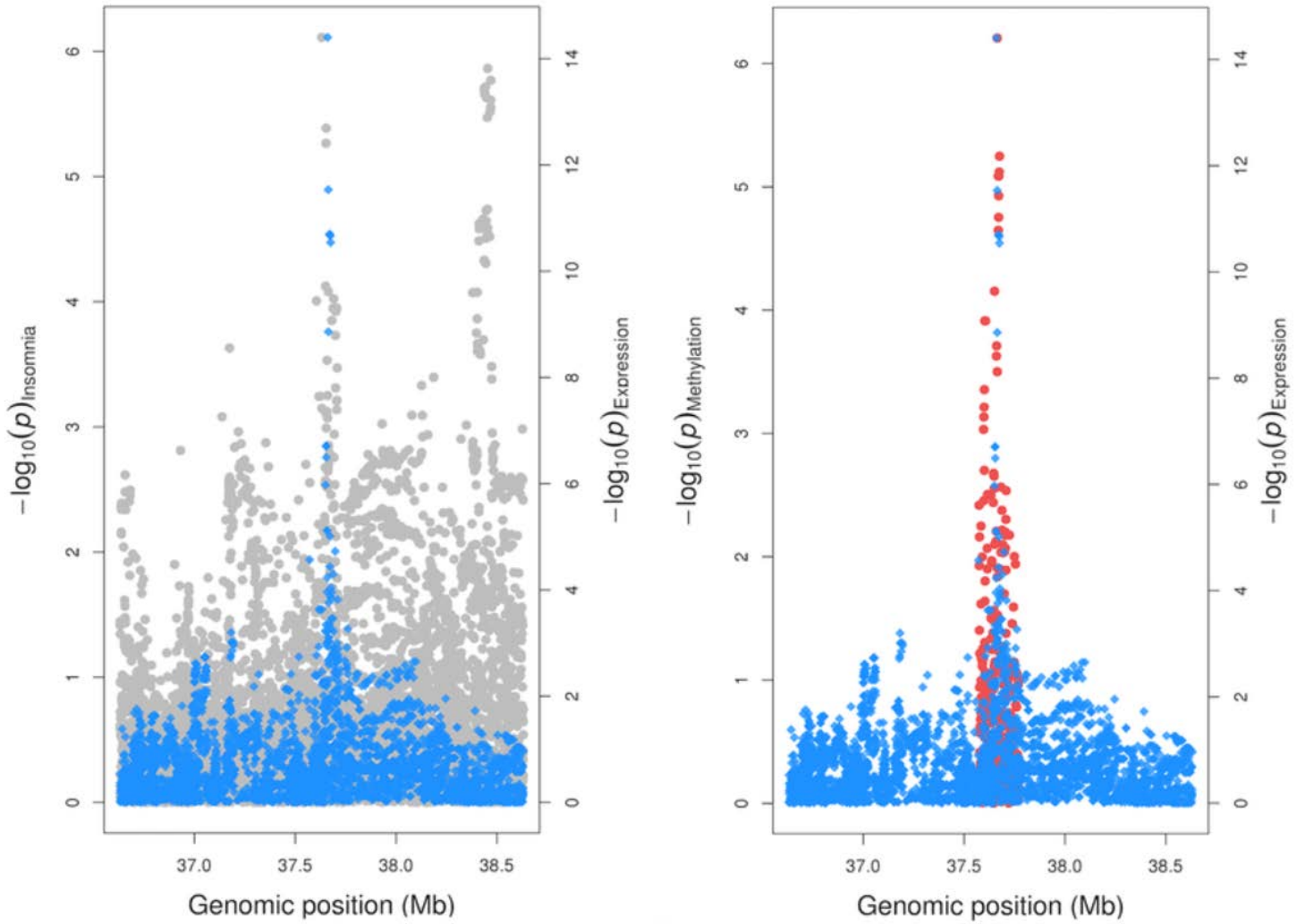


https://www.nature.com/articles/s41398-019-0437-2

COLOCALIZATION

Misa Graff

June 6, 2024

What is colocalization?

Genetic colocalization is a statistical procedure that determines if two or more traits share the same genetic signals at a specific locus in genome-wide association studies (GWAS).

Colocalization analysis tries to differentiate between two scenarios:

(1) Distinct causal variants - The causal variant for trait A is different from the causal variant for trait B, but they are both at the same locus.

(2) Shared signal(s) - The causal variant(s) for trait A and trait B are shared

Why perform colocalization?

Genetic association studies have found evidence that human disease risk or other traits are under the influence of genetic variants.

Many of these studies are publicly available.

Focusing on whether different traits are under influence of the same variants can help us understand:

- how variants lead to differences in disease risk,
- etiology of disease, and/or
- molecular basis of disease.

Here are some example questions that can be addressed through genetic colocalization analysis:

1. Do two traits that are genetically correlated share common genetic determinants?
2. Can we identify shared genetic architecture between a disease phenotype and a related quantitative trait?
3. Are two diseases with overlapping symptoms or pathways genetically linked through shared risk variants?
4. Is a particular gene implicated in the etiology of two seemingly unrelated phenotypes?
5. Can we pinpoint specific genetic variants driving the association between a protein and a complex trait?
6. Are there pleiotropic effects of a genetic variant on multiple phenotypes, and if so, what are the affected traits?
7. Are there shared genetic factors contributing to both the severity and age of onset of a disease?
8. Can we elucidate the genetic basis of co-occurring traits or comorbid conditions?
9. Are there genetic factors influencing both the response to a drug and susceptibility to adverse effects?

Examples of tools used for colocalization

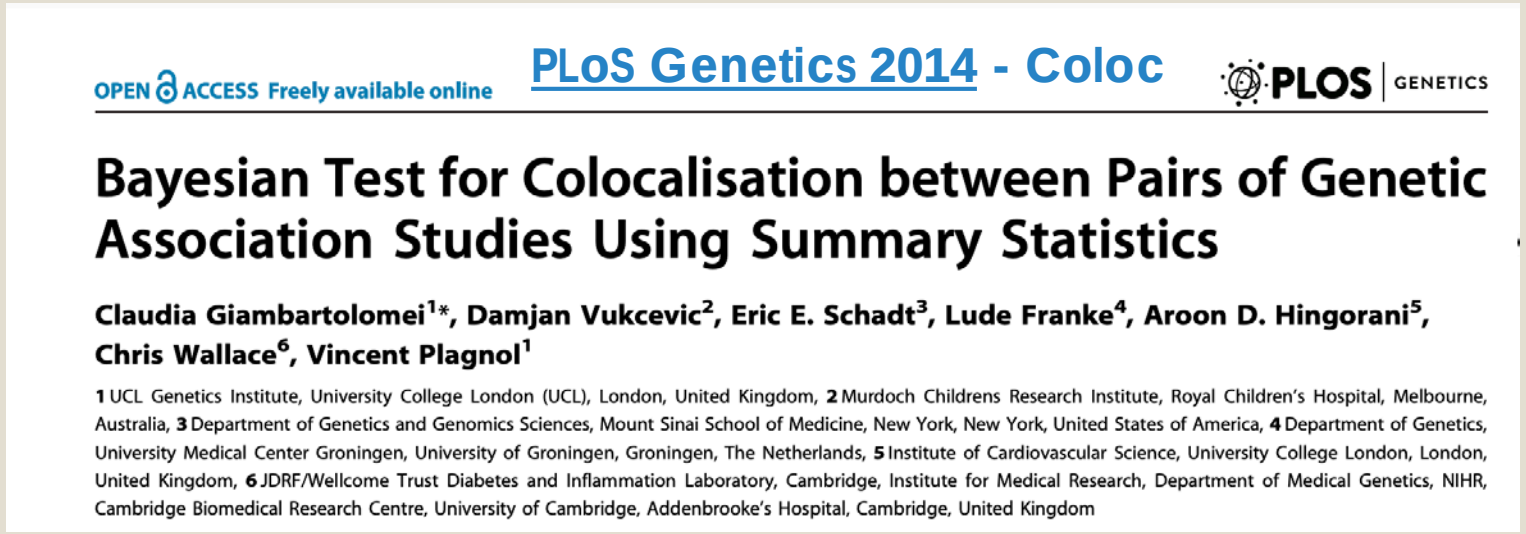
•**Coloc**: Coloc is an R package that implements Bayesian methods for colocalization analysis. It allows users to assess evidence for colocalization of genetic signals between two traits or phenotypes based on summary statistics from genome-wide association studies (GWAS).

•**Coloc-SuSiE:** Coloc-SuSiE aims to improve the accuracy and resolution of identifying shared causal variants between traits while accounting for the complex correlation structure of genetic variants and traits.

•**eCAVIAR**: Enhanced CAVIAR (CAusal Variants Identification in Associated Regions) is a tool for colocalization analysis that integrates GWAS summary statistics with functional genomic annotations to prioritize potentially causal variants within associated regions.

•**HEIDI**: Heterogeneity in Dependent Instruments (HEIDI) is a method for colocalization analysis that distinguishes between shared causal variants and pleiotropy.

•**GCTA-COJO**: Genome-wide Complex Trait Analysis (GCTA) Conditional and Joint Analysis (COJO) is a tool for conditional and joint analysis of GWAS summary statistics. While primarily designed for conditional analysis, it can also be used for colocalization analysis to identify shared causal variants between traits.



OPEN ACCESS Freely available online

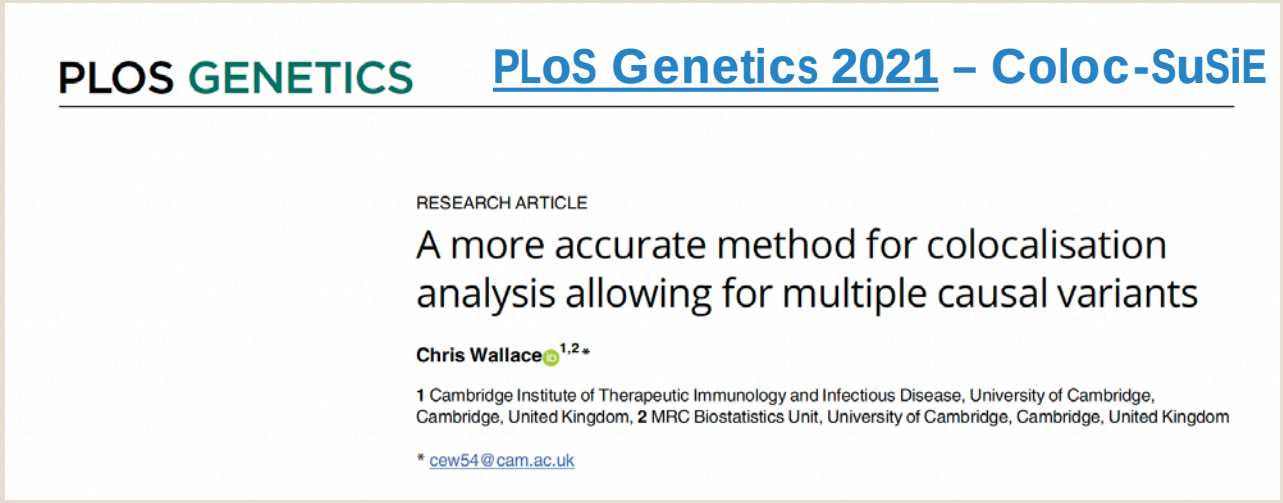
PLoS Genetics 2014 - Coloc

PLOS | GENETICS

Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics

Claudia Giambartolomei[1]*, Damjan Vukcevic[2], Eric E. Schadt[3], Lude Franke[4], Aroon D. Hingorani[5], Chris Wallace[6], Vincent Plagnol[1]

1 UCL Genetics Institute, University College London (UCL), London, United Kingdom, **2** Murdoch Childrens Research Institute, Royal Children's Hospital, Melbourne, Australia, **3** Department of Genetics and Genomics Sciences, Mount Sinai School of Medicine, New York, New York, United States of America, **4** Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands, **5** Institute of Cardiovascular Science, University College London, London, United Kingdom, **6** JDRF/Wellcome Trust Diabetes and Inflammation Laboratory, Cambridge, Institute for Medical Research, Department of Medical Genetics, NIHR, Cambridge Biomedical Research Centre, University of Cambridge, Addenbrooke's Hospital, Cambridge, United Kingdom



PLOS GENETICS

PLoS Genetics 2021 – Coloc-SuSiE

RESEARCH ARTICLE

A more accurate method for colocalisation analysis allowing for multiple causal variants

Chris Wallace[1,2]*

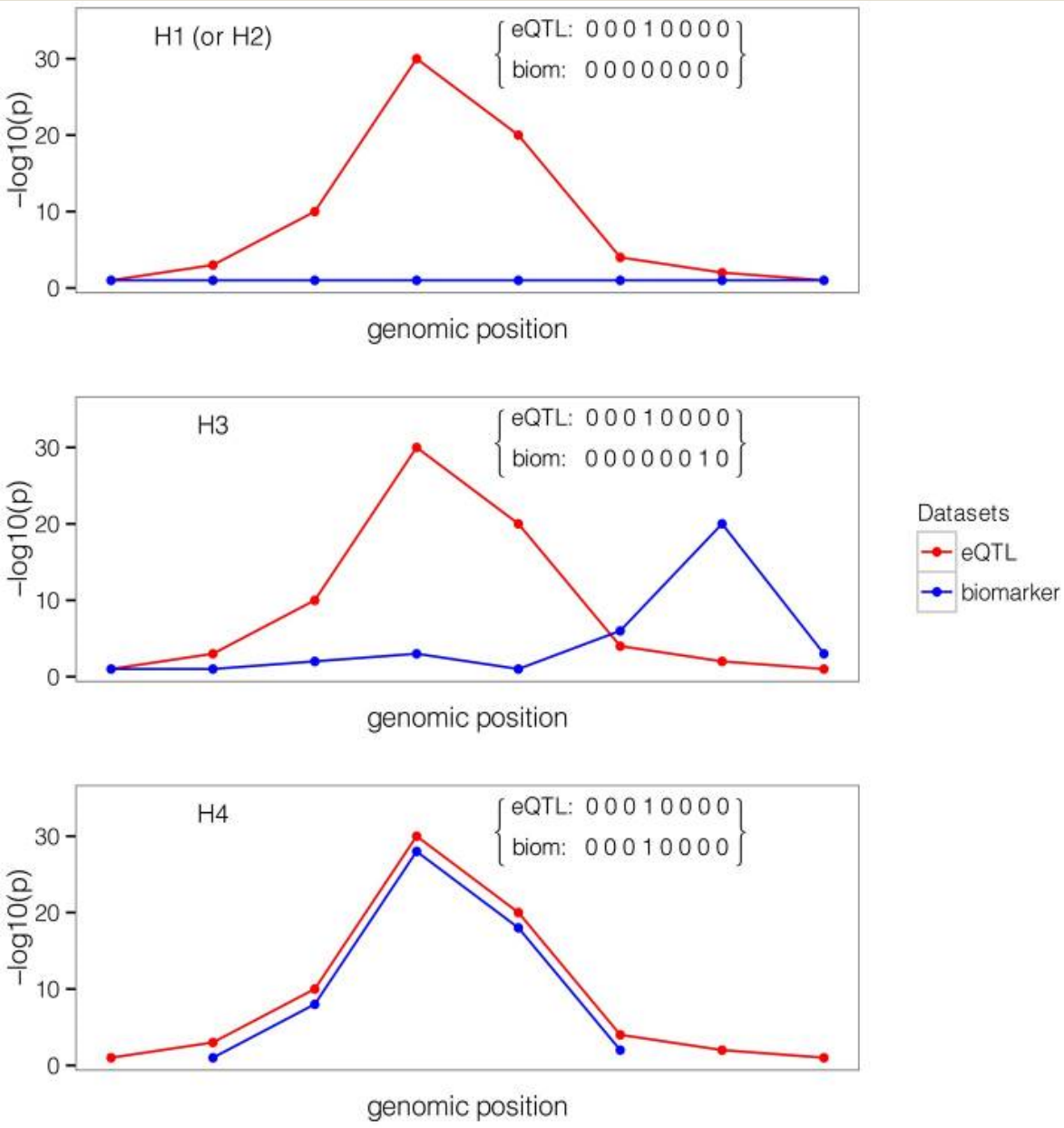
1 Cambridge Institute of Therapeutic Immunology and Infectious Disease, University of Cambridge, Cambridge, United Kingdom, **2** MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom

* cew54@cam.ac.uk

Coloc and Coloc-SuSiE Package

The coloc package can be used to perform genetic colocalisation analysis of two potentially related phenotypes, to ask whether they share common genetic causal variant(s) in a given region.

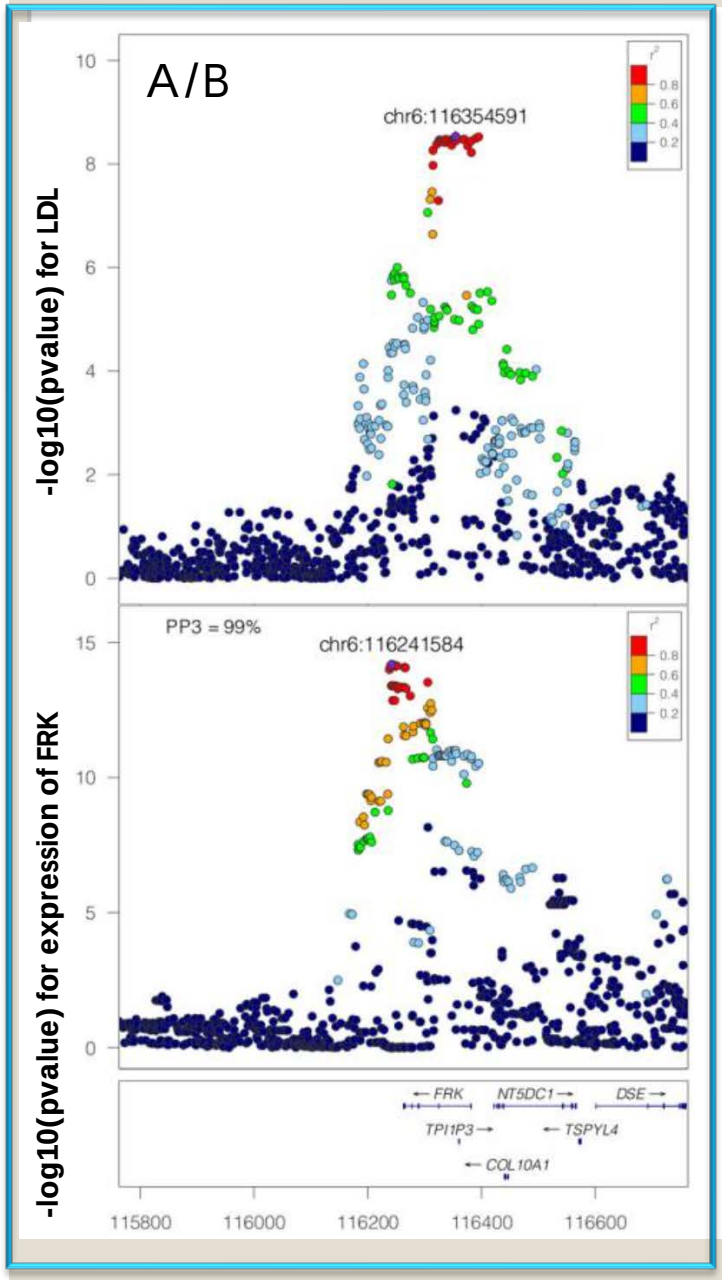
- https://cran.r-project.org/web/packages/coloc/vignettes/a02_data.html

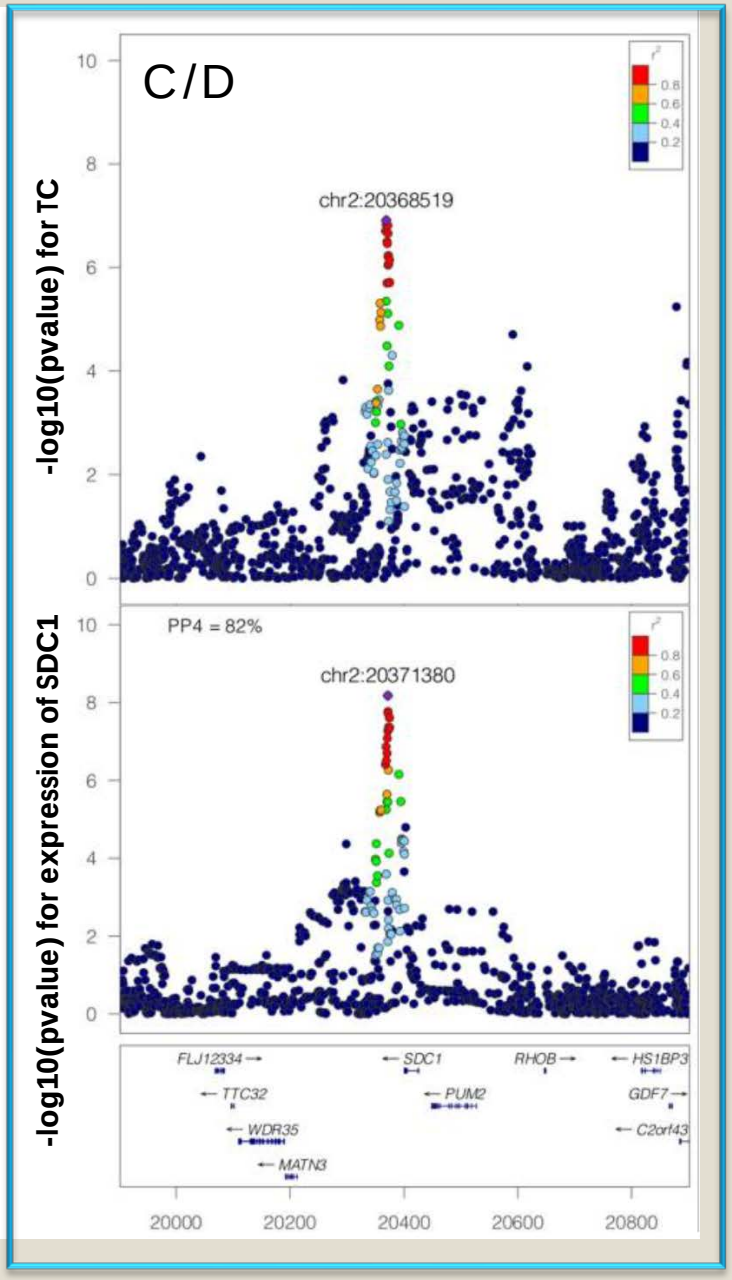


PLoS Genetics 2014 - Coloc

Example of one configuration under different hypotheses. A configuration is represented by one binary vector for each trait of (0,1) values of length n=8, the number of shared variants in a region. The value of 1 means that the variant is causally involved in disease, 0 that it is not. The first plot shows the case where only one dataset shows an association. The second plot shows that the causal SNP is different for the biomarker dataset (**biom**) compared to the expression dataset (**eQTL**). The third plot shows the configuration where the single causal variant is the fourth one.

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





PLoS Genetics 2014 - Coloc

A/B show

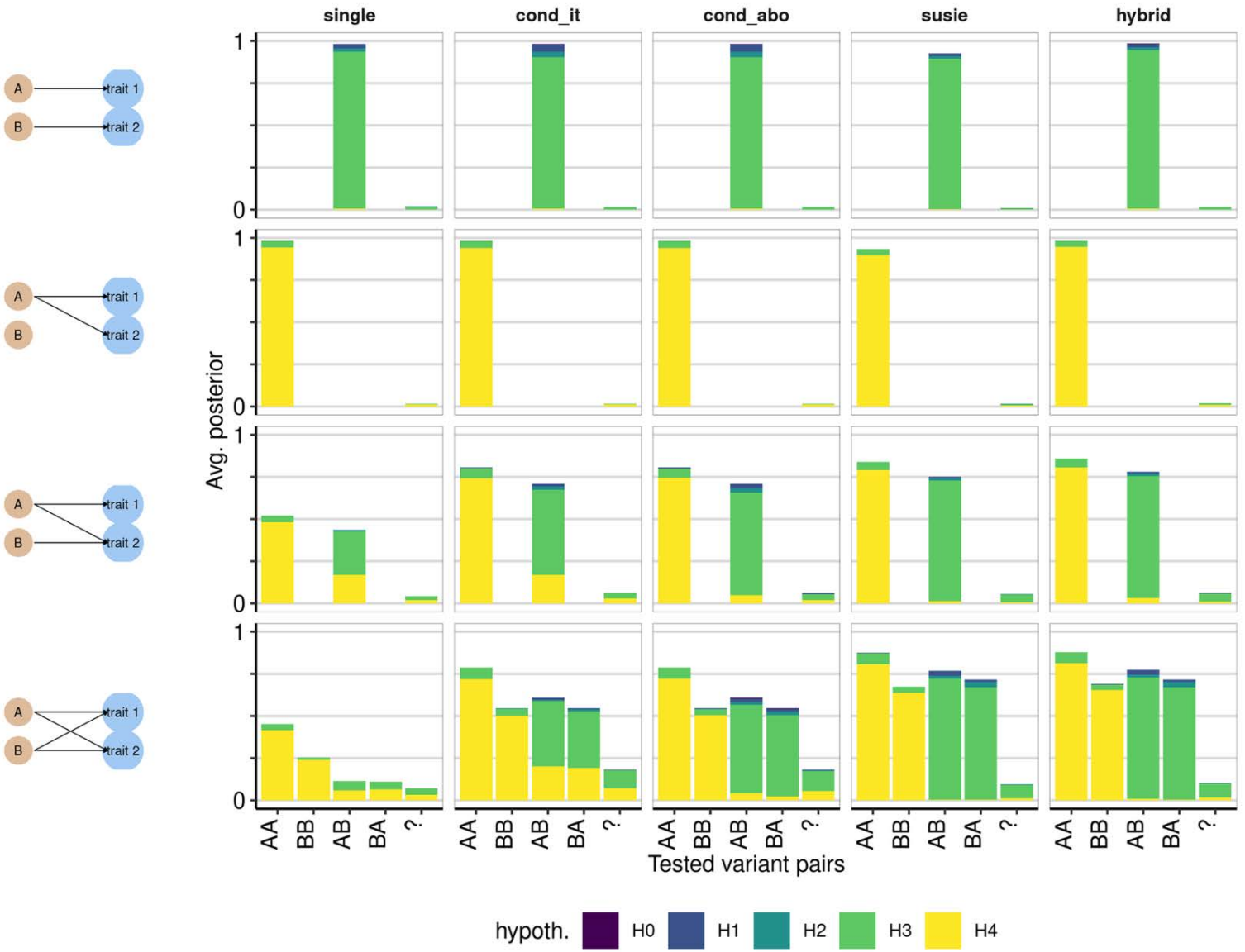
- H_3 : Association with trait 1 and trait 2, two independent SNPs
- PP3=99%

C/D show

- H_4 : Association with trait 1 and trait 2, one shared SNP
- PP4=82%

PLoS Genetics 2021 – Coloc-SuSiE

Figure 1.



Average posterior probability distributions in simulated data.
Each row is a different scenario:

- **Top row:** shows a scenario where traits 1 and 2 have distinct causal variants A and B.
- **Second row:** shows a scenario where traits 1 and 2 have one shared causal variants A.
- **Third row:** shows a scenario where traits 1 and 2 have one shared causal variants A and trait 2 has a second distinct causal variant B (not shared with trait 1).
- **Fourth row:** shows a scenario where traits 1 and 2 have two shared causal variants A and variant B.

Columns indicate the different analysis methods:

- **single** – coloc assuming 1 causal variant,
- **cond_it** indicating that coloc-conditioning was run in iterative mode, and
- **cond_abo** indicating it was run in "all but one" mode,
- **susie** indicating coloc-SuSiE, and
- **hybrid,** indicating both single coloc and coloc-SuSiE

Using Coloc and Coloc-SuSiE

- The coloc package can be used to perform genetic colocalisation analysis of two potentially related phenotypes, to ask whether they share common genetic causal variant(s) in a given region.
- **What is needed for input?**
  - Summary statistics of two traits for your region of interest
    - "trait" can be mRNA expression quantitative trait locus (eQTL), or protein quantitative trait locus (pQTL), or methylation quantitative trait locus (mQTL), etc.
  - **LD (for coloc-SuSiE)**

- https://chr1swallace.github.io/coloc/
- https://cran.r-project.org/web/packages/coloc/vignettes/a02_data.html

What is needed for input?

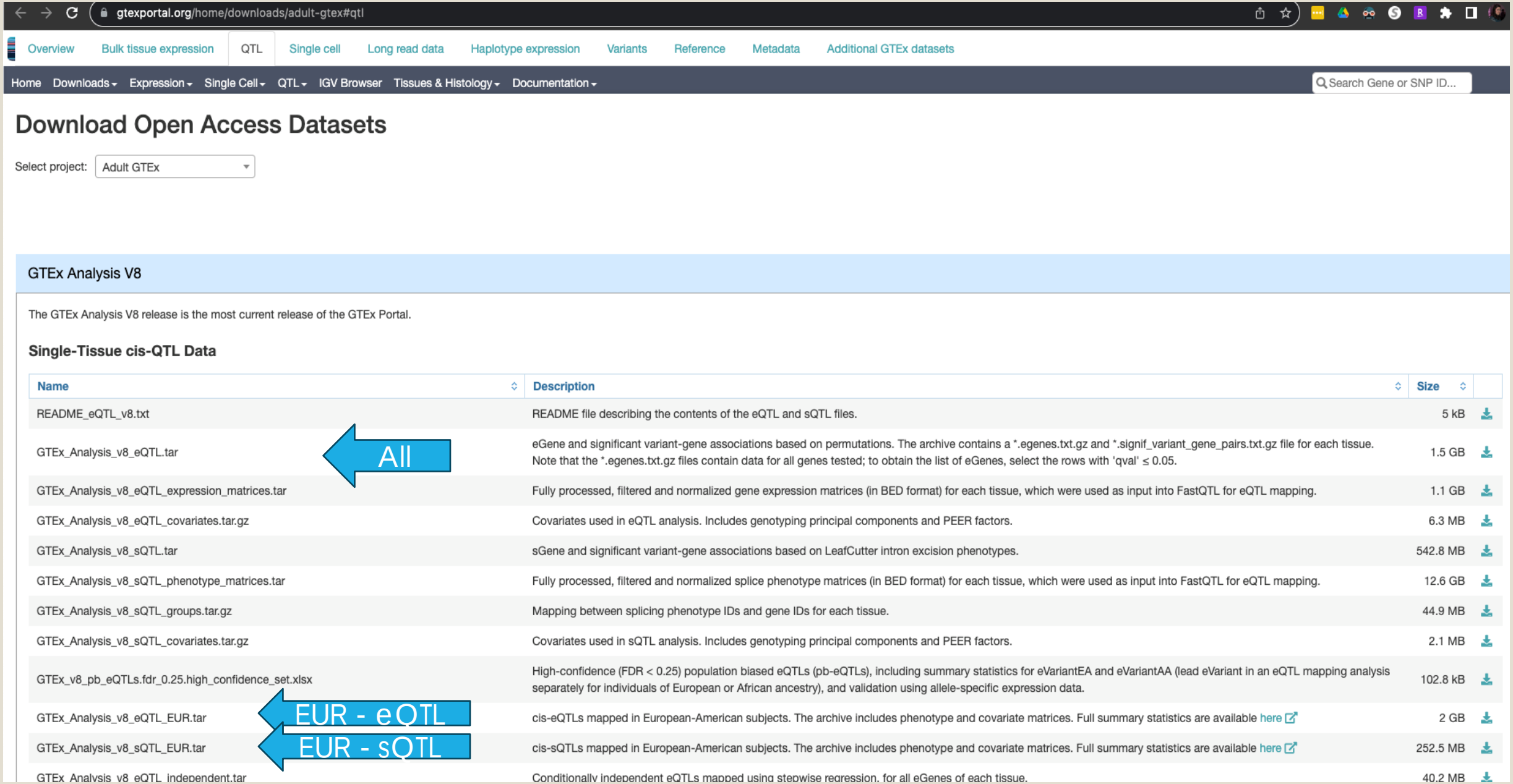
- **Summary statistics of two traits for your region of interest**
  - “trait” can be mRNA expression quantitative trait locus (eQTL), or protein quantitative trait locus (pQTL), or methylation quantitative trait locus (mQTL), etc.
- **LD matrix (for coloc-SuSiE),** required if assuming >1 causal variant

Some omics QTL sources

- **eQTL : other catalogs**
  - https://gtexportal.org/home/downloads/adult-gtex#qtl
  - https://www.ebi.ac.uk/eqtl/Data_access/
  - https://www.eqtlgen.org
- **pQTLs : mainly paper by paper right now**
  - https://www.ebi.ac.uk/gwas/publications/29875488
  - https://gtexportal.org/home/downloads/egtex/proteomics (14 people; 32 tissues)
- **meQTLs or mQTLs (methylation) : GTEx and others**
  - https://gtexportal.org/home/downloads/egtex/methylation (987 samples; 9 tissues)
  - https://www.epigenomicslab.com/online-data-resources/
- **mtQTLs or mQTLs (metabolomics): mainly paper by paper right now**
  - https://www.nature.com/articles/s41467-017-01972-9
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9081120/

Getting GTEx files for eQTLs

- https://gtexportal.org/home/downloads/adult-gtex#qtl
  - Count of individuals varies by tissue type (54 total tissues)
    - ~20 bladder
    - ~100-200 brain, small intestine, pituitary, liver
    - ~300-400 stomach, heart, aortic artery,
    - ~ 500 adipose, lung, skin, thyroid
    - ~700-800 whole blood, muscle
  - As an example, let's look here:
  - gtex-resources/GTEx_Analysis_v8_QTLs/GTEx_Analysis_v8_eQTL_all_associations


gtexportal.org/home/downloads/adult-gtex#qtl
Overview
Bulk tissue expression
QTL
Single cell
Long read data
Haplotype expression
Variants
Reference
Metadata
Additional GTEx datasets
Home
Downloads
Expression
Single Cell
QTL
IGV Browser
Tissues & Histology
Documentation
Search Gene or SNP ID...
Download Open Access Datasets
Select project:
Adult GTEx
GTEx Analysis V8
The GTEx Analysis V8 release is the most current release of the GTEx Portal.
Single-Tissue cis-QTL Data
Name
Description
Size
README_eQTL_v8.txt
README file describing the contents of the eQTL and sQTL files.
5 kB
GTEx_Analysis_v8_eQTL.tar
All
eGene and significant variant-gene associations based on permutations. The archive contains a *.egenes.txt.gz and *.signif_variant_gene_pairs.txt.gz file for each tissue. Note that the *.egenes.txt.gz files contain data for all genes tested; to obtain the list of eGenes, select the rows with 'qval' ≤ 0.05.
1.5 GB
GTEx_Analysis_v8_eQTL_expression_matrices.tar
Fully processed, filtered and normalized gene expression matrices (in BED format) for each tissue, which were used as input into FastQTL for eQTL mapping.
1.1 GB
GTEx_Analysis_v8_eQTL_covariates.tar.gz
Covariates used in eQTL analysis. Includes genotyping principal components and PEER factors.
6.3 MB
GTEx_Analysis_v8_sQTL.tar
sGene and significant variant-gene associations based on LeafCutter intron excision phenotypes.
542.8 MB
GTEx_Analysis_v8_sQTL_phenotype_matrices.tar
Fully processed, filtered and normalized splice phenotype matrices (in BED format) for each tissue, which were used as input into FastQTL for eQTL mapping.
12.6 GB
GTEx_Analysis_v8_sQTL_groups.tar.gz
Mapping between splicing phenotype IDs and gene IDs for each tissue.
44.9 MB
GTEx_Analysis_v8_sQTL_covariates.tar.gz
Covariates used in sQTL analysis. Includes genotyping principal components and PEER factors.
2.1 MB
GTEx_v8_pb_eQTLs.fdr_0.25.high_confidence_set.xlsx
High-confidence (FDR < 0.25) population biased eQTLs (pb-eQTLs), including summary statistics for eVariantEA and eVariantAA (lead eVariant in an eQTL mapping analysis separately for individuals of European or African ancestry), and validation using allele-specific expression data.
102.8 kB
GTEx_Analysis_v8_eQTL_EUR.tar
EUR - eQTL
cis-eQTLs mapped in European-American subjects. The archive includes phenotype and covariate matrices. Full summary statistics are available here
2 GB
GTEx_Analysis_v8_sQTL_EUR.tar
EUR - sQTL
cis-sQTLs mapped in European-American subjects. The archive includes phenotype and covariate matrices. Full summary statistics are available here
252.5 MB
GTEx_Analysis_v8_eQTL_independent.tar
Conditionally independent eQTLs mapped using stepwise regression, for all eGenes of each tissue.
40.2 MB

Once you have your summary statistics selected, what do you do?

- Select a REGION
  - Do **not** subset by
    - Significance
    - MAF
    - Dataset overlap
  - Only include the region of interest
    - this is not a whole chromosome
    - Usually a gene region, or specific size where something is happening in your trait of interest (e.g. 250kb 500kb, etc)

Required input – PER REGION

- effect size: β (beta)
- its uncertainty: varβ (varbeta).
  - varβ = (standard error of β)^2
- SNP IDs: snp IDs needs to match across summary datasets and be unique
- Positions: (position) are useful if you want to use coloc's plot_dataset().
- type of outcome of trait: 'cc' for case-control, or 'quant' for quantitative
- sdY, : the standard deviation of the outcome, if missing can estimate using:
  - N
  - MAF

As a reminder these are the Hypotheses evaluated:

H_0: neither trait has a genetic association in the region

H_1: only trait 1 has a genetic association in the region

H_2: only trait 2 has a genetic association in the region

H_3: both traits are associated, but with different causal variants

H_4: both traits are associated and share a single causal variant

Colocalization typically defined as:
H4 PP >0.5

We will use coloc with a small example

- Assumes 1 causal variant per locus
- Can colocalize using p-values or beta and variance of beta
- When using beta, allele order matters!

EXAMPLE:

- Trait 1: Looking at ONE locus associated with height in a Hispanic GWAS
- Trait 2: One nearby gene from GTEx eQTLs

We will also use coloc.susie with the same small example

- Assumes >1 causal variant per locus
- Colocalize using beta and variance of beta
- Additionally requires an LD matrix
  - What LD matrix?
  - Your own data?, 1000 genomes? Subset of 1000 genomes?
  - What if associations are from different populations?
    - Use a different LD matrices!

EXAMPLE:

- Trait 1: Looking at ONE locus associated with height in a Hispanic GWAS
- Trait 2: One nearby gene from GTEx eQTLs
- LD from 1000 Genomes AMR: for Height
- LD from 1000 Genomes EUR: for GTEx eQTLs

What the input files look like

```
> print(head(eqtl))
            phenotype_id           variant_id tss_distance REF ALT   maf_trc     beta_asc beta_se_asc  tstat_asc   pval_asc samples_asc
rs1005307 ENSG00000159363 chr1_17122574_C_T_b38       110646   C   T 0.3462687  0.028255275  0.01526278  1.8512529 0.06488703         392
rs1005753 ENSG00000159363 chr1_17118274_G_T_b38       106346   G   T 0.4276120  0.007713964  0.01507085  0.5118467 0.60904729         392
rs1007150 ENSG00000159363 chr1_16294688_C_T_b38      -717240   C   T 0.2820895 -0.012296460  0.01726603 -0.7121764 0.47678025         392
rs1007887 ENSG00000159363 chr1_16182534_C_T_b38      -829394   C   T 0.4022388 -0.004990482  0.01537443 -0.3245963 0.74566025         392
rs1008529 ENSG00000159363 chr1_16256191_G_A_b38      -755737   G   A 0.2753732 -0.009709161  0.01707139 -0.5687387 0.56986000         392
rs1010069 ENSG00000159363 chr1_16026442_G_A_b38      -985486   G   A 0.4753732  0.011110632  0.01444001  0.7694337 0.44210046         392
          dof_asc      rsID      varbeta
rs1005307     391 rs1005307 0.0002329526
rs1005753     391 rs1005753 0.0002271305
rs1007150     391 rs1007150 0.0002981158
rs1007887     391 rs1007887 0.0002363731
rs1008529     391 rs1008529 0.0002914324
rs1010069     391 rs1010069 0.0002085139
> print(head(gwas))
          MarkerName Chromosome Position EA NEA    EAF Nsample HetISq HetPval  beta_0   se_0  p_value   varbeta
rs1005307  rs1005307          1 17449069  T   C 0.6574 43042.8    0.0 0.72030  0.0133 0.0062 0.031470 3.844e-05
rs1005753  rs1005753          1 17444769  T   G 0.6225 48664.9    0.0 0.65860  0.0175 0.0058 0.002582 3.364e-05
rs1007150  rs1007150          1 16621183  T   C 0.5798 49771.0    0.0 0.58170 -0.0178 0.0055 0.001369 3.025e-05
rs1007887  rs1007887          1 16509029  T   C 0.5015 50536.0   24.7 0.08499 -0.0105 0.0054 0.053240 2.916e-05
rs1008529  rs1008529          1 16582686  A   G 0.5990 48665.0   18.4 0.16110 -0.0197 0.0057 0.000493 3.249e-05
rs1010069  rs1010069          1 16352937  A   G 0.4081 55053.8    0.0 0.87840  0.0082 0.0054 0.127600 2.916e-05
>
```

Summary of output

Summary of results based on p-values

```
> print(my.res)
Coloc analysis of trait 1, trait 2

SNP Priors
   p1    p2   p12
1e-04 1e-04 1e-05

Hypothesis Priors
         H0     H1     H2        H3      H4
 0.09124558 0.3682 0.3682 0.1355344 0.03682

Posterior
       nsnps           H0           H1           H2           H3           H4
3.682000e+03 1.026948e-13 7.764444e-08 3.293488e-08 2.392499e-02 9.760749e-01
```

Summary of results based on betas and variance of betas

```
> print(my.res.bvar)
Coloc analysis of trait 1, trait 2

SNP Priors
   p1    p2   p12
1e-04 1e-04 1e-05

Hypothesis Priors
         H0     H1     H2        H3      H4
 0.09124558 0.3682 0.3682 0.1355344 0.03682

Posterior
       nsnps           H0           H1           H2           H3           H4
3.682000e+03 1.884775e-15 1.227796e-09 4.814382e-08 3.039265e-02 9.696073e-01
```

Subset output to find top SNP that colocalizes

Summary of results based on p-values

```
> print(subset(my.res$results,SNP.PP.H4>0.1))
          snp position pvalues.df1 MAF.df1 N.df1        V.df1    z.df1     r.df1 lABF.df1  pvalues.df2   MAF.df2 N.df2       V.df2    z.df2
3613 rs9435734 16984695   1.283e-12  0.5443 50000 4.031648e-05 7.096129 0.9982114 21.96934 3.199793e-12 0.4962686   392 0.005102325 6.968658
        r.df2 lABF.df2 internal.sum.lABF SNP.PP.H4
3613 0.8151487  18.9486          40.91794 0.6201555
> print(subset(my.res.bvar$results, SNP.PP.H4>0.1))
          snp position     V.df1   z.df1     r.df1 lABF.df1        V.df2    z.df2     r.df2 lABF.df2 internal.sum.lABF SNP.PP.H4
3613 rs9435734 16984695 2.809e-05 7.09434 0.9987531  21.7899 0.0001136237 7.195579 0.9949754 23.11139          44.90129 0.6152423
```

Summary of results based on betas and variance of betas

Additional resources

- ColocQuaiL
  - The ColocQuiaL pipeline provides a framework to perform colocalization analysis of GWAS signals with expression quantitative trait loci (eQTL) and splicing quantitative trait loci (sQTL) to connect GWAS signals to candidate causal genes at scale across the genome and returns summary files and locus visualization plots to allow for detailed review of the results.
  - https://academic.oup.com/bioinformatics/article/38/18/4409/6650620?login=false
  - https://github.com/bvoightlab/ColocQuiaL

THANKS!

- Hung-Hsin Chen
- Heather Highland
- Carlos González-Carballo
- Fernando Rivas
- Paulina Baca
- Jesus Alegre
- Jaime Brumen

- All of YOU!!