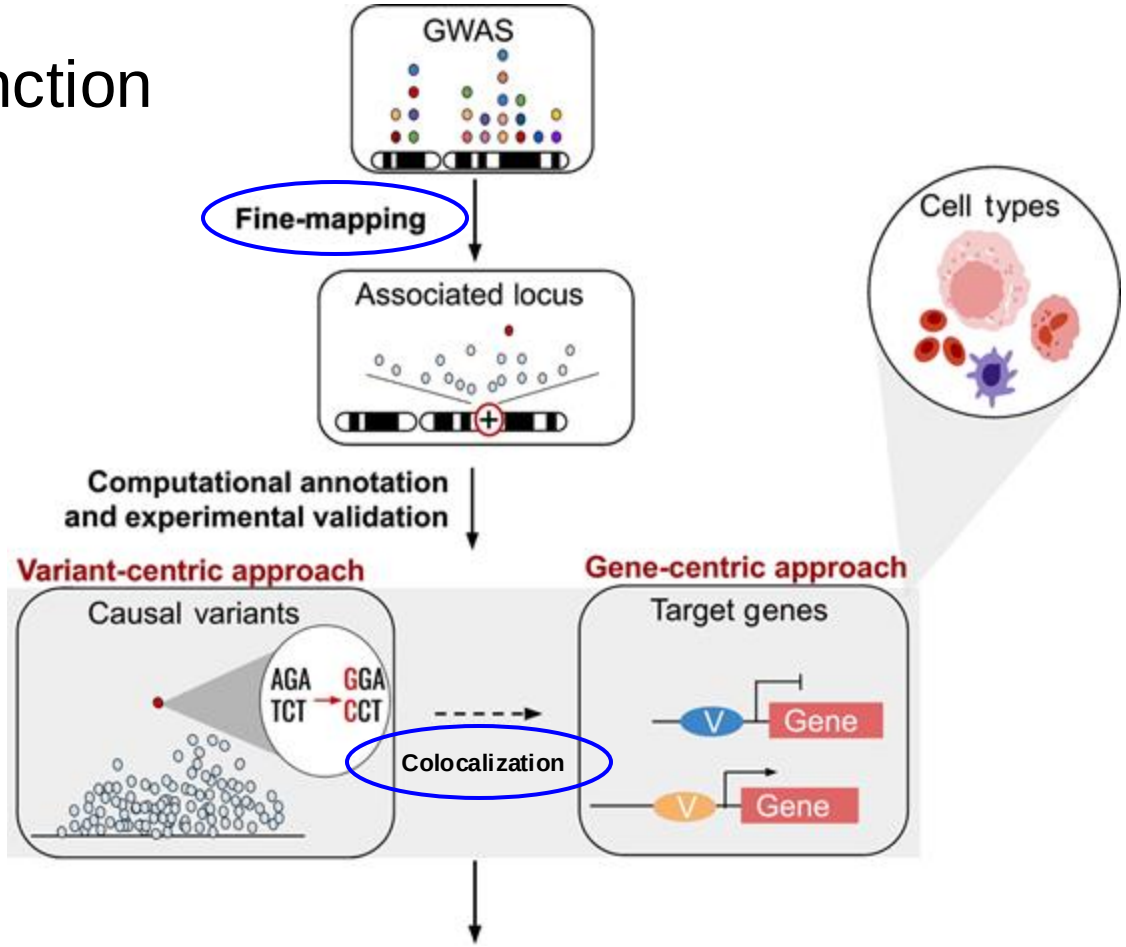
Fine-mapping and colocalization

Ran Cui, Ph.D.
Neale lab and Daly lab at Broad Institute
Analytic and Translational Genetics Unit, MGH

Outline

- A context: Variant to Function model (V2F)
- Fine-mapping
  - The goal of fine-mapping
  - Factors that influence fine-mapping
  - Overview of methods and benchmarking
  - Multi-cohort fine-mapping
  - Resources
- Colocalization
  - The goal of colocalization
  - Methods and outputs
  - How well does it work
  - Resources

Variant to Function (V2F)



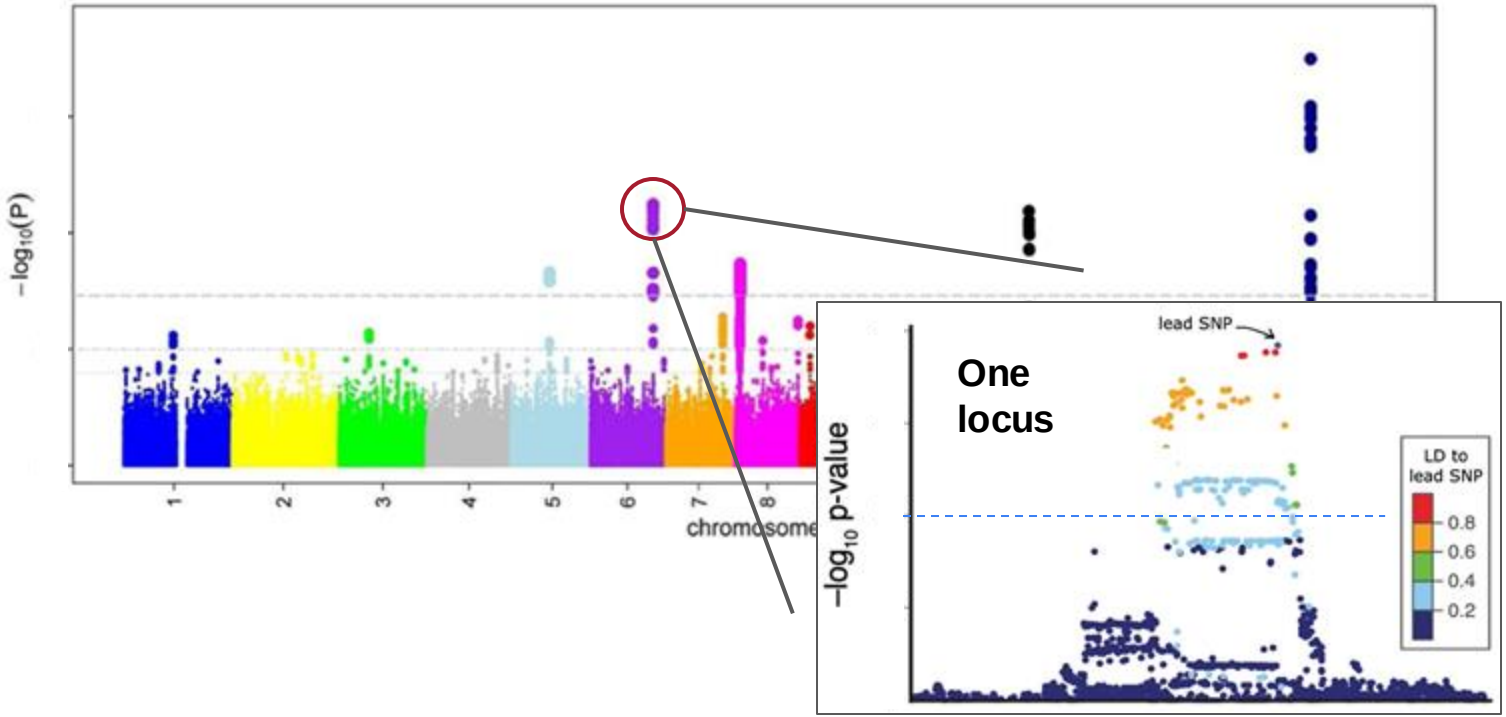
Mechanisms → Therapeutics

Nandakumar et al. Trends Genet. 2021

Part 1: Fine-mapping

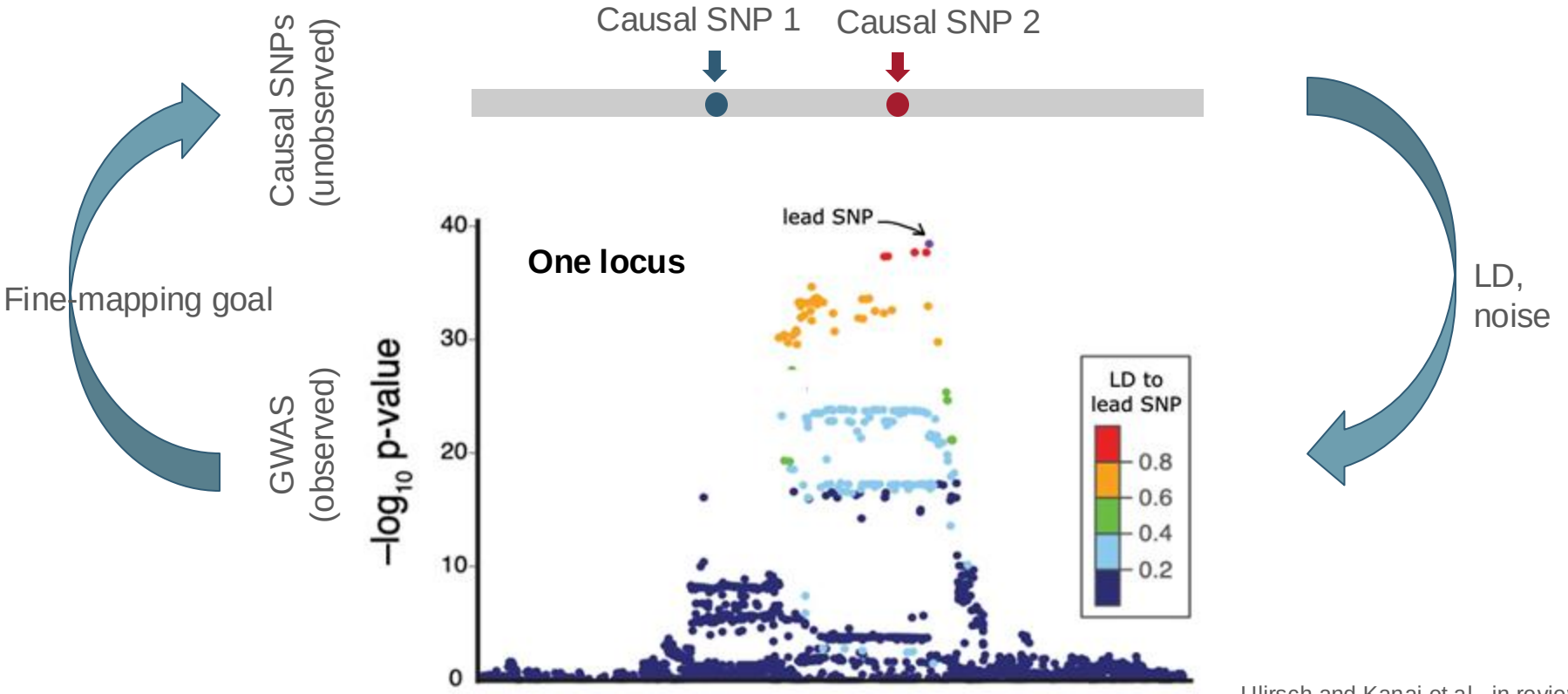
The goal of fine-mapping

GWAS aims to detect association



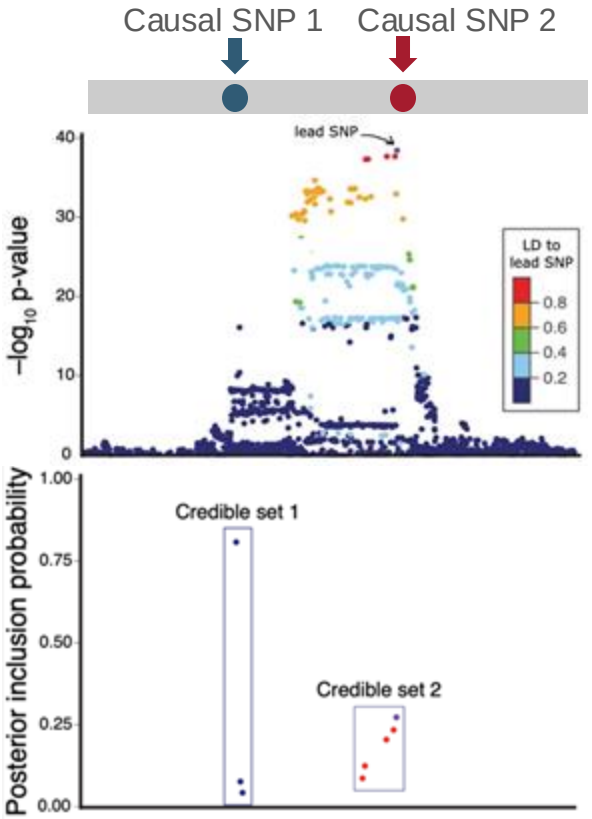
Association is not necessarily causation (cause: influence target trait in a nontrivial way)

Fine-mapping aims to nominate causal variants



Ulirsch and Kanai et al., in review

Fine-mapping outputs PIP and credible sets



Goal

Outputs

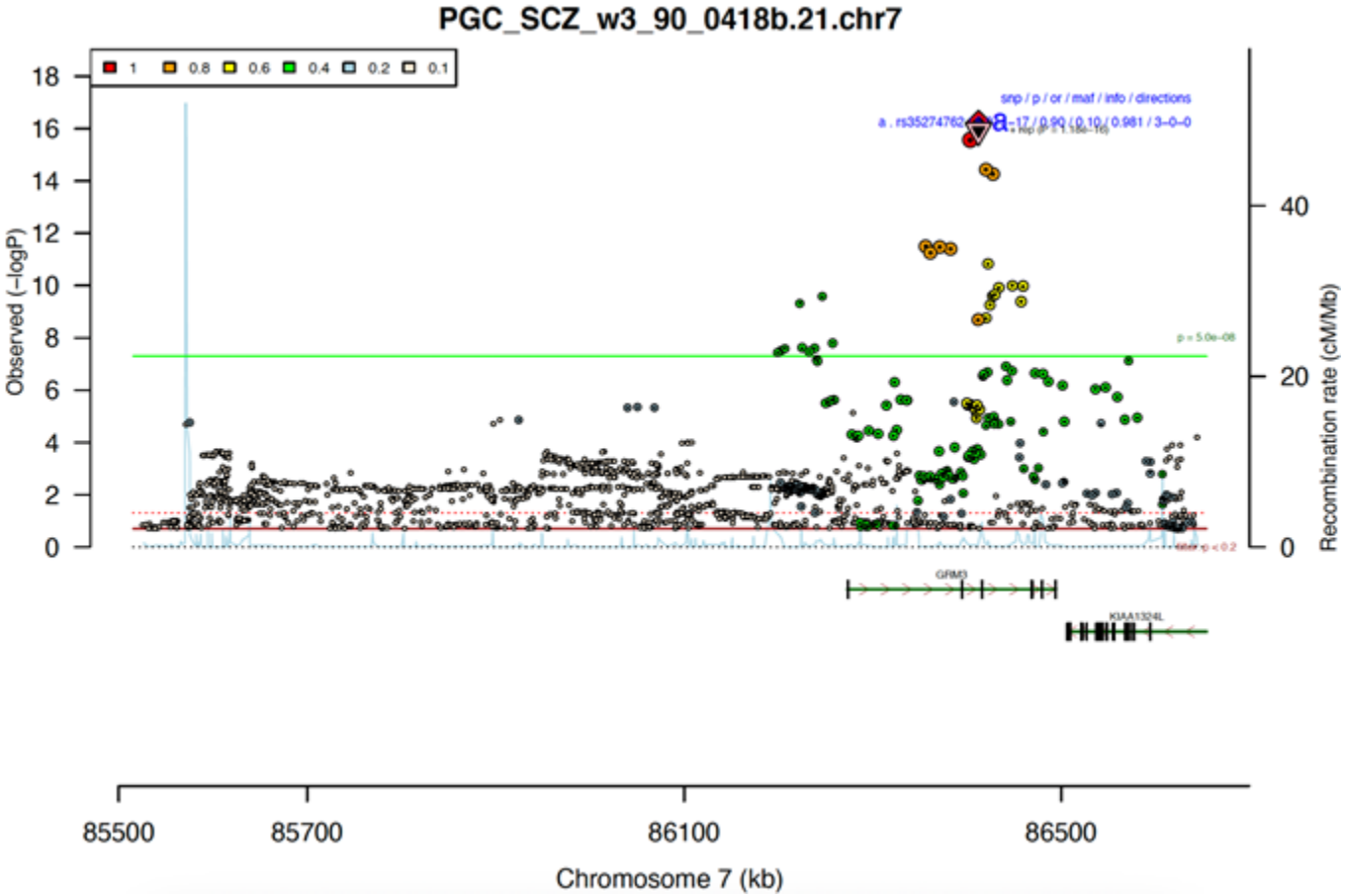
Outputs

Posterior Inclusion Probability (PIP) is the probability that a SNPs is causal given the observed data.

Credible sets capture the uncertainty of the identity of causal SNP due to LD.

Ulirsch and Kanai et al., in review

A simpler case

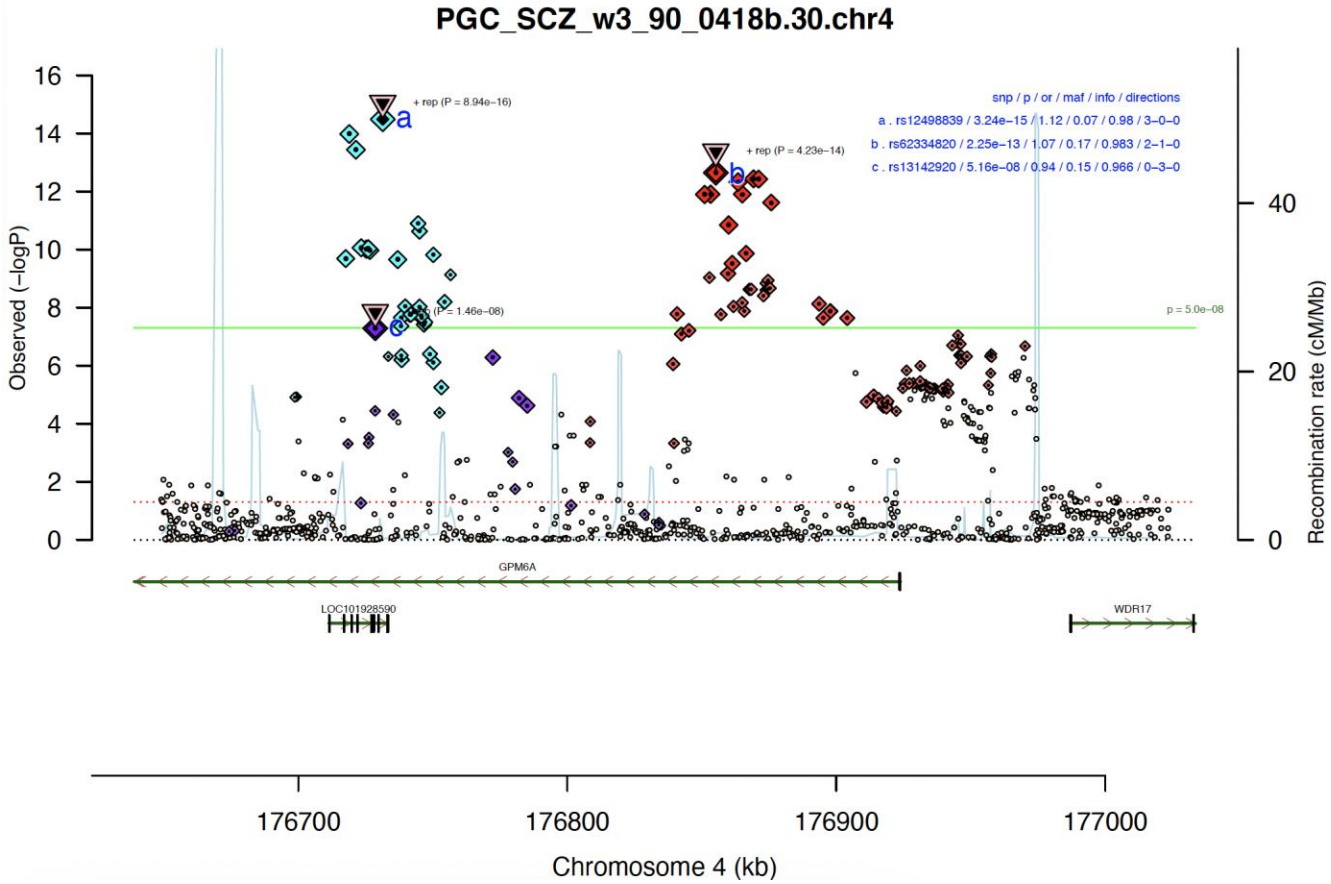


PIPs:
rs6943762: 0.591
rs35274762: 0.362

Corresponding gene:
GRM3

Trubetskoy et al. 2022 Nature

Multiple independent signals

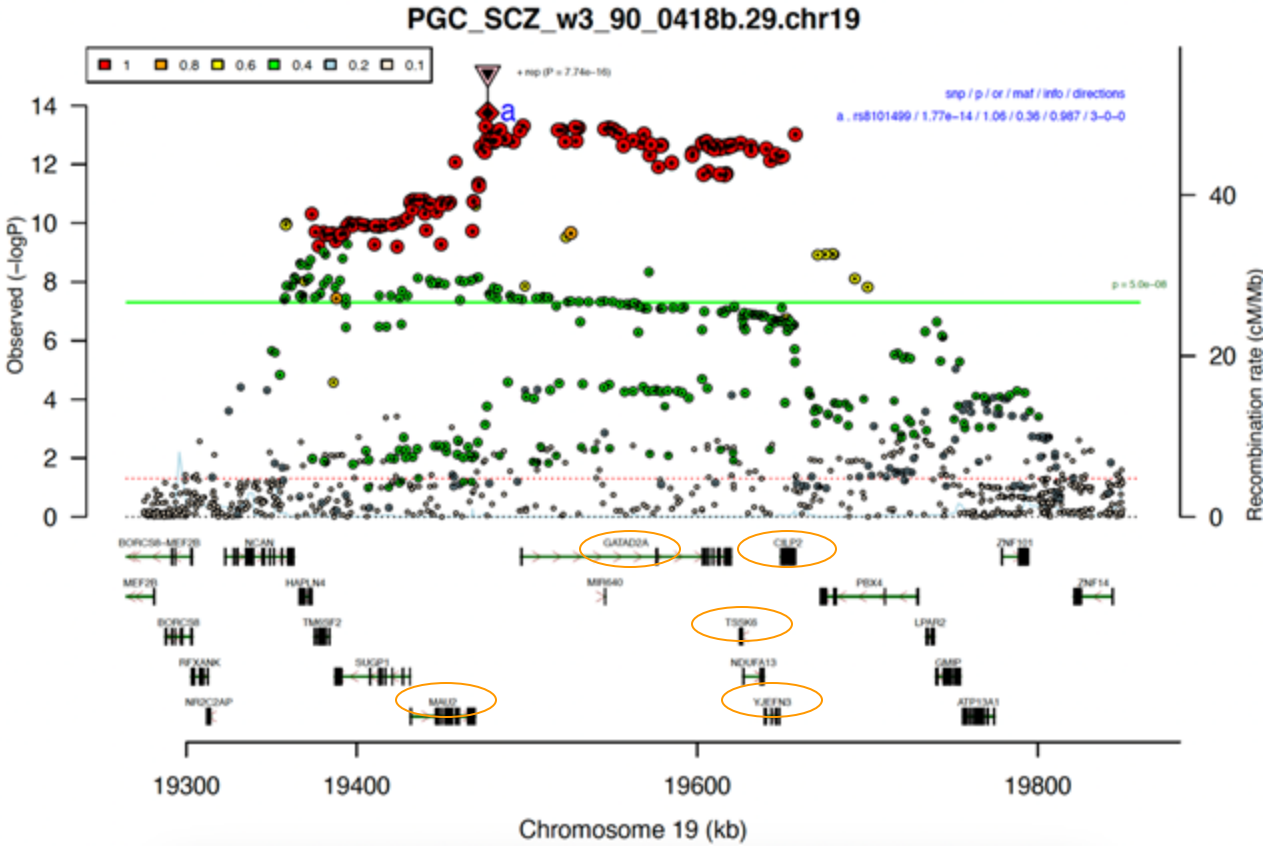


PIPs:
(a) rs12498839: 0.33
(b) rs62334820: 0.134
(c) rs13142920: 0.75

Corresponding gene:
GPM6A

Trubetskoy et al. 2022 Nature

A more complicated case



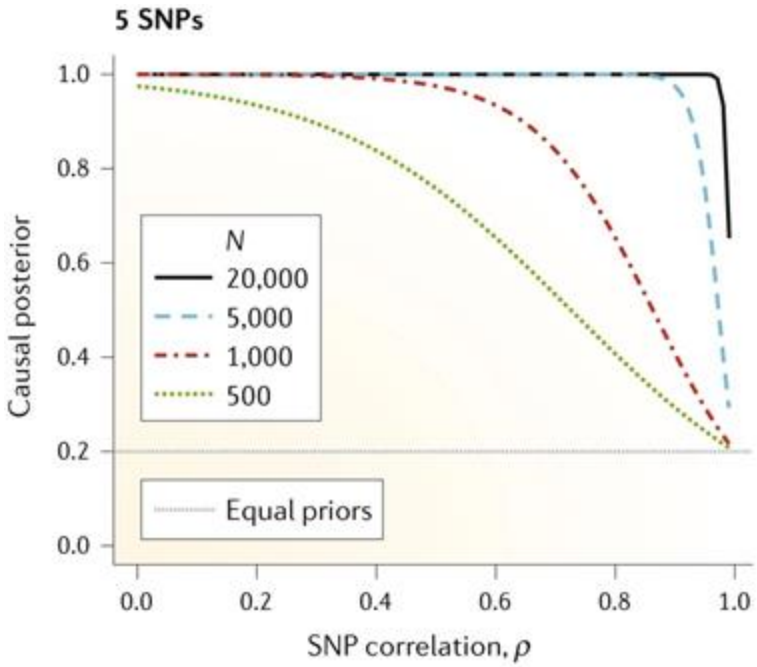
Credible sets spanning:
MAU2
GATAD2A
TSSK6
YJEFN3
CILP2

Trubetskoy et al. 2022 Nature

Factors that influence fine-mapping

Factors that influence fine-mapping

- Linkage Disequilibrium (LD)
- Sample size



Simulation settings:

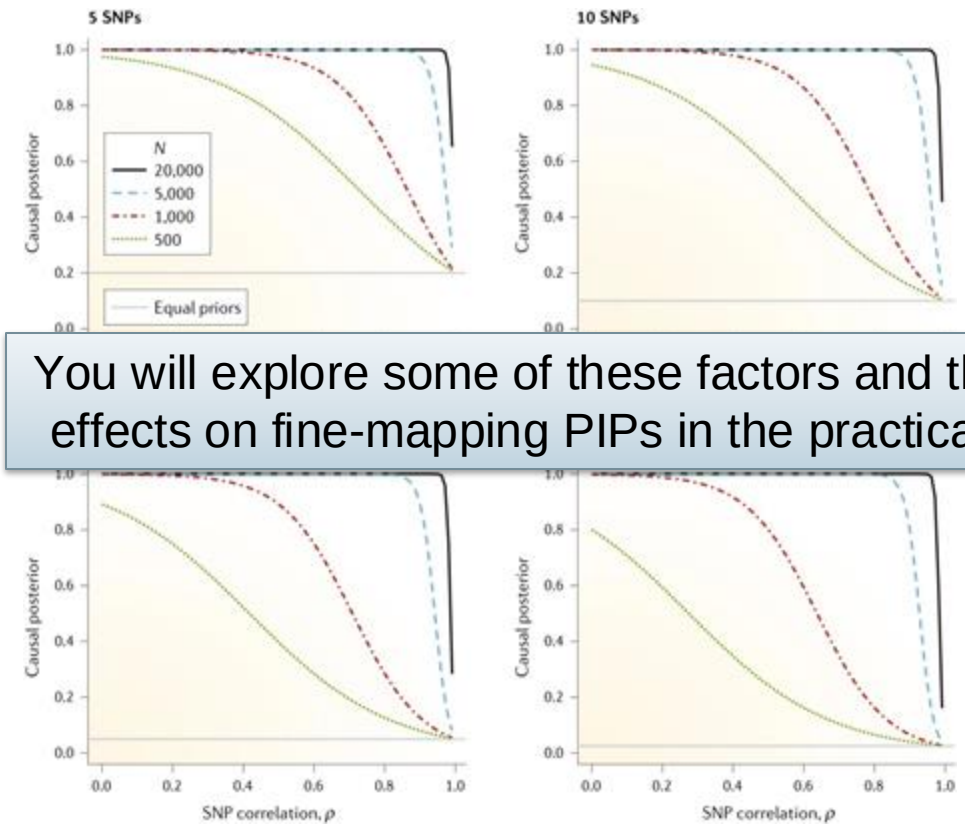
- Single causal SNP;
- All SNPs are correlated with correlation ρ;
- Region contains 5 - 40 SNPs.

Schaid et al. 2018 Nat. Rev. Genet.

Factors that influence fine-mapping

- Linkage Disequilibrium (LD)
- Sample size
- SNP density

- Number of causal variants
- Effect sizes
- Missing causal variants



You will explore some of these factors and their effects on fine-mapping PIPs in the practicals.

Simulation settings:
- Single causal SNP;
- All SNPs are correlated with correlation ρ;
- Region contains 5 - 40 SNPs.

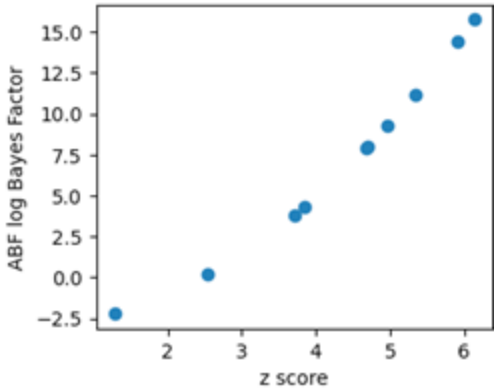
Schaid et al. 2018 Nat. Rev. Genet.

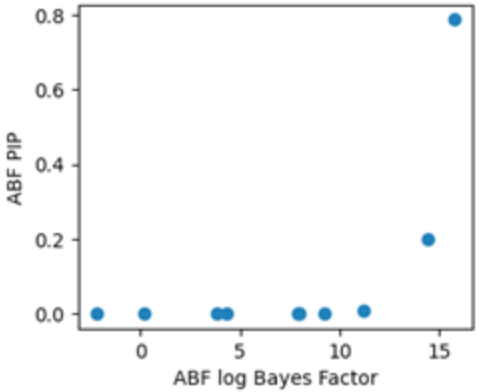
Overview of methods

(Approximate) Bayes Factor method: ABF

$$BF_i = \frac{p(\text{Data} \,|\, \text{SNP } i \text{ is causal})}{p(\text{Data} \,|\, \text{no SNP is causal})} \qquad PIP_i := Pr(\gamma_i = 1 | y, X) = \frac{BF_i}{\sum_k BF_k}$$

In this case, Bayes factors are roughly monotonic transformations of z scores. Highest PIP is almost always given to the most significant variant.



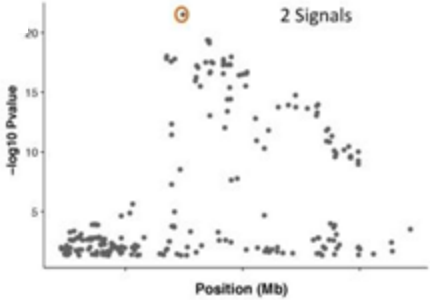


Maller et al. 2012 NatGenet.

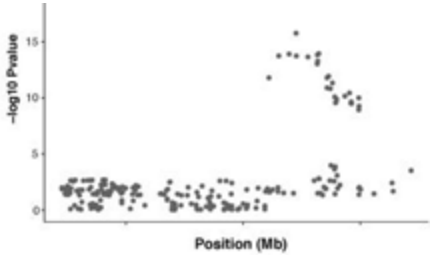
Generalizing to multiple causal SNPs

Conditional analysis:

COJO+ABF



Condition out the lead SNP



Spain et al. 2015 Hum Mol Genet.

Direct modeling of casual configurations:

CAVIAR, CAVIARBF,

FINEMAP, DAP-G

Causal configuration γ :

0	1	0	1	0	0	0	1	0	0

Compute:

$$p(\gamma|y, X)$$

Benner et al. 2016 Bioinformatics

Sum of single effects:

SuSiE

Single effect causal configurations

γ_1 :	0	1	0	0	0	0	0	0	0	0
γ_2 :	0	0	0	1	0	0	0	0	0	0
⋮										
γ_L :	0	0	0	0	0	0	0	1	0	0

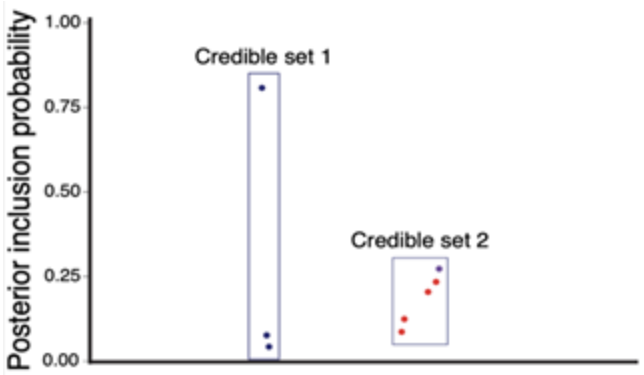
Using Iterative Bayesian Stepwise Selection (IBSS) to compute posterior distributions:

$$p(\gamma_i \,|y, X, \gamma_{-i}, \beta_{-\mathrm{i}})$$

Wang et al. 2020 JRSS

Two ways to define credible sets

1. Aims to capture one causal SNP in each credible set of minimal size and report as many credible sets as the data supports (SuSiE).
2. Aims to find the smallest set of SNPs that contain all the causal SNPs. (ABF, CAVIAR, FINEMAP etc.)



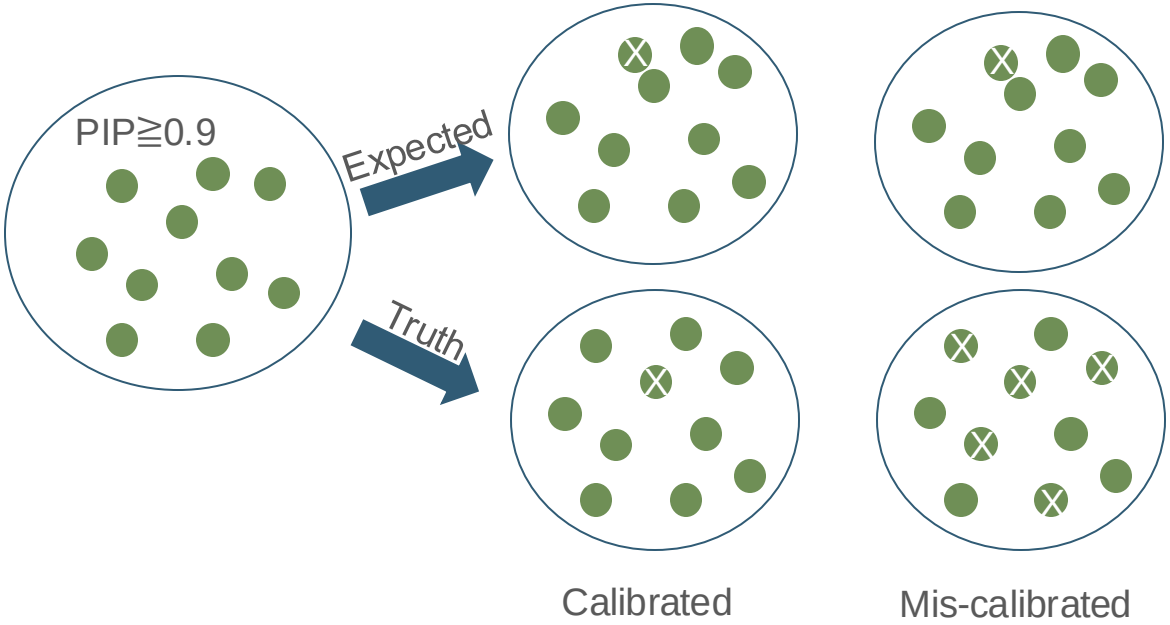
The two definitions coincide when assuming single causal variant per locus

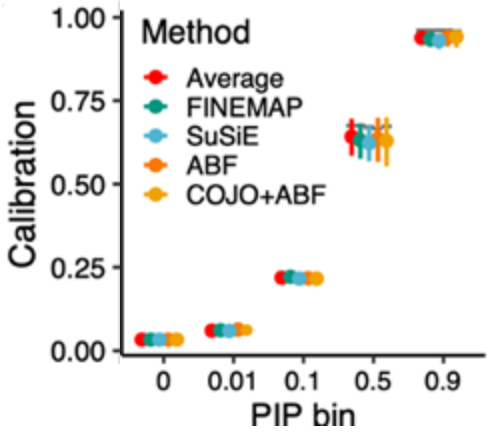
Benchmarking

Benchmarking fine-mapping: in simulations

Two main metrics:

- **Calibration**: Of variants with PIP≅x%, are x% truly causal?
- **Recall**: What proportion of all causal variants are captured by the x% variants with highest PIP?



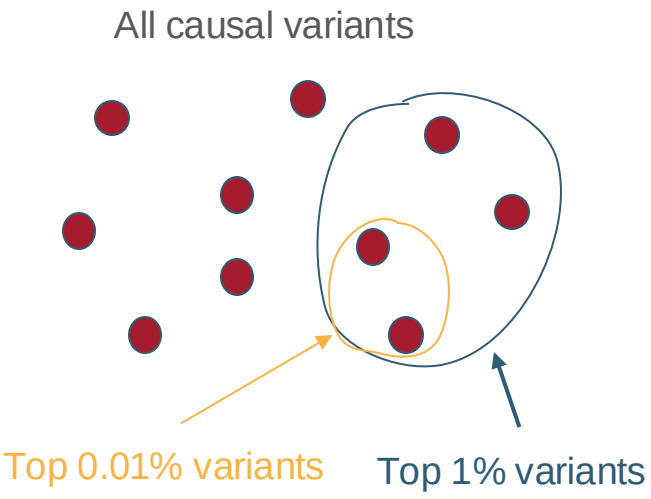


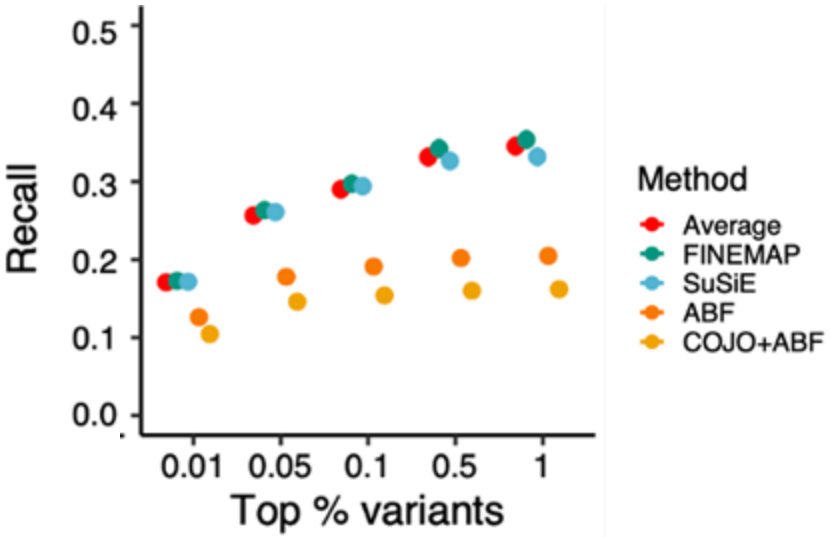
Ulirsch and Kanai et al., in review

Benchmarking fine-mapping: in simulations

Two main metrics:

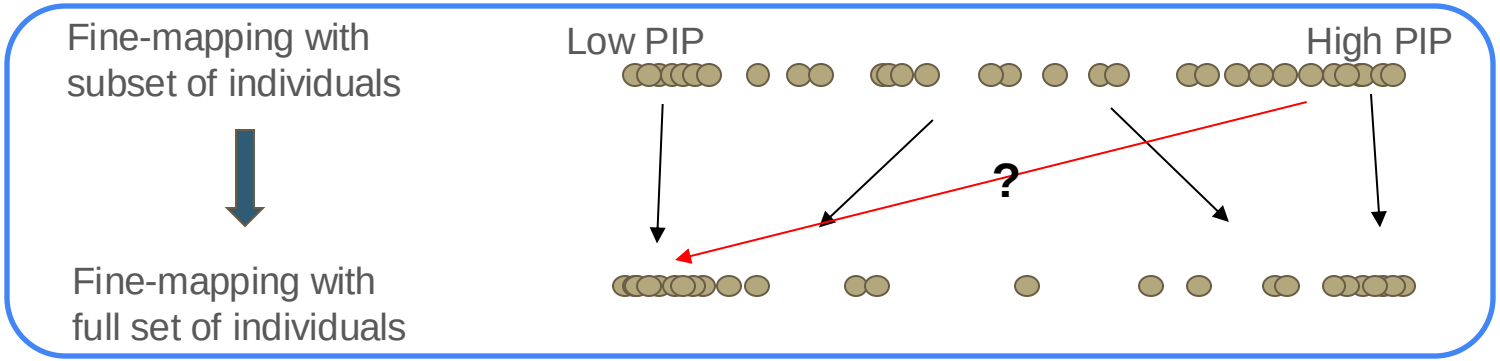
- **Calibration**: Of variants with PIP=x%, are x% truly causal?
- **Recall**: What proportion of all causal variants are captured by the x% variants with highest PIP?





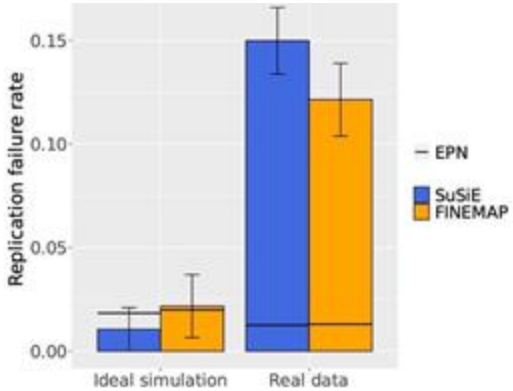
Ulirsch and Kanai et al., in review

Benchmarking fine-mapping: in real data



Replication failure: when a variant's PIP drops from high (>0.9) to low (<0.1) when sample size increases.

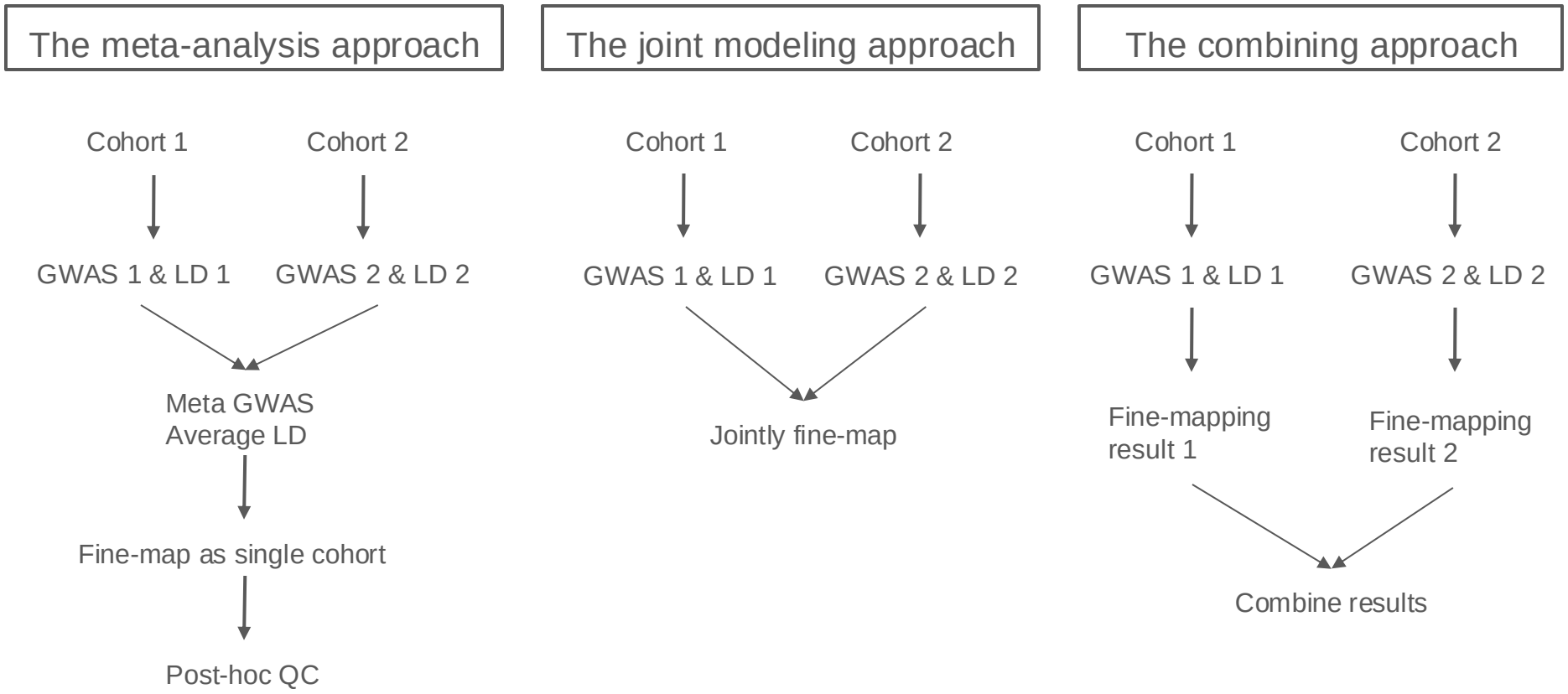
RFR (Replication Failure Rate): the proportion of replication failures.



Cui et al. 2023 NatGenet.

Multi-cohort fine-mapping

Multi-cohort fine-mapping



Resources

Fine-mapping resources

Published fine-mapping results by large studies:

- FinnGen: https://finngen.gitbook.io/documentation/methods/finemapping
- UK Biobank: https://www.finucanelab.org/data
- PGC Schizophrenia study: Supplementary Table 11

LD resources:

- PolyFun published UK Biobank LD matrices.
- Pan-UKBB published multi ancestry LD matrices.

Fine-mapping pipelines:

- FinnGen pipeline: https://github.com/FINNGEN/finemapping-pipeline
- UK Biobank pipeline: https://github.com/mkanai/finemapping-pipeline

Derivations:

- ABF paper, CAVIARBF paper, Schaid et al. NatRev.

Use caution when applying fine-mapping

- Beware of reference LD, follow Weissbrod et al. 2020 NG guidelines.
- Meta-analysis fine-mapping is tricky, see Kanai et al. 2022 Cell Genomics.
- Don’t forget to use covariate-adjusted LD when the cohort has more complex population structure, e.g. admixture. See Pan-UKBB LD documents.
- Keep in mind that model misspecifications and missing causal variants exist in real data applications. Use caution when interpreting fine-mapping results.
- Run different methods if you can.

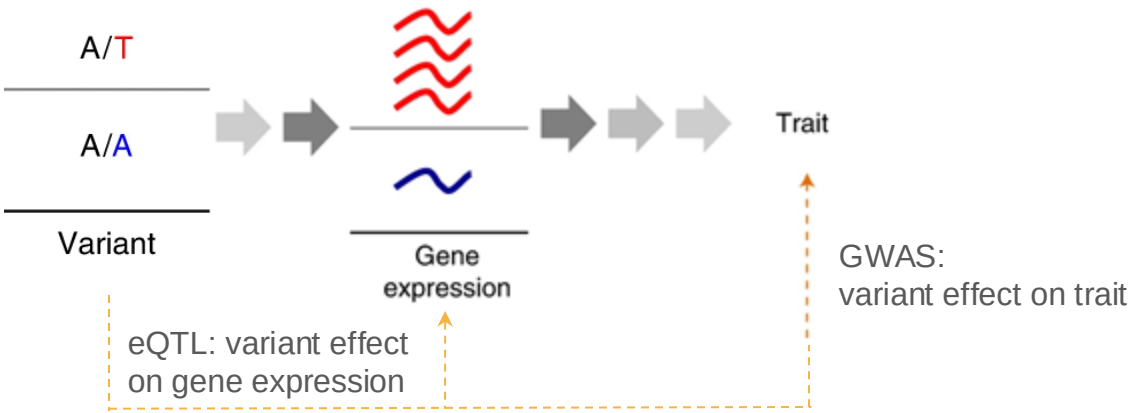
Part 2: Colocalization

The goal of colocalization

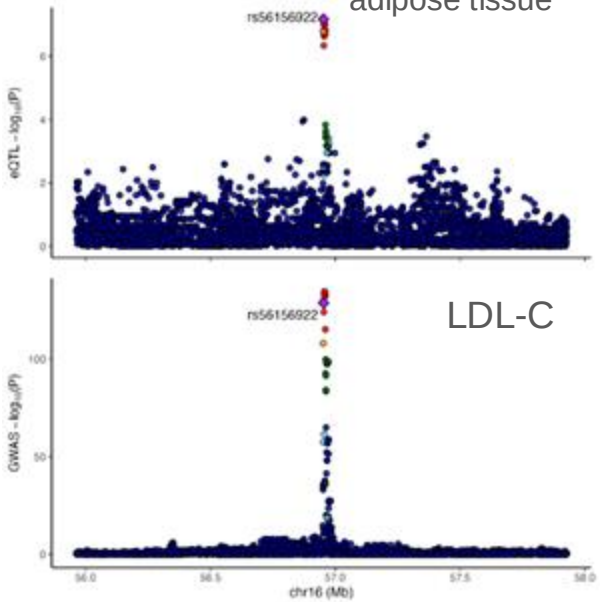
High-level goal: Prioritize gene and cell type targets for functional follow-ups.

Some assumptions are made here.

Specific goal: To test if two associations at a locus share the <u>same causal variant</u>.



CETP locus in adipose tissue



Modified from Baca et al. 2022 NatGenet

Hakim et al. 2025 Genes

Colocalization is closely related to fine-mapping

$$Pr(\gamma_i = 1, d_i = 1 | y, X, y_{eQTL}, X_{eQTL})$$

SNP i is causal for trait

SNP i is causal for expression

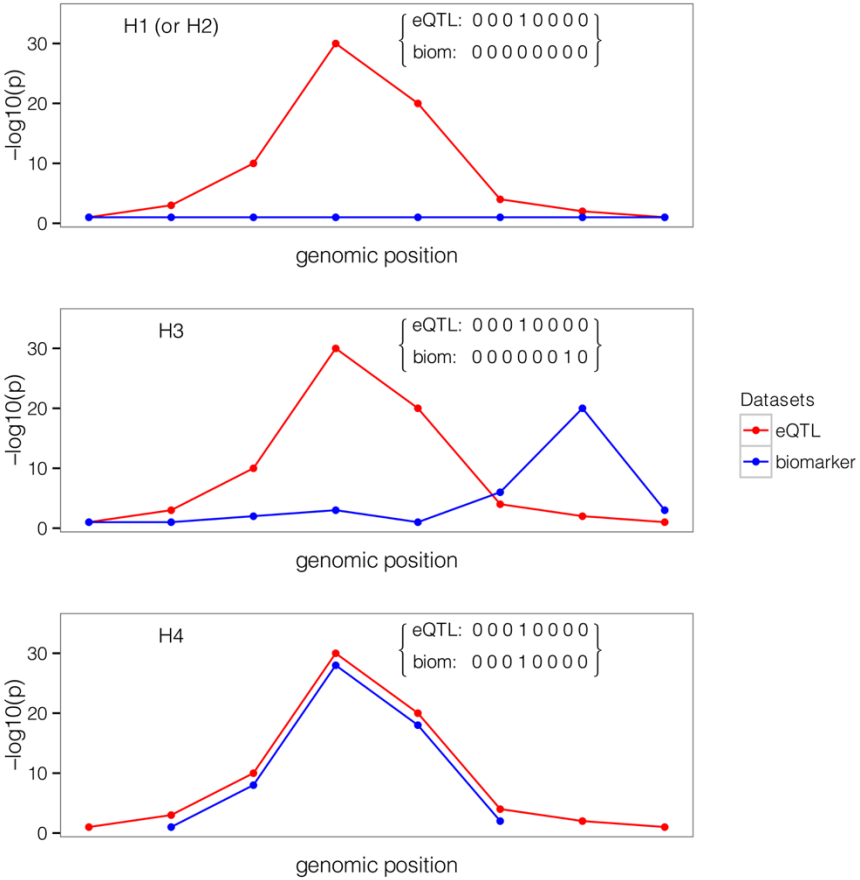
Data from GWAS analysis

Data from eQTL analysis

$$Pr(\gamma_i = 1 | y, X) \ \& \ Pr(d_i = 1 | y_{eQTL}, X_{eQTL})$$

Colocalization methods and outputs

coloc outputs



coloc considers 5 hypotheses:

- H_0: No association with either trait.
- H_1: Association with trait 1, none for trait 2.
- H_2: Association with trait 2, none for trait 1.
- H_3: Different SNPs associated with trait 1 and 2.
- H_4: Same SNP associated with trait 1 and 2.

coloc outputs posterior probabilities for all H_i:

$$p(H_0|Data), p(H_1|Data), \cdots, p(H_4|Data)$$

Giambartolomei et al. 2014 PLOS Genetics

Colocalization methods and their outputs

- coloc
  - coloc.abf: Giambartolomei et al. 2013 PLoS Genet. ⇨ Posterior probabilities of 5 hypotheses
  - coloc.susie: Wallace et al. 2021 PLoS Genet.
- eCAVIAR
  - Hormozdiari et al. 2016 AJHG ⇨ CLPP for each variant
- enloc/fastENLOC
  - Wen et al. 2017 PLoS Genet. ⇨ SNP and regional level colocalization probabilities (SCP and RCP)
  - Pividori et al. 2020 Sci. Adv.
  - Hukku et al. 2022 AJHG

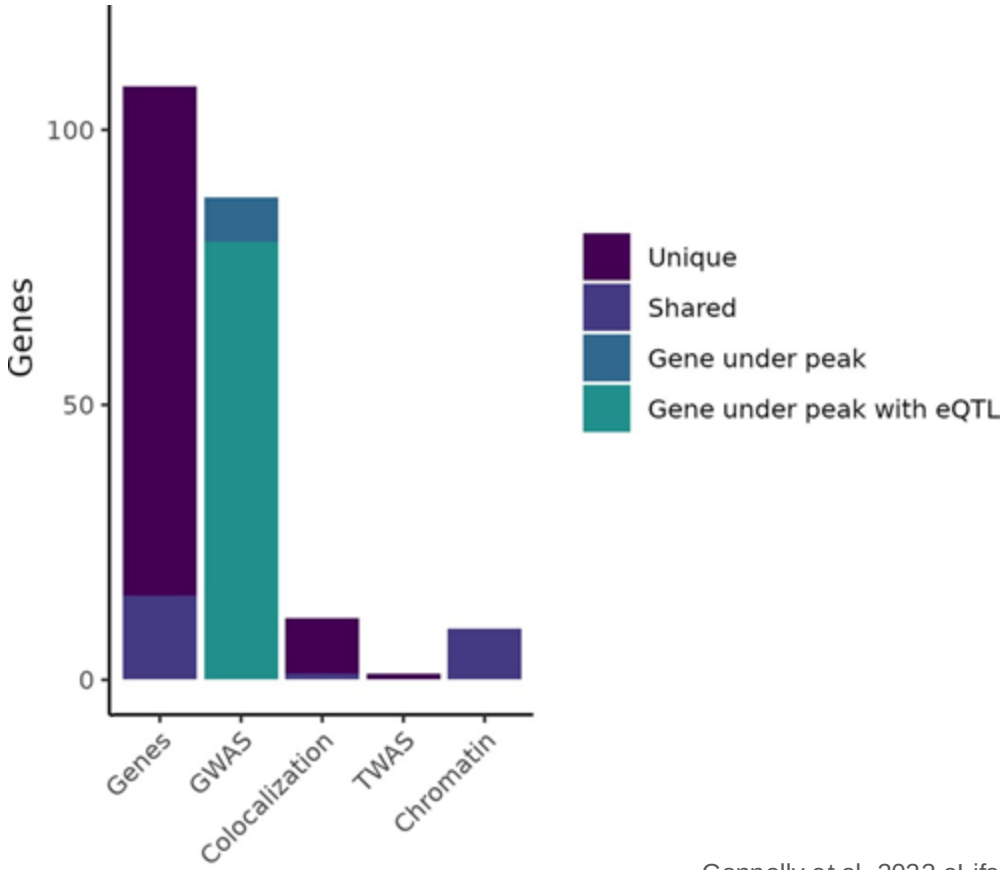
How well does colocalization work

It doesn’t work as well as we had hoped

We have reasons to believe that most trait associations should be eQTLs:

- Trait associated SNPs are more likely to be eQTLs.
- A large fraction of heritability resides in regions with gene regulatory potential.

However, only 5-40% of trait associations colocalize with eQTLs using various methods.



Connally et al. 2023 eLife

Some possible explanations

- Power
- Cell type
- Cell context
- Other molecular phenotypes (sQTL, pQTL etc.)
- Gene by environment interaction (GxE)
- Development
- Fine-mapping inaccuracies
- Re-examine assumptions (Mostafavi et al. 2023 NatGenet)

Resources

Published fine-mapping results by large studies:

- https://gtexportal.org/home
- GTEx fine-mapping results
- twas_hub.org

Visualization tool:

- Locuscomparer

Derivations:

- coloc paper, eCAVIAR paper, enloc paper, TWAS paper

Qualtrics link:
https://qimr.az1.qualtrics.com/jfe/form/SV_5bifqSVk0lrbCd0