nature genetics

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41588-019-0404-0

In the format provided by the authors and unedited.

Abundant associations with gene expression complicate GWAS follow-up

Boxiang Liu[1,2,9*], Michael J. Gloudemans[2,3,9], Abhiram S. Rao[2,4], Erik Ingelsson[5,6,7] and Stephen B. Montgomery[2,8*]

[1]Department of Biology, Stanford University, Stanford, CA, USA. [2]Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA. [3]Biomedical Informatics Training Program, Stanford University School of Medicine, Stanford, CA, USA. [4]Department of Bioengineering, Stanford University, Stanford, CA, USA. [5]Department of Medicine, Division of Cardiovascular Medicine, Stanford University School of Medicine, Stanford, CA, USA. [6]Stanford Cardiovascular Institute, Stanford University, Stanford, CA, USA. [7]Stanford Diabetes Research Center, Stanford University, Stanford, CA, USA. [8]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. [9]These authors contributed equally: Boxiang Liu, Michael J. Gloudemans.
*e-mail: jollier.liu@gmail.com; smontgom@stanford.edu

Supplementary Material for Abundant Associations with Gene Expression Complicate GWAS Follow-up

Boxiang Liu[1,2,*], Mike J. Gloudemans[2,3,*], Abhiram S. Rao[2,4], Erik Ingelsson[5,6,7], and Stephen B. Montgomery[2,8]

[1]Department of Biology, Stanford University
[2]Department of Pathology, Stanford University School of Medicine
[3]Biomedical Informatics Training Program, Stanford University School of Medicine
[4]Department of Bioengineering, Stanford University
[5]Department of Medicine, Division of Cardiovascular Medicine, Stanford University School of Medicine, Stanford, CA 94305
[6]Stanford Cardiovascular Institute, Stanford University, Stanford, CA 94305
[7]Stanford Diabetes Research Center, Stanford University, Stanford, CA 94305
[8]Department of Genetics, Stanford University School of Medicine
*These authors contributed equally

March 27, 2019

List of Figures

1 Supplementary Methods

1.1 Review of GWAS literature

To determine the proportion of GWAS literature that used eQTL reference datasets and colocalization methods, we surveyed GWAS studies published on *Nature Genetics* between January 2017 to August 2018. We found 63 papers in total (Supplementary Table 1), 50 of which used eQTL reference database and 15 used colocalization methods. We define colocalization methods as any model that compare the distribution of GWAS and eQTL summary statistics, such as coloc [1] and eCAVIAR [2] (see Supplementary Table 2 for a full list of methods). Further, we considered non-model-based methods, such as visualizing eQTL and GWAS effect sizes with scatter plots, as colocalization methods as well. Supplementary Table 1 contains the details of colocalization methods used by each GWAS study.

1.2 The LocusCompare Web Server and the LocusCompareR R package

The LocusCompare web server is implemented in Shiny v1.1.0 with MySQL v5.6.25 as the database. The current database schema is depicted in Supplementary Figure 6. The GWAS and eQTL tables stores information about variant rsID, trait, and p-value. Additional variant-level information from

1000 Genomes phase 3 and gene-level information from GENCODE v19 are stored in separate tables to avoid redundancy. The LocusCompare plot requires LD r^2 information to color each data point. We calculate ancestry-specific LD for individuals of African, East Asian, European, South Asian, and Native American descent. We used plink1.9 with the option "–keep-allele-order –maf 0.01 –keep <individuals_file> –r2 –ld-window 9999999 –ld-window-kb 10000". To speed up the web server, we store all pairwise LD information in the MySQL database.

The LocusCompare web server is designed for single queries and manual exploration. To facilitate batch queries and programmatic access, we developed the LocusCompareR R package with instruction on the GitHub page (https://github.com/boxiangliu/locuscomparer). Similar to the web server, the R package will query the MySQL database for LD information.

1.3 Colocalization analysis

To identify the subset of genomic loci and the associated genes to test for colocalization, we started with our list of all GWAS traits. Since the direction of effect is required to run FINEMAP [3] and eCAVIAR [2], we removed all GWAS traits with unspecified direction of effect. For each remaining GWAS, we selected all loci with an nominal $p < 5 \cdot 10^{-8}$, as long as the lead SNP at the locus was at least 1MB from all other selected lead SNPs. In cases of conflict, the SNP with stronger association was always selected first as the lead SNP for that locus. For each of these loci, we then identified the set of all gene/tissue combinations for which the GWAS lead SNP was a cis-eQTL associated with the expression of that gene in that tissue ($p < 10^{-6}$), similar to the criteria that would be used in a naive eQTL lookup without colocalization testing.

For all trait/locus/gene/tissue combinations that passed the above cutoffs, we took the subset containing all SNPs at the locus that were tested in both the GWAS and eQTL studies, and that were also present in the 1000 Genomes VCF [4]. Whenever possible, we aligned directions of effect for the eQTL and the GWAS to the ref/alt direction found in the 1000 Genomes VCF. We then ran FINEMAP [3] to produce posterior causal probabilities for each of these SNPs, in both the GWAS and the eQTL studies. We used the full 1000 Genomes VCF as a reference for the LD statistics in all studies, and we limited the number of causal variants at each of the GWAS and eQTL loci to a maximum of 1 for computational feasibility. We then analyzed these causal probabilities with a custom script to compute the colocalization posterior probability (CLPP) for the entire locus, as described in the eCAVIAR method:

$$CLPP = \sum_{i=1}^{N} g_i \cdot e_i$$

Where g_i is the probability that the i-th SNP is the causal variant for the GWAS, e_i is the probability that the i-th SNP is the causal variant for the eQTL trait, and N is the total number of variants at the locus.

The authors of eCAVIAR suggested that in practice, CLPP > 0.01 indicates a reasonably high probability of colocalization. Naturally, higher-CLPP loci within the same study typically have a higher probability of sharing the same causal variant. However, we do not recommend comparing CLPP scores across different GWAS studies, as differing SNP densities and LD compositions complicates this comparison.

The loci shown on the Colocalization page of the website for a given GWAS-eQTL combination include the full set of all genes that passed the p-value cutoffs in that pairing; that is, they show all colocalization tests performed at that locus, regardless of whether or not they showed colocalization.

The complete wrapper for the colocalization analysis is freely available online with detailed instructions at `https://bitbucket.org/mgloud/production_coloc_pipeline`. We tested approximately 83,000 trait/locus/gene/tissue combinations for the full set of GWAS, which took roughly a week when running on eight separate threads. Although eCAVIAR [2] has its own fine-mapping functionality, we used FINEMAP [3] for this step instead because it runs significantly faster with very similar overall results.

1.4 Data for main and supplementary figures

For each figure, Supplementary Table 3 list the studies on which the x-axis and y-axis were based.

Figure	**x-axis**	**y-axis**
Fig. 1, S1 and S2	Zhao, W. et al. Nat Genet (2017)	GTEx v6p eQTL
Fig. S3	Nelson, C.P. et al. Nat Genet (2017)	Kanai, M. et al. Nat Genet (2018)
Fig. S4	Nikpay, M. et al. Nat Genet (2015)	GTEx v6p eQTL
Fig. S5	Nikpay, M. et al. Nat Genet (2015)	GTEx v6p eQTL

Table 3: Studies used for each figure

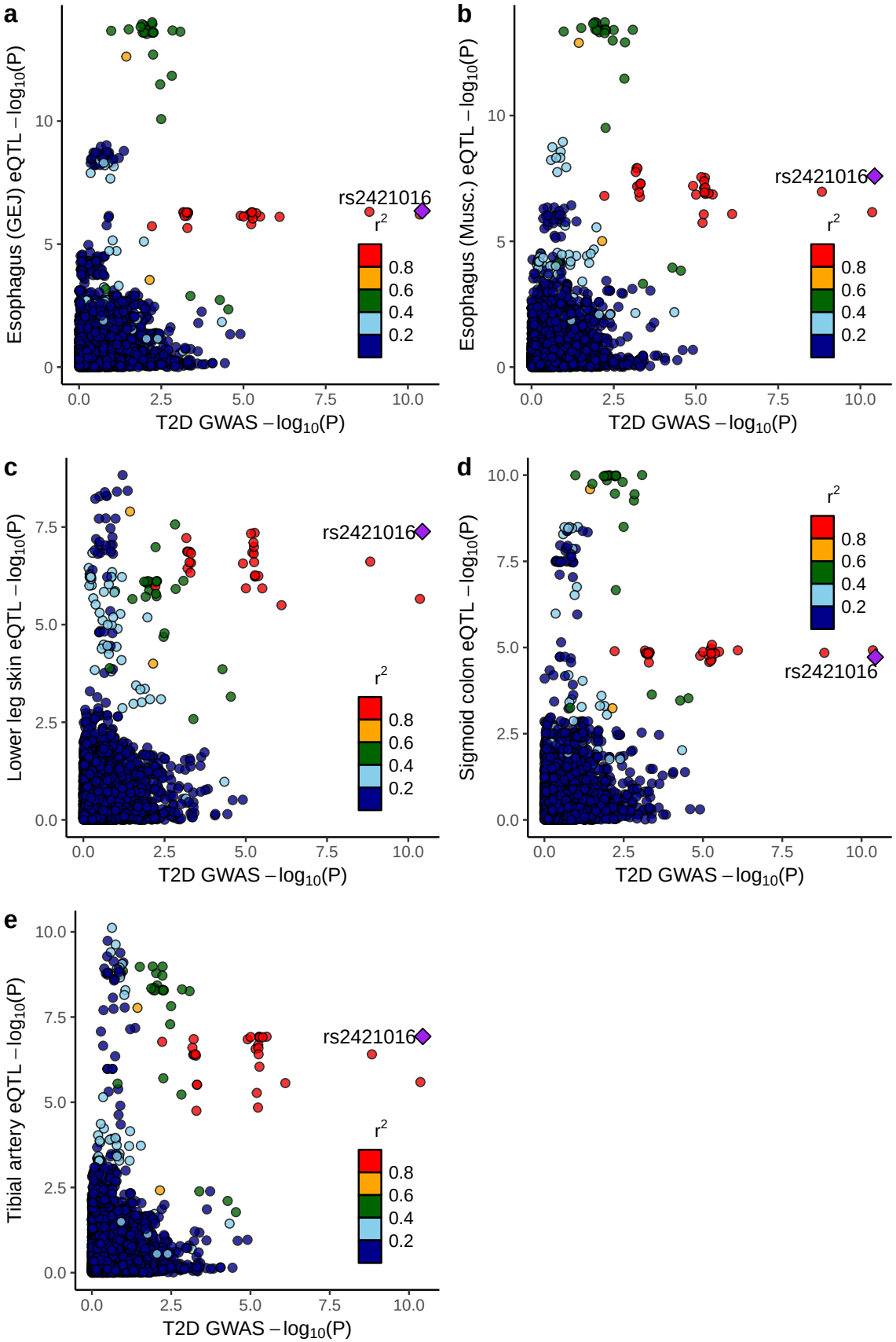

Figure 1: ***ARMS2* colocalization across tissues**. In all five tissues including esophagus (gastro-esophageal junction), esophagus (muscularis), lower-leg skin, sigmoid colon, and tibial artery, *ARMS2* shows two independent peaks towards the top-left and the bottom-right corners.
The eQTL p-values were extracted from the GTEx Esophagus - Gastroesophageal Junction (n = 127 individuals) and Esophagus - Muscularis (n = 218 individuals), Skin - Sun-Exposed Lower Leg (n = 302 individuals), Colon - Sigmoid (n = 124), and Artery - Tibial (n = 285) datasets based on a simple linear regression model. The GWAS p-values were extracted from Zhao *et al* [5] (n_{case} = 73,337 and $n_{control}$ = 192,341 individuals) based on a logistic regression model and meta-analysis.

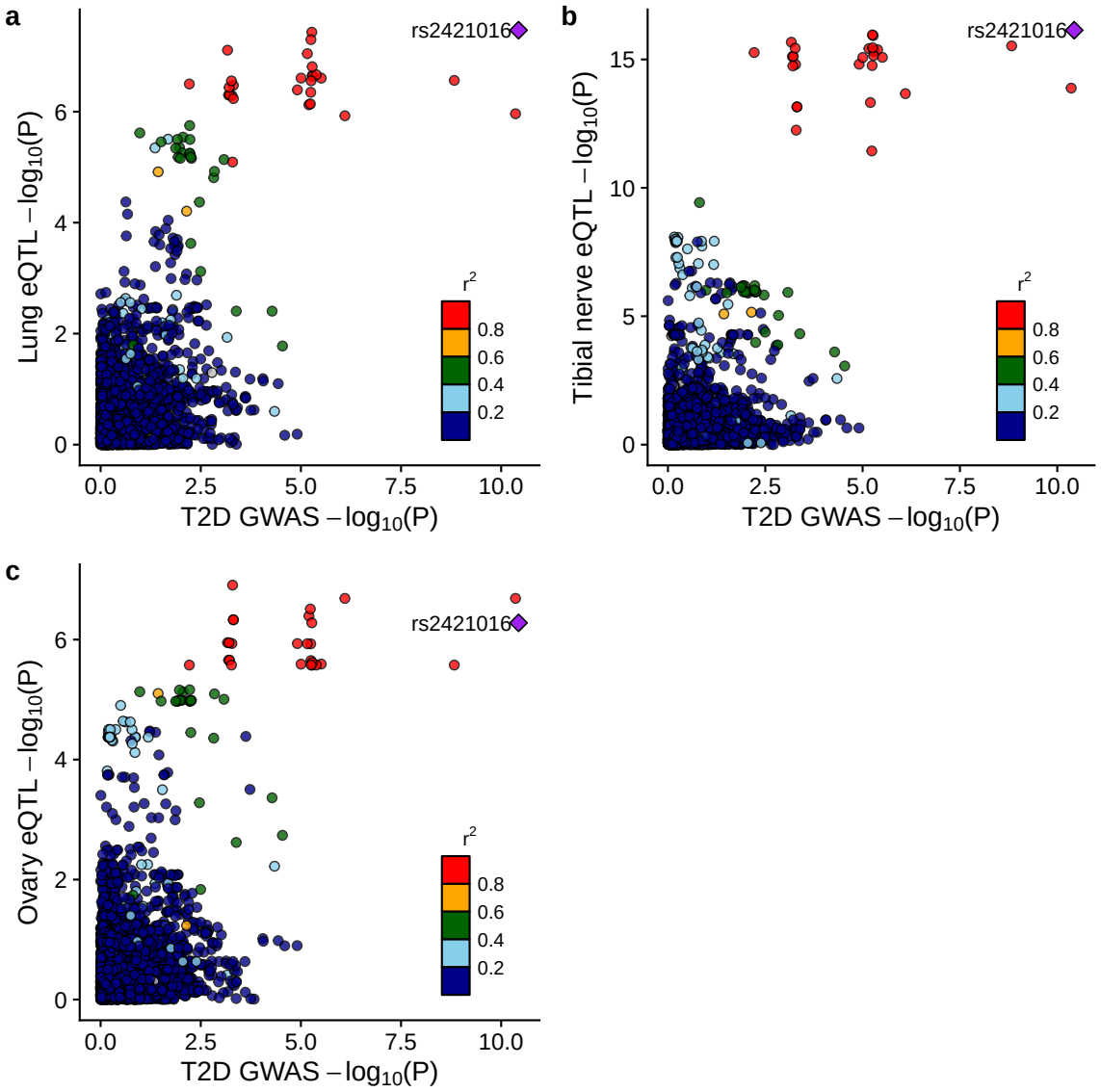

Figure 2: ***PLEKHA1* colocalization across tissues.** In all three tissues including lung, tibial nerve and ovary, *PLEKHA1* shows clear colocalization patterns. The eQTL p-values were extracted from the GTEx Lung (n = 278), Nerve - Tibial (n = 256), and Ovary (n = 85) datasets based on a simple linear regression model. The GWAS p-values were extracted from Zhao *et al* [5] (n = 265,678 individuals) based on a logistic regression model and meta-analysis.

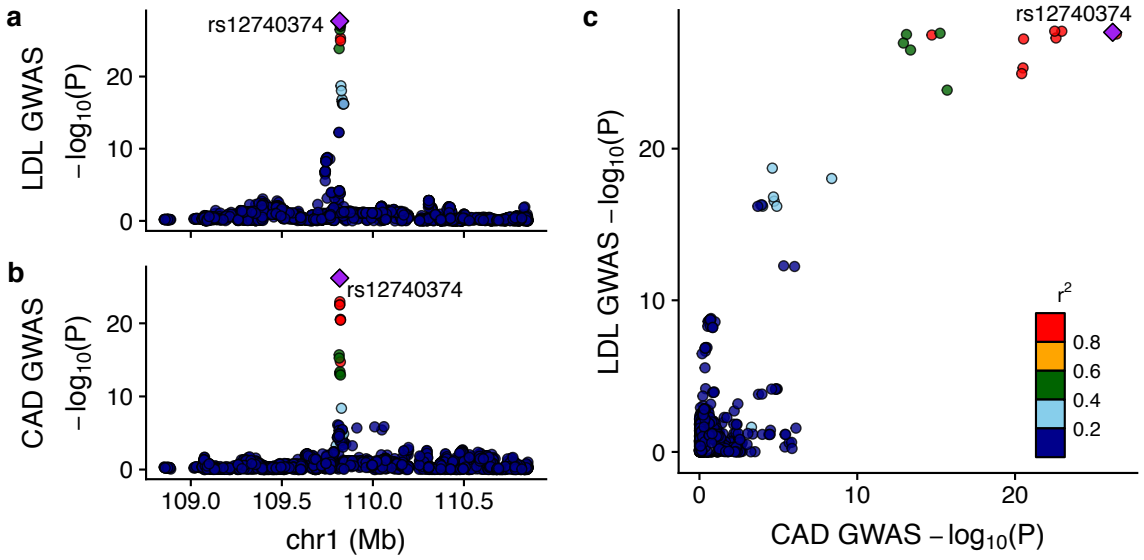

Figure 3: **LocusCompare demonstrates a pleiotropic effect.** A well-known pleiotropic effect showcases that the SNP rs12740374 is associated with LDL cholesterol and coronary artery disease risks. The Coronary Artery Disease (CAD) GWAS p-values were extracted from Nelson *et al* [6] (n_{case} = 10,801 and $n_{control}$ = 137,914 individuals) based on a logistic regression model and meta-analysis. The LDL GWAS p-values were extracted from Kanai *et al* [7] (n = 162,255 individuals) based on a linear regression model.

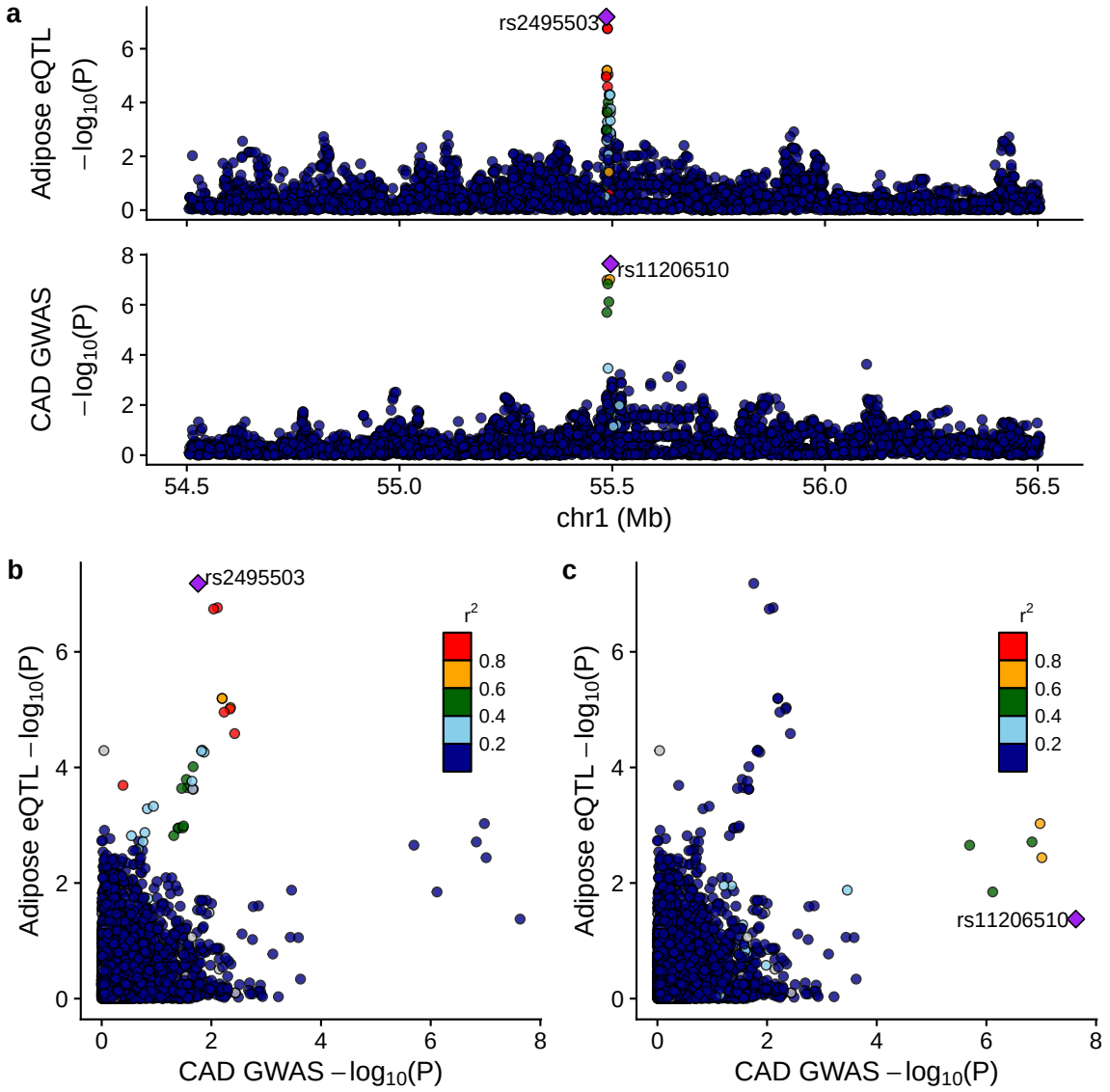

Figure 4: **LocusCompare helps dissect an apparent colocalization.** Manhattan plots show an apparent colocalization between CAD GWAS and *PSCK9* eQTL in visceral adipose. However, the lead variants are different across two studies. LocusCompare shows that the lead variants are independent. The eQTL p-values were extracted from the GTEx Adipose - Visceral Omentum (n = 185) based on a linear regression model. The GWAS p-values were extracted from Nikpay *et al* [8] (n_{case} = 60,801 and $n_{control}$ = 123,504 individuals) based on a logistic regression model and meta-analysis.

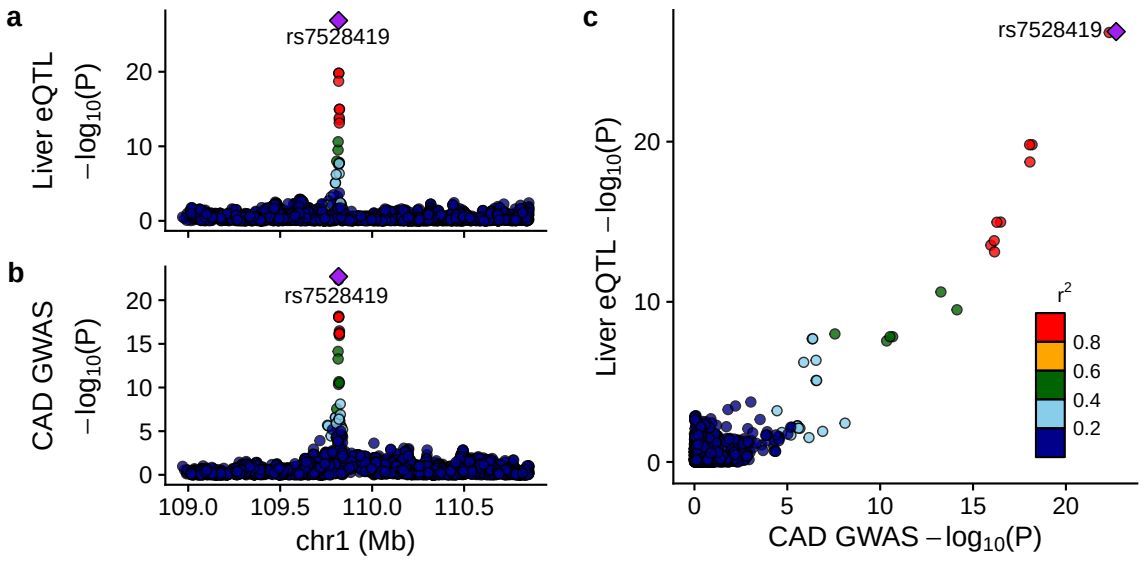

Figure 5: **Example of a *bona fide* colocalization signal.** Colocalization between *SORT1* locus in Coronary Artery Disease GWAS by Nikpay *et al.* (2015) and *SORT1* eQTL in Liver. The eQTL p-values were extracted from the GTEx Liver (n = 97) based on a linear regression model. The GWAS p-values were extracted from Nikpay *et al* [8] (n_{case} = 60,801 and $n_{control}$ = 123,504 individuals) based on a logistic regression model and meta-analysis.

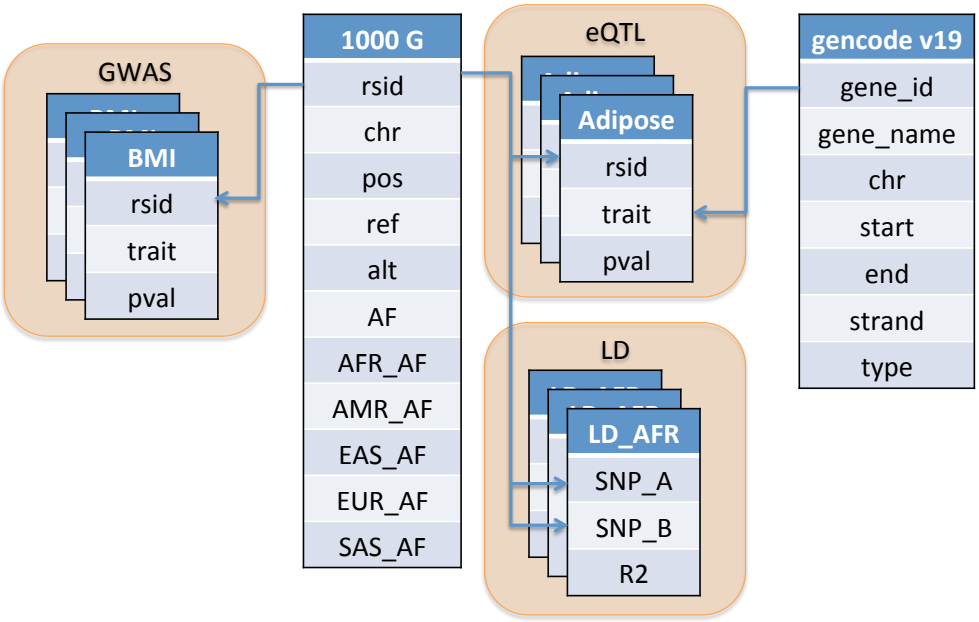

Figure 6: **LocusCompare database schema**. The LocusCompare database consists of five types of tables: 1) Variant 2) Gene 3) LD 4) GWAS and 5) QTL. Each table is indexed by a multiple keys to allow fast access and has foreign keys to ensure referential integrity.

References

[1] Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS genetics* **10**, e1004383 (2014).

[2] Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *The American Journal of Human Genetics* **99**, 1245–1260 (2016).

[3] Benner, C. *et al.* Finemap: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).

[4] 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

[5] Zhao, W. *et al.* Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. *Nature Genetics* **49**, 1450–1457 (2017).

[6] Nelson, C. P. *et al.* Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nature Genetics* **49**, 1385 (2017).

[7] Kanai, M. *et al.* Genetic analysis of quantitative traits in the japanese population links cell types to complex human diseases. *Nat Genet* **50**, 390–400 (2018).

[8] The CARDIoGRAMplusC4D Consortium. A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics* **47**, 1121–1130 (2015).