Post-GWAS Analysis

Fine-mapping and colocalization

Statistical Genetics Workshop

NIH NATIONAL CANCER INSTITUTE

Sept 27th, 2023

Agenda

1. *What is post GWAS analysis*

2. *Fine mapping*
   - *Single causal variant*
   - *Multiple causal variants*
   - *Incorporating genomic annotations*
   - *Using summary statistics*

3. *Colocalization*
   - *Differences with fine mapping*
   - *Methods and Interpretation*

Genome-wide association studies

March 2023: GWAS catalog

GWAS: Identifies associations

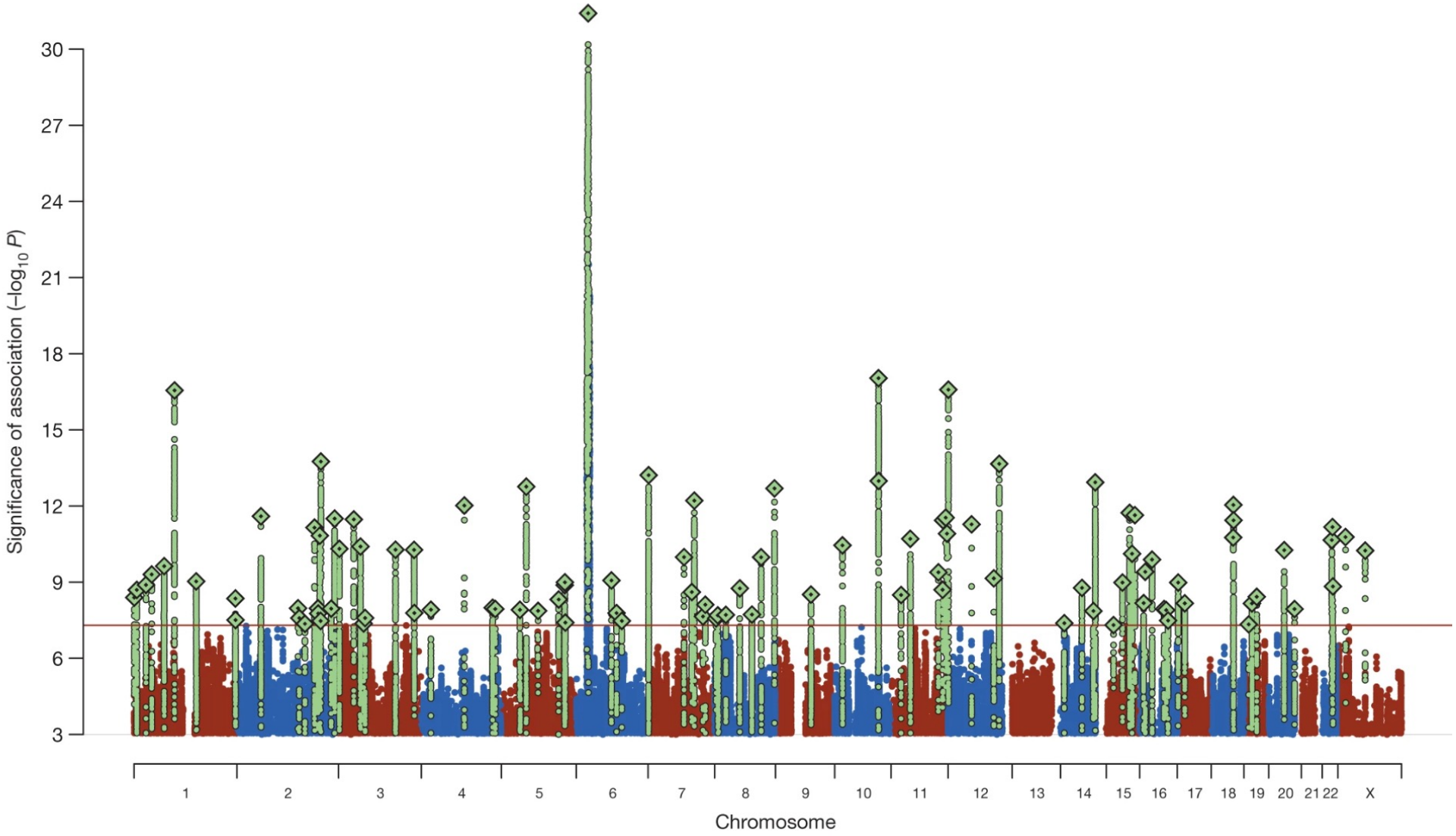


Image from: Schizophrenia working group (2014); *Nat. Gen.*

Many questions remain unanswered

- GWAS **identifies associations** not true causality
- Need to tease apart "**true signals**" from signals arising due to high correlation with the true signals
- Cannot tell us how the variants affect the outcome

Statistical Causality vs Biological causality

- A variant is **statistically causal** for the outcome if the true underlying statistical effect of the SNP on the outcome is non-zero

- A SNP is **biologically causal** for the outcome if the SNP influences phenotypic variation (or case control status) through a biological mechanism

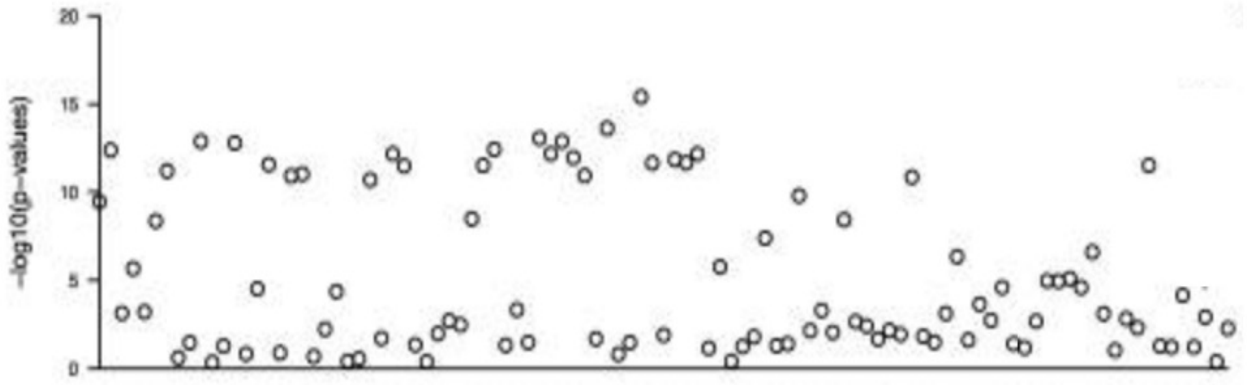
- Here we only look at statistical causality

- The purpose of *in silico* analyses and identifying statistical causality is to propose a smaller number of candidate variants to be investigated downstream.

Fine mapping: Why is it important?

- Find candidate causal variants
  - causal regulatory and coding variations
- Pinpoint regulation and mechanisms
  - Suggest perturbation and knock out experiments
- Understand genetic architecture
  - Enrichment of genetic features
  - Similarities and differences across diseases or ancestry-groups
  - Prediction

Looking in-depth at a locus



Locations of variants on a chromosome

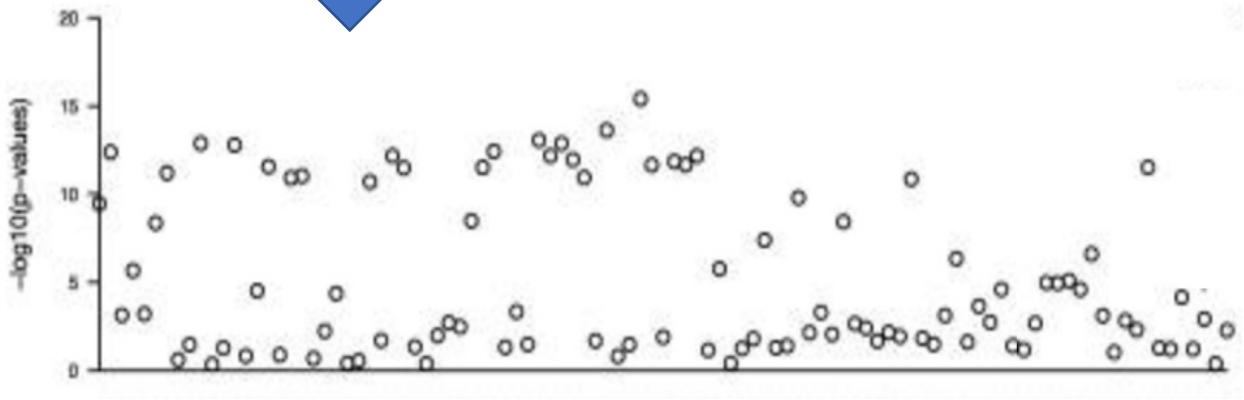
- How does this GWAS association pattern arise
- Hypothesis: simple underlying causal structure
- Which SNP(s) is(are) causal?

Underlying “Statistically” causal SNPs

True Causal Status

- Correlation between SNPs (LD)
- Sampling noise/error

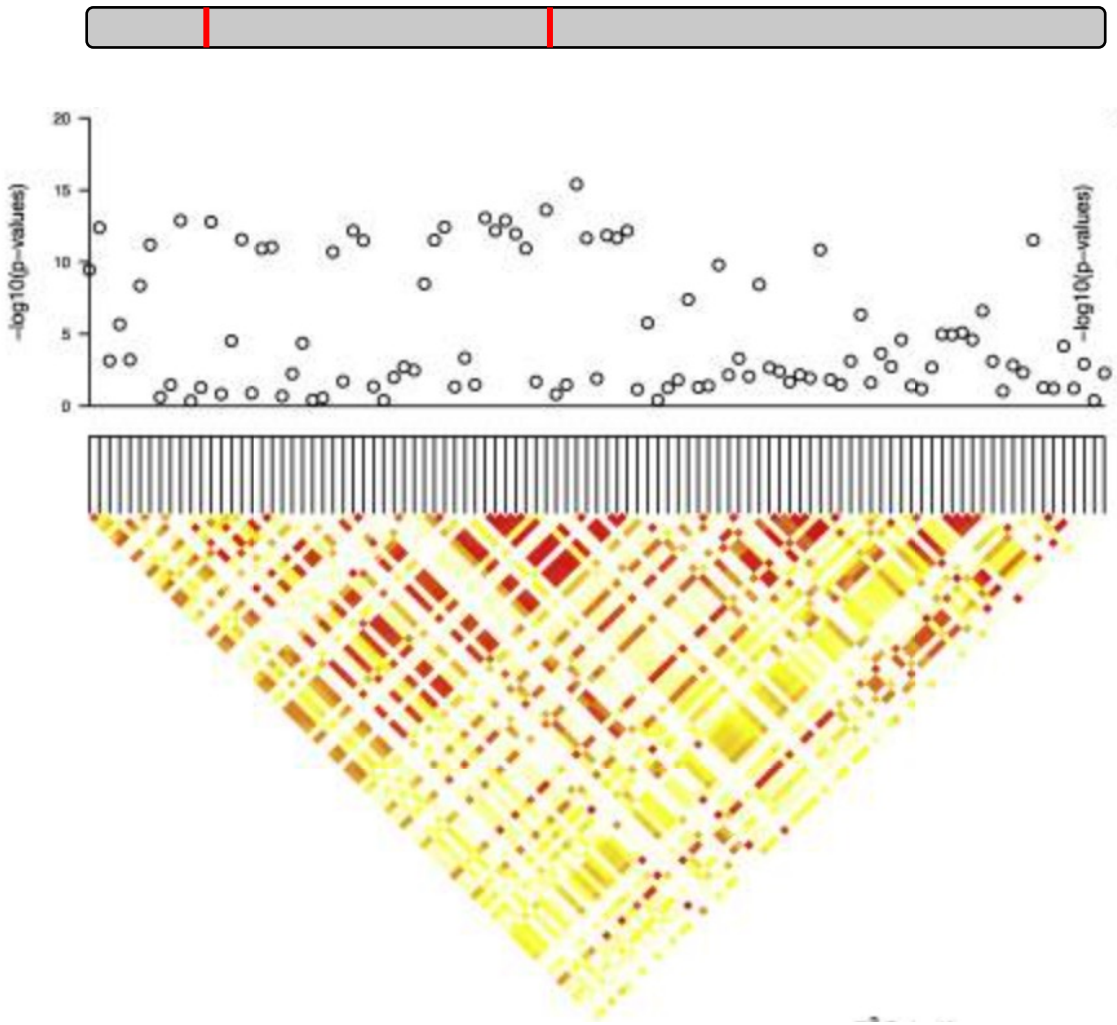
Observed association patterns in GWAS



Locations of variants on a chromosome

Influence of LD

True Causal Status



SNPs in high LD (correlated) with the true causal variant(s) can potentially be significant marginally

Often the SNP with the lowest p-value in a locus (called sentinel/index SNP) is in high LD (correlated) with the true causal SNP(s)

Sentinel SNPs might not be causal but “tags” the true causal SNP(s)

The true causal SNP(s) is possibly in a set of SNPs that are in high LD with the sentinel SNP

Significant SNP vs causal SNP

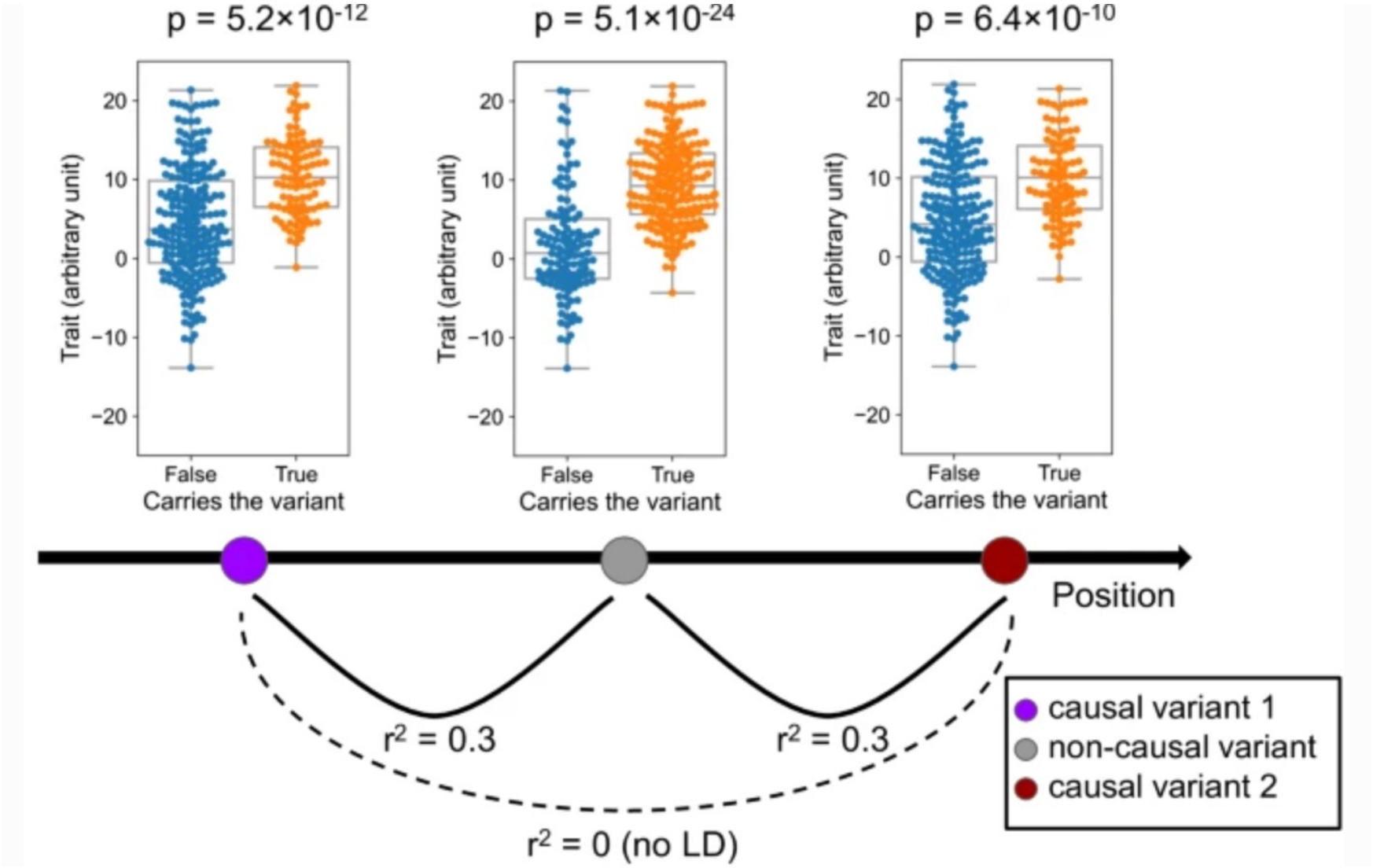


Image from: Wang et al (2022); *Semin Immunopathol*

Significant SNP vs causal SNP

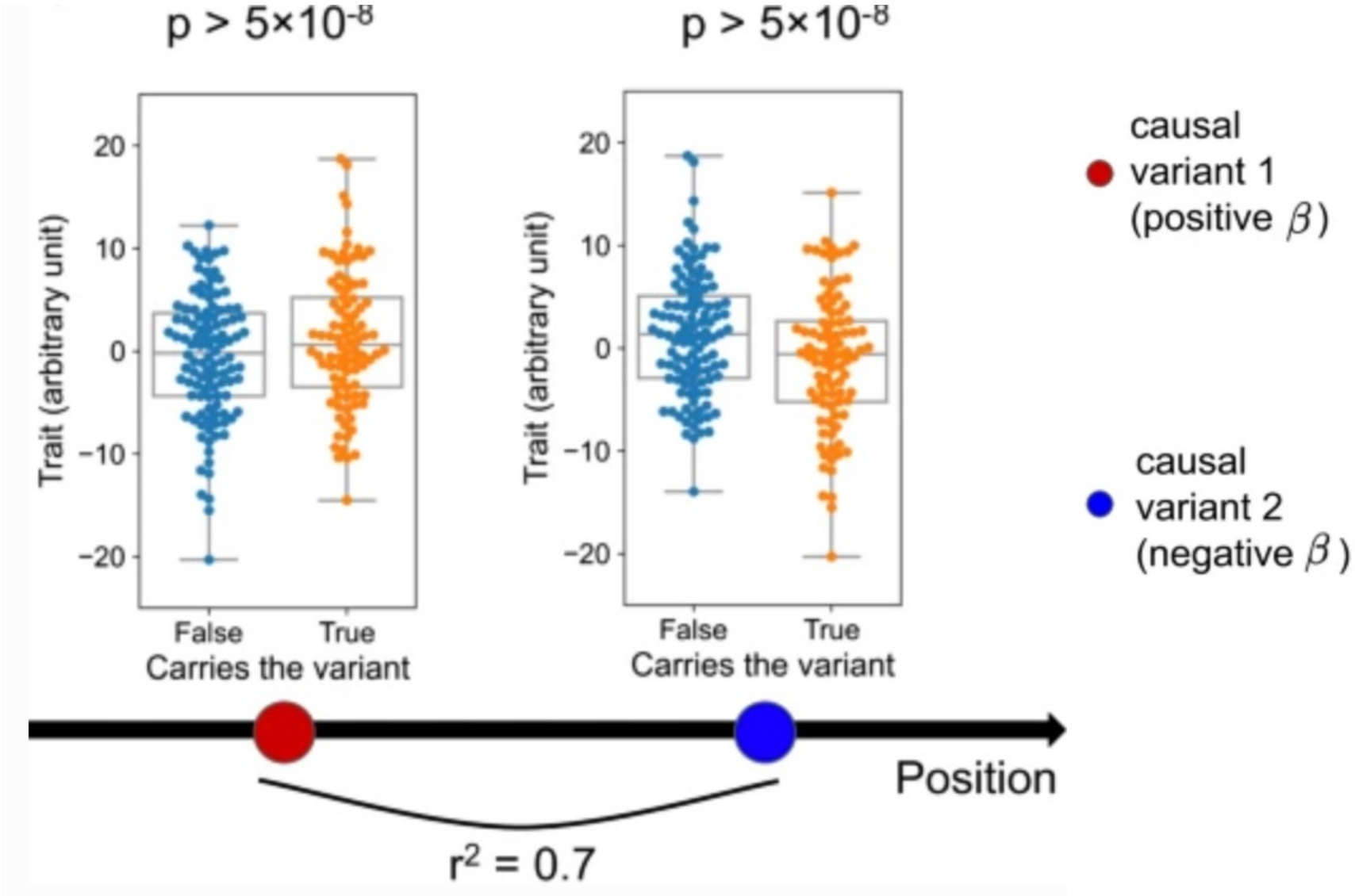
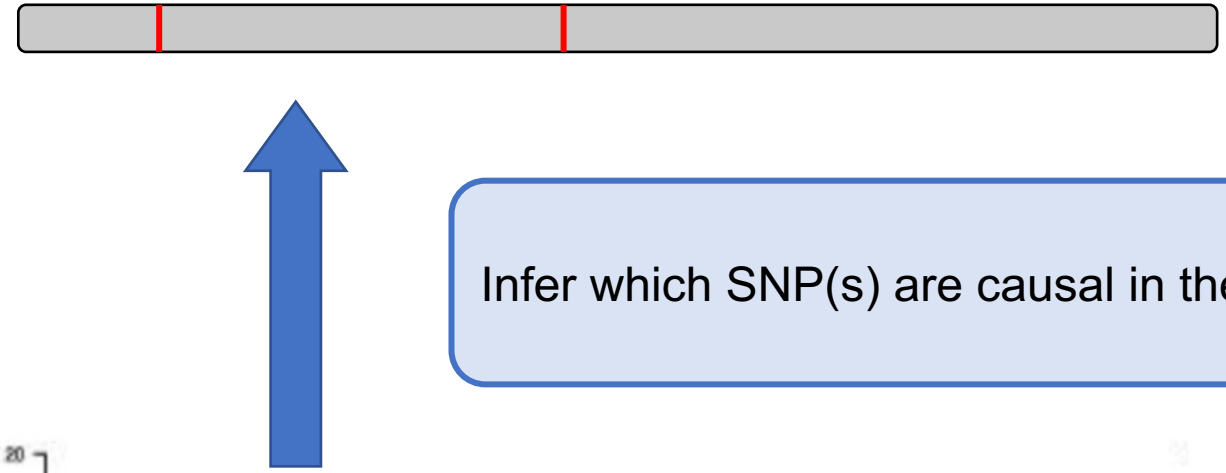


Image from: Wang et al (2022); *Semin Immunopathol*

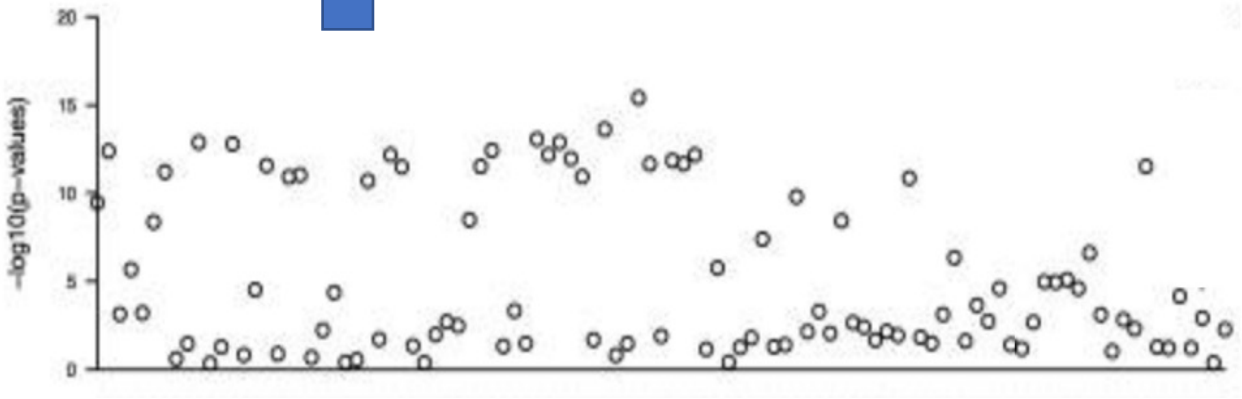
Goal of Fine mapping

True Causal Status



Infer which SNP(s) are causal in the locus

Observed association patterns



Locations of variants on a chromosome

Fine mapping: Topics for today

- **What we want to calculate**

- **How to calculate**: Methods
  - Single causal variant
  - Multiple causal variant
  - Integrating Functional data
  - Using GWAS summary statistics vs individual level data

- **What are the outputs**

What we want to get

- Probability that a variant is causal

- A small set of variants that contains a causal variant with a given probability

Fine mapping: Single causal variant

Simple idea for fine-mapping

Suppose we have **individual level data** for n individuals (genotype and phenotype)

$$Y = \alpha + G_1\beta_1 + G_2\beta_2 + \cdots + G_p\beta_p + \boldsymbol{\varepsilon}$$

where α: intercept (might include non genetic covariates)

G_i: genotype/dosage for ith variant in the locus across n individuals

β_i: **joint effect size** of the ith variant

$\boldsymbol{\varepsilon}$: gaussian error term

Solution: $$\beta = (\mathrm{G}'\mathrm{G})^{-1}G'y$$

Takes care of the correlation between the variants

Choose the SNPs that have corresponding high absolute value of β

$$\beta = (\mathrm{G}'\mathrm{G})^{-1} G'y$$

Dependency between variants

Estimated using LD

In-sample LD or LD estimated from a reference panel

Interaction between outcome and SNPs

Can be estimated from GWAS summary statistics: effect size and SE

No direct probabilistic interpretation on the causal status of the SNPs

Numerical problems in calculations

Alternative solution: Bayesian Inference

What we want to estimate in Bayesian Methods

- **Posterior Inclusion probability (PIP)**: Probability that a variant is causal (given a particular causal configuration)
  - Other names like PPA etc. are also used sometimes but PIP is the most common.

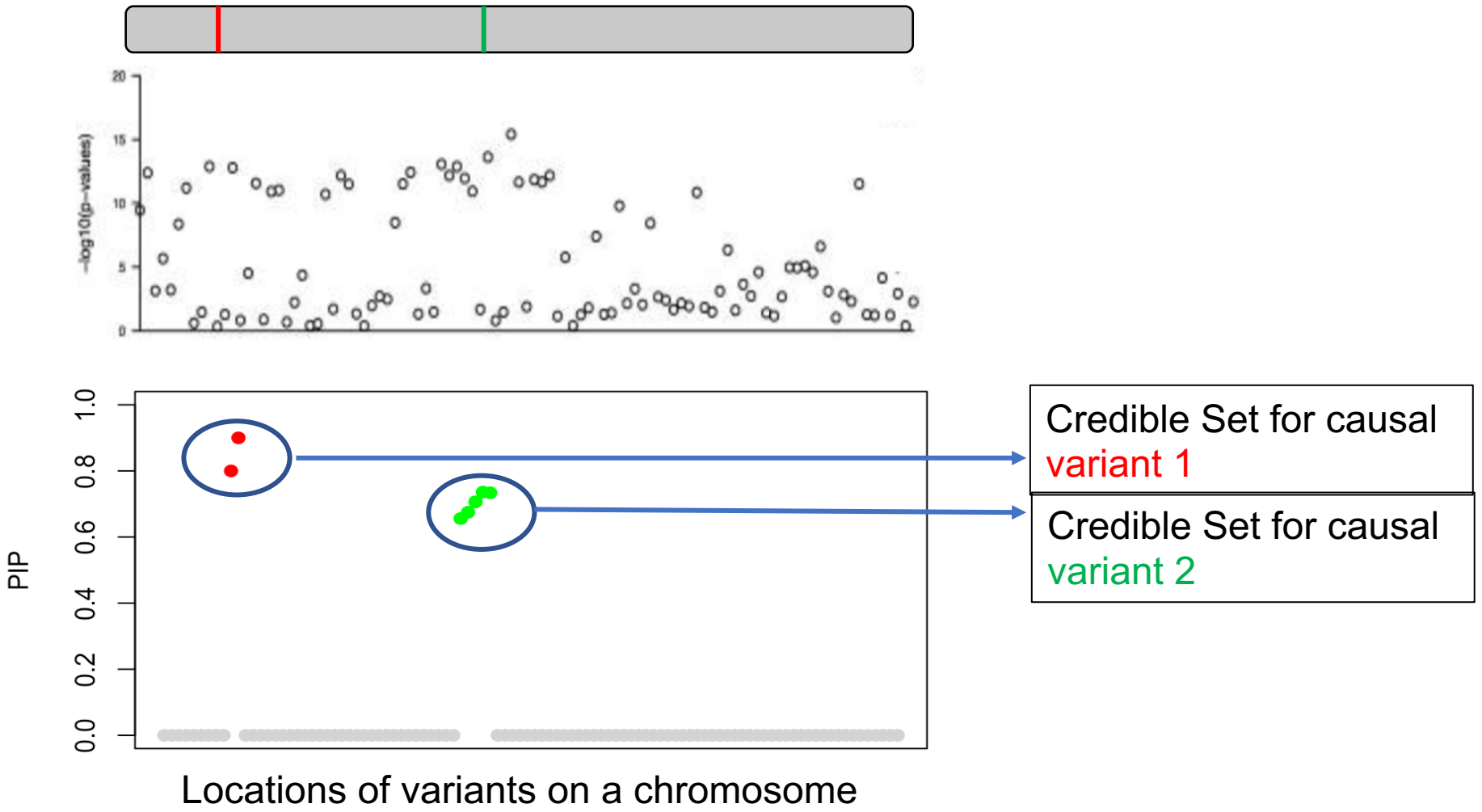
- **Credible Set (CS)**: smallest set of variants that contains a causal variant with a given probability
  - Typically, we look at 95% or 99% credible sets
  - Analogous to confidence intervals (in frequentist hypothesis testing)

Output of Fine mapping

True Causal Status

Observed association patterns

Inferred posterior probability



How to compute these

- $\text{PIP}_j = P\left(j^{th}\ SNP\ is\ causal \mid data\ \right)$

- Bayes factor for SNP S
  - **information/evidence in favor of the hypothesis that SNP S is causal** against the null hypothesis that no SNP is causal

- Under single causal variant assumption (Wakefield 2009)

$$\text{ABF} = \sqrt{\frac{V+W}{V}}\exp\left(-\frac{z^2}{2}\frac{W}{(V+W)}\right)$$

- Z: GWAS t-statistics, V: estimated GWAS SE and W: prior variance on effect size

- Can be computed directly from GWAS summary statistics and **does not require LD information**

How to compute PIP and CS

PIP: Calculate bayes factors and compute the PIP for each variant as scaled bayes factors

$$PIP_j = \frac{BF_j}{\sum_k BF_k}$$

Credible set: S is a 95% CS if Prob(causal variant in S) ≥ 0.95

To construct **smallest 95% CS**: Order PIPs in decreasing order and cumulatively add them until the sum of PIPs is ≥ 0.95

Example

SNP	PVALUE	PIP
rs2912781	9.39E-22	0.52705
rs10736303	1.96E-21	0.24881
rs7895676	2.79E-21	0.17922
rs12356902	8.10E-05	0.02005
rs146141333	2.07E-03	0.00675
rs2912778	7.96E-21	0.00495

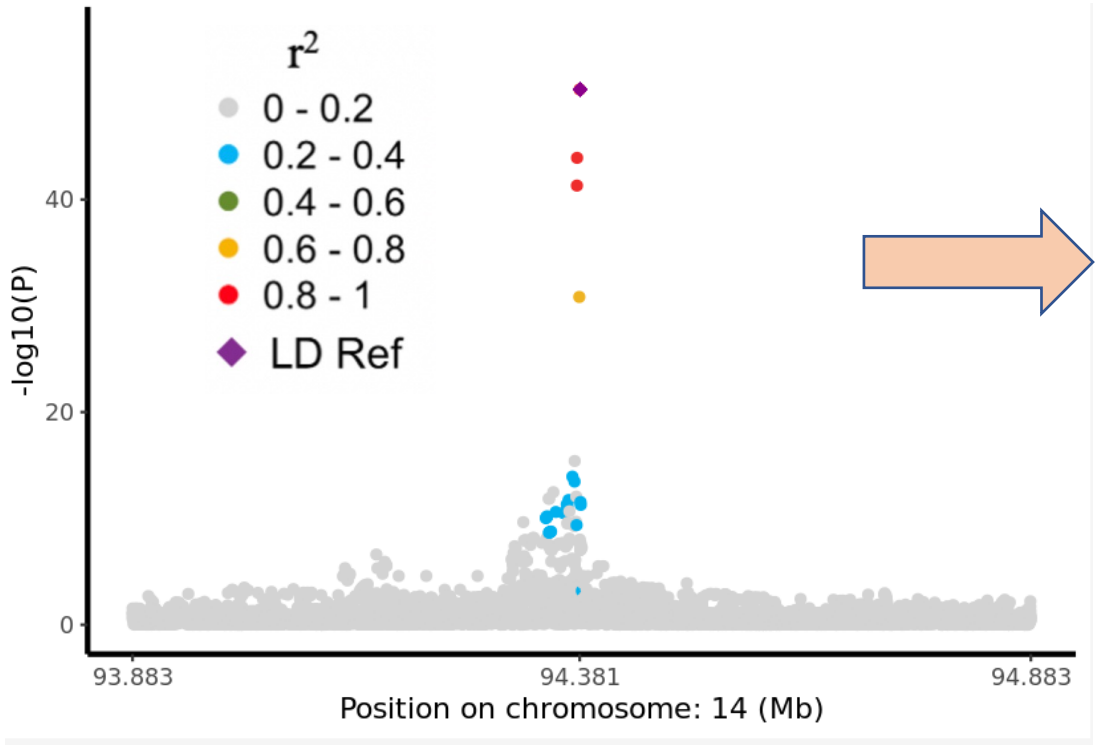
SNPs ordered by PIP

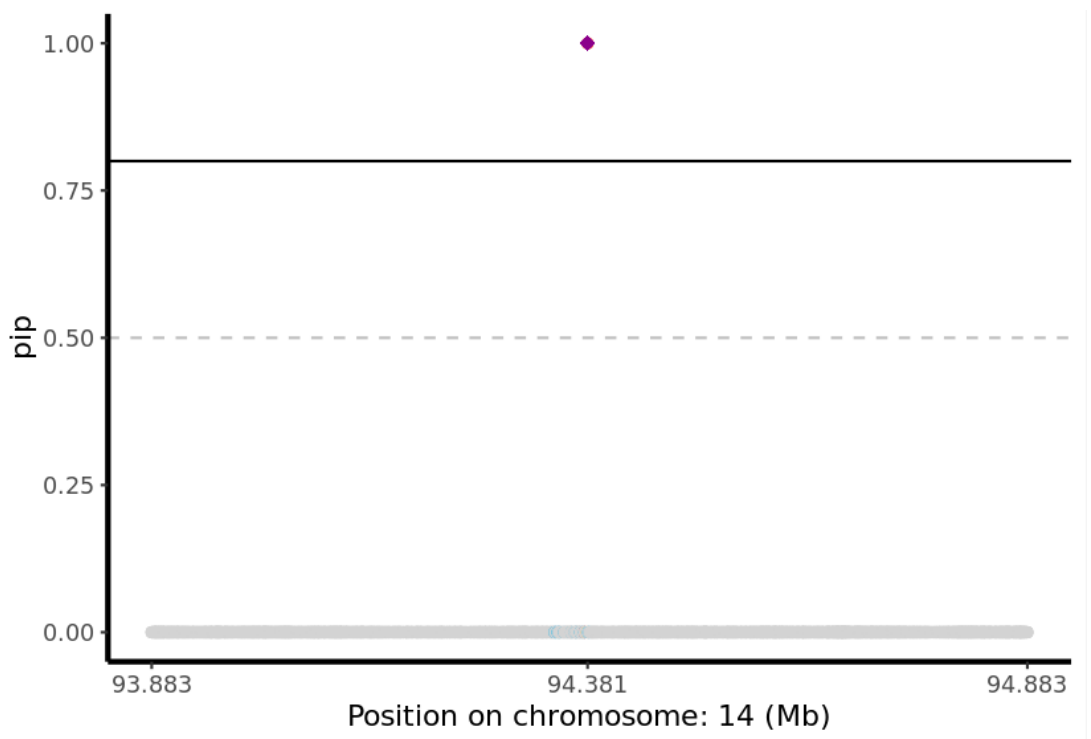
SNP	PVALUE	PIP	
rs2912781	9.39E-22	0.52705	0.52705
rs10736303	1.96E-21	0.24881	0.77586
rs7895676	2.79E-21	0.17922	0.95508
rs12356902	8.10E-05	0.02005	0.97513
rs146141333	2.07E-03	0.00675	0.98188
rs2912778	7.96E-21	0.00495	0.98683

95% Credible set contains 3 SNPs

Example: Singleton CS

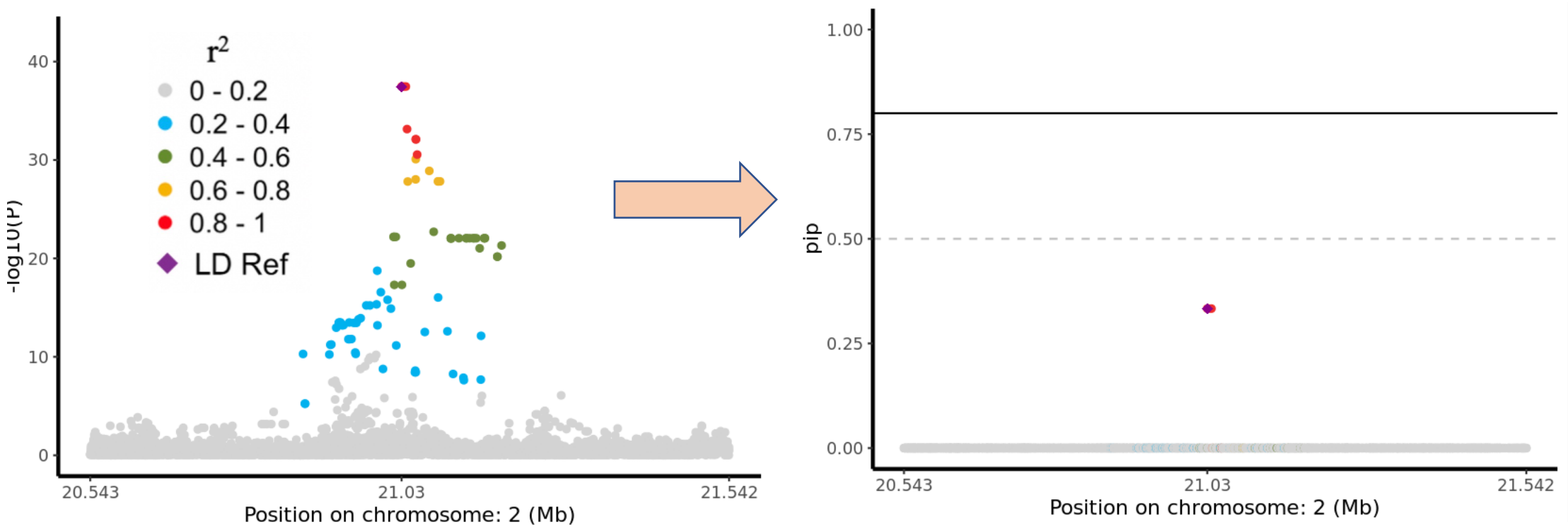
Clear difference in GWAS association even among high LD SNPs



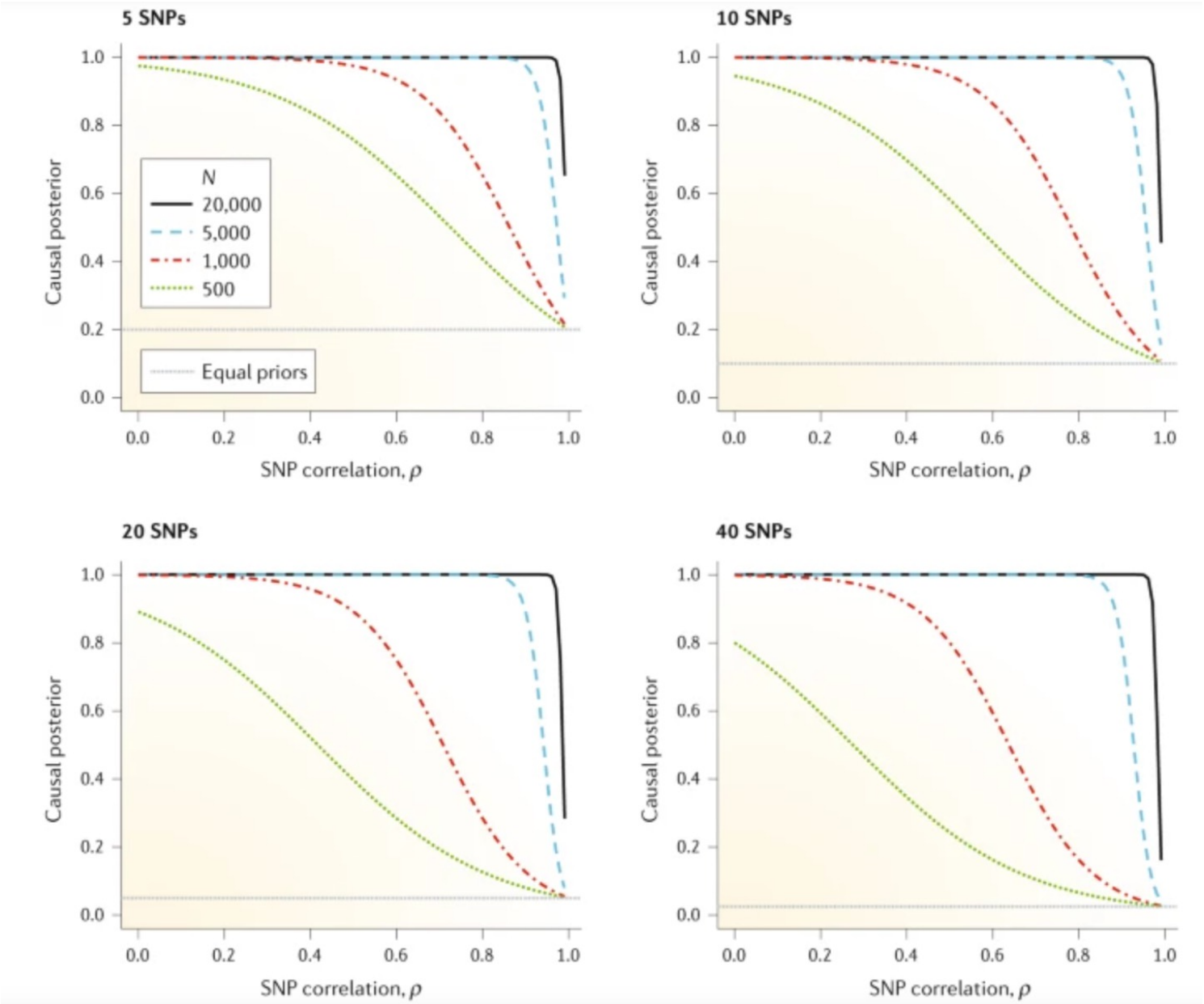


Example: Non-singleton CS

Minor differences in GWAS association among high LD SNPs



Factors influencing fine mapping



- LD structure in the region
  - Many SNPs in high LD with causal variants PIP would be almost same for all the variants and CS would contain many SNPs
- Sample size of initial GWAS
  - The study needs to be well powered to be informative
- Effect size
  - High effect size causal effects are easier to distinguish

Image from: Schaid et al (2018); *Nat. Rev. Genetics*

Fine mapping: Multiple causal variants

Multiple causal variants

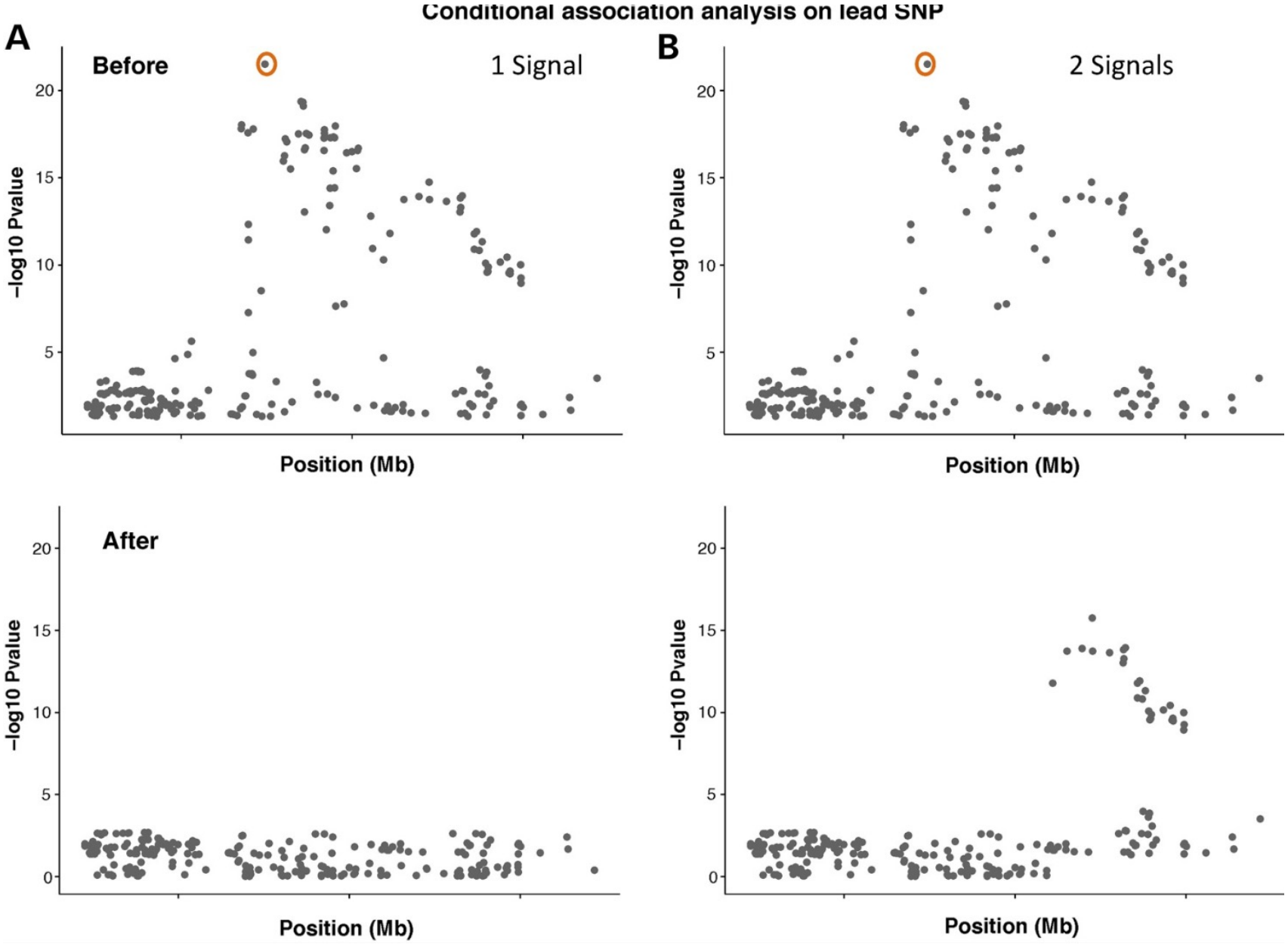
- **Single causal variant might not be optimal** model for many scenarios especially for molecular phenotypes

- Multiple causal variants methods developed to incorporate a **wider range of causal scenarios** (might not be exhaustive)

- How to determine if the association signal at a locus is driven by a single causal variant or multiple causal variant?
  - **Conditional Regression methods**

Conditional forward stepwise regression

- **Step 1**: Take the sentinel SNP in the region and initialize the set S_c
- **Step 2**: Rerun the GWAS analysis with the sentinel SNP as an additional covariate
  - If SNP_x is significant then, $S_c = S_c \cup SNP_x$
  - If no SNP is significant then stop
- Repeat Step 2 until no additional SNPs emerge as significance
- Ideally S_c should give tag SNPs that have independent associations with the outcome



Conditional regression works best if the signals are **separable** and **not many SNPs are in LD with multiple causal variants**

Image from: Spain et al. (2015); *Hum. Mol. Gen.*

Considerations for conditional regression

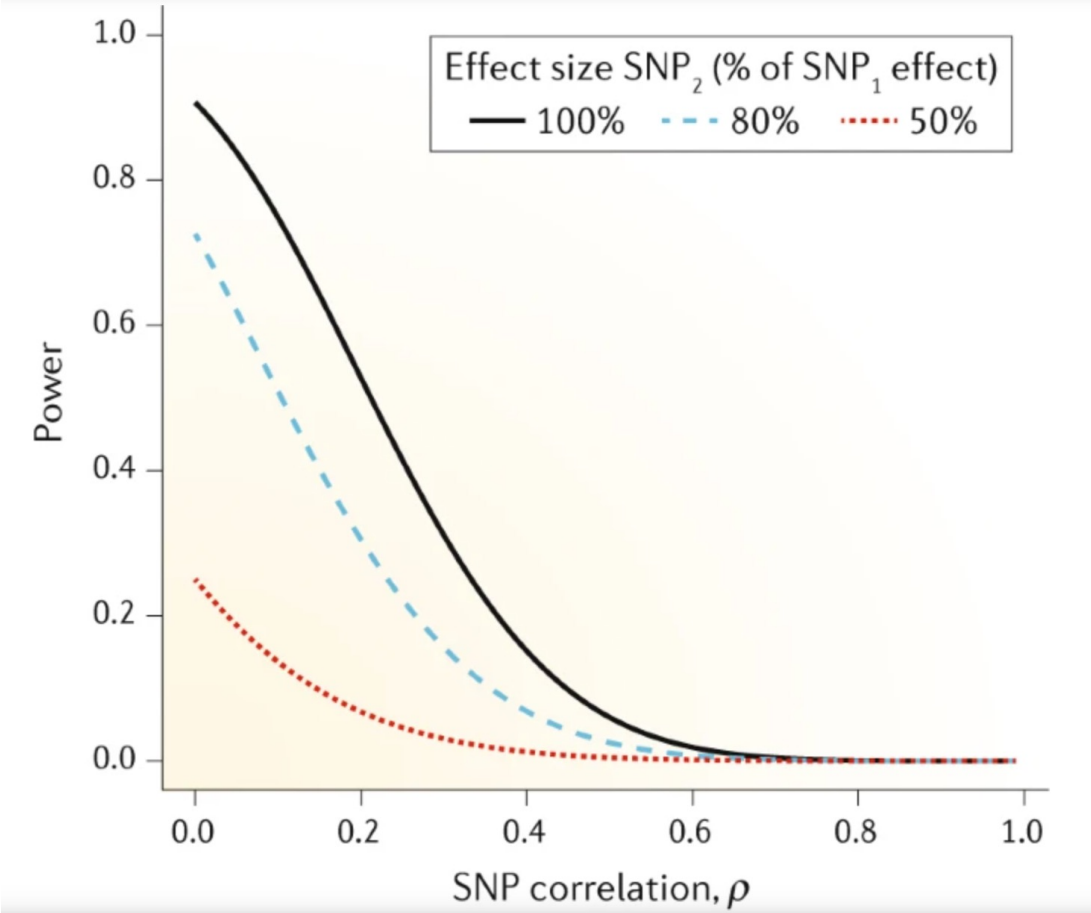


Image from: Schaid et al (2018); *Nat. Rev. Genetics*

Low statistical power with **increasing number of steps** (need to control for sequential testing)

Many SNPs + using lenient cutoffs to include SNPs at each step makes the algorithm is highly unstable.

Simulation (Schaid et al.) with primary signal heritability 1% and sample size such that the power to detect the primary signal is 90%

Very low power with increasing LD with the primary signal

Very low power with weaker secondary signal

Splitting might be difficult

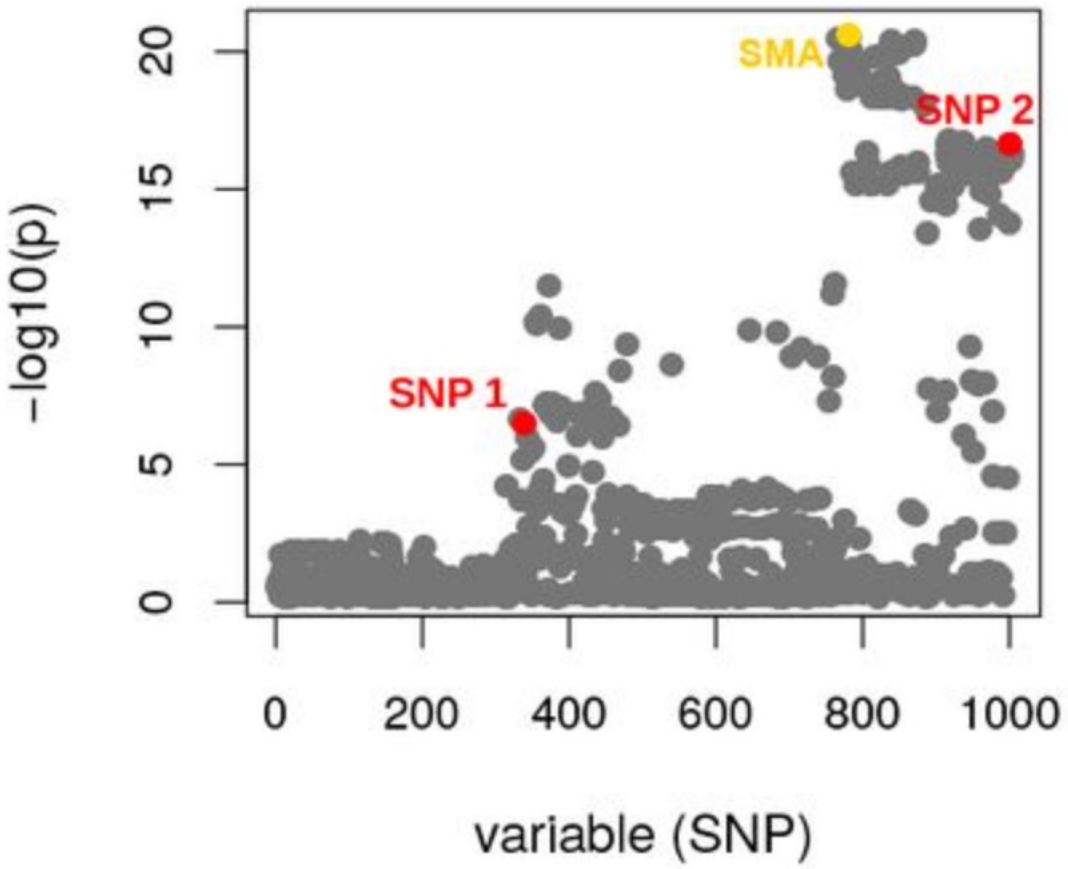


Image from: Wang et al (2019); *JRSS-B*

Complicated LD patterns in the region and between the causal variants

Splitting the region into sub-regions driven by single causal variant might be difficult

Comprehensive and newer methods jointly model the data rather than splitting into sub-regions

Estimating the PIP

Rather than splitting into single causal variant sub-regions, we jointly model all possible causal configurations

In the region/locus there are “m” SNPs

M_c: a binary vector (mx1) indicating a causal scenario

2^m possible configurations

Each variant has **same prior probability** to be causal

Computationally expensive with increasing m

Computational challenges

Calculations become intractable exponentially with increasing number of causal variants

Majority of methods limit the number of causal variants: 3-6

Method	Author	Computational trick
PAINTOR	Kichaev et al (2014)	Maximum likelihood estimation using EM algorithm
CAVIAR	Hormozdiari et al (2014)	Limits the number of causal variants and/or limit locus size
FINEMAP	Benner et al (2016)	Finds most likely configurations through stochastic model searching
DAP-G	Wen et al (2016)	Finds most likely configurations through fast deterministic algorithm
SuSiE	Wang et al (2019)	Variational inference based on iterative Bayesian conditional regression

Fine mapping:

Integrating genomic annotations

Incorporating informative prior

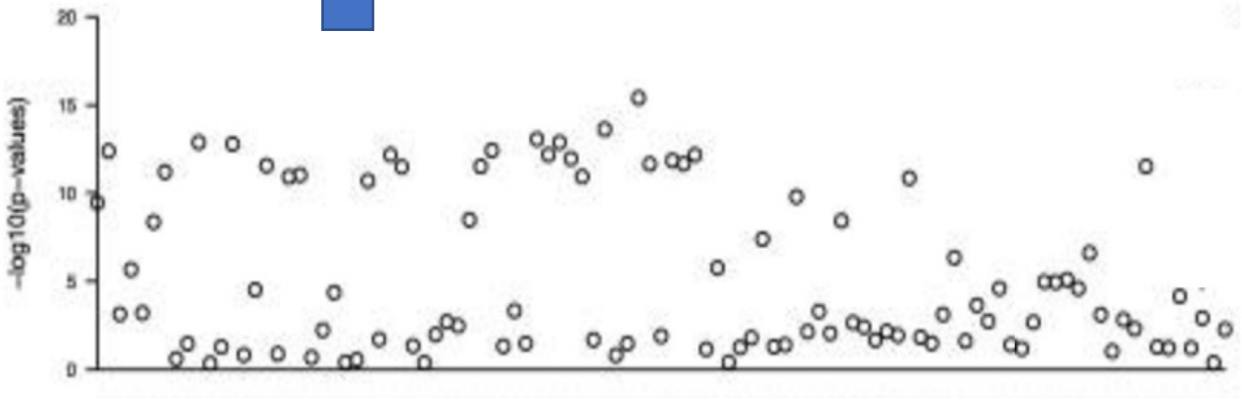
- From genomic studies we know certain types of variants are **more likely to be biologically causal**
- Higher chances of variants with certain **genomic features** to be causal
- **Adjust the prior** to incorporate such genomic features of annotation
- These are referred to ordinarily as **“functional” prior** (priors from functional annotations)

Goal of Fine mapping with Genomic Annotations

True Causal Status

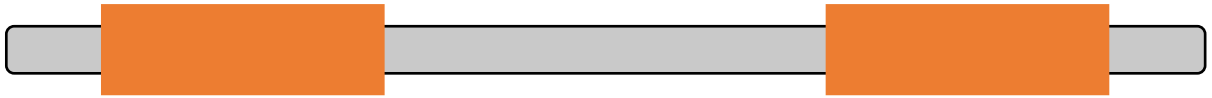
Infer which SNP(s) are causal in the locus

Observed association patterns



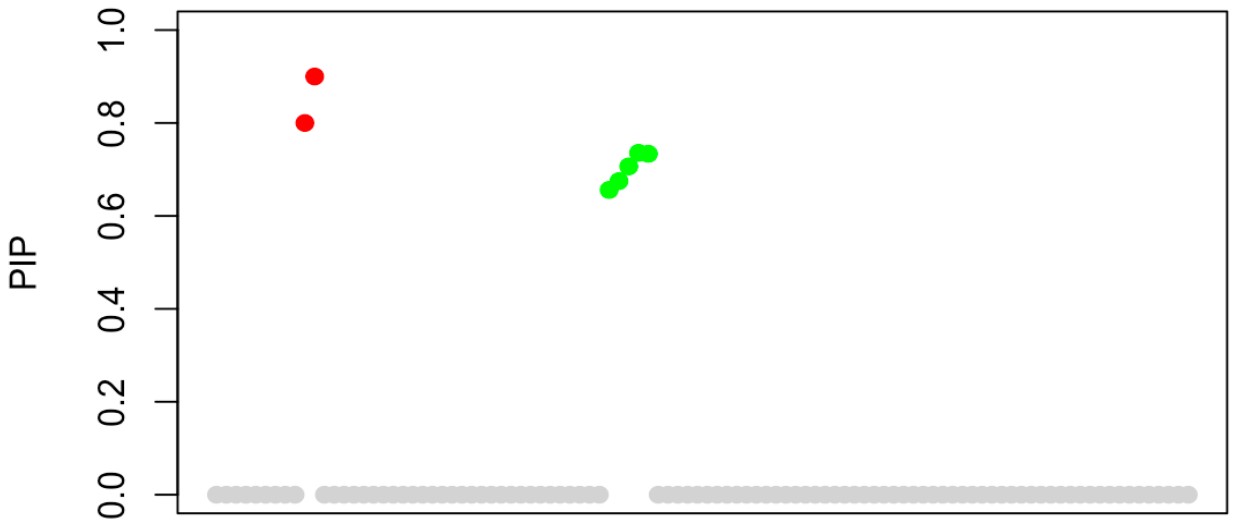
Locations of variants on a chromosome

Genomic Features



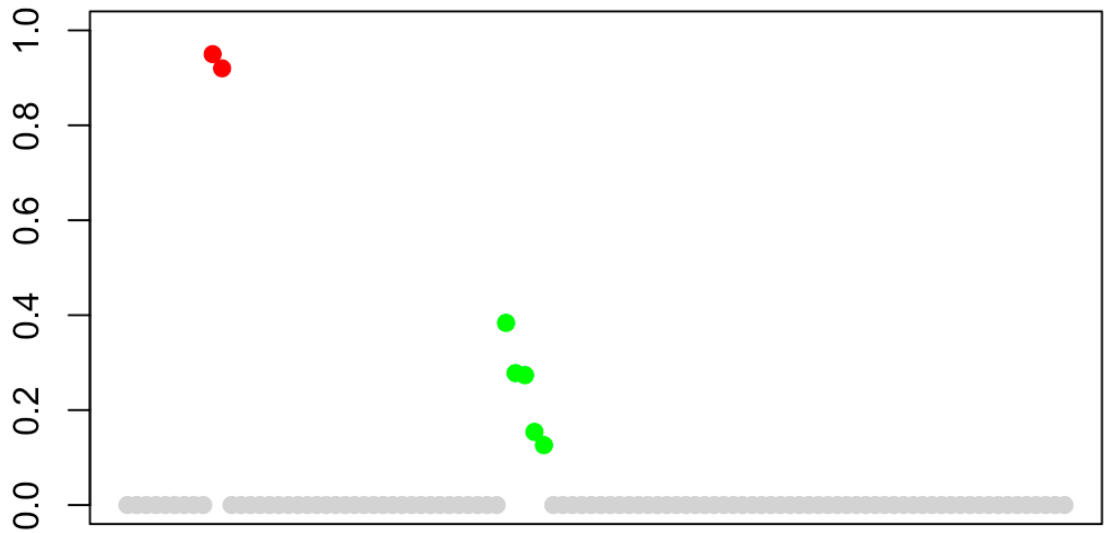
Difference in PIP

No Functional information



Non informative prior

Fine mapping with Functional information



Informative/Functional prior

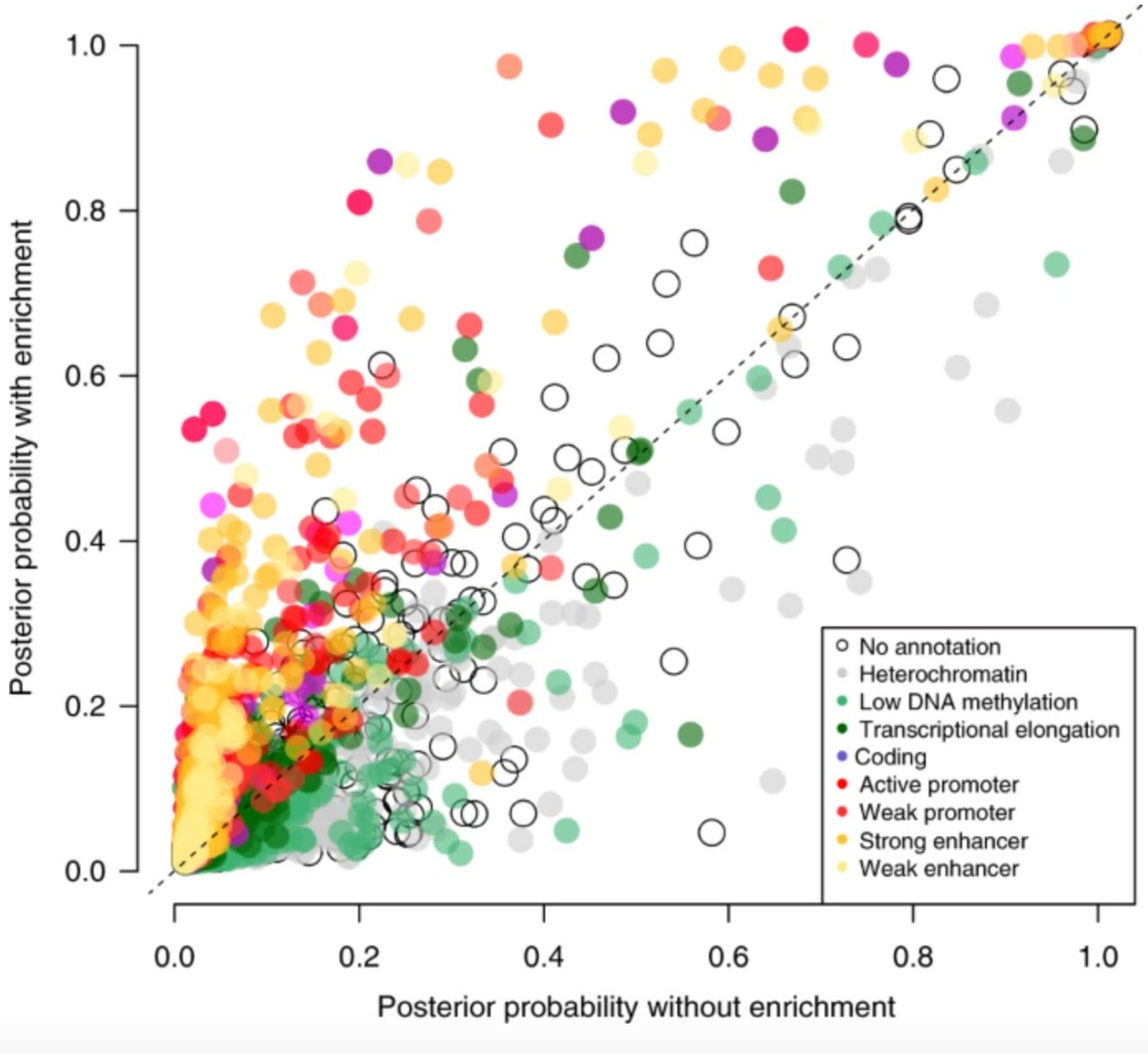
Methods that incorporate functional priors

Method	Author	
fGWAS	Pickrell et al 2014	Single causal variant & many annotations
PAINTOR	Kichaev et al 2014	Multiple causal variants (<6) & many annotations
CAVIARBF	Chen et al 2016	Multiple causal variants & many annotations
PolyFun	Weissbrod et al 2021	Using polygenic enrichment of functional categories as priors (from S-LDSC)

Commonly used Annotation sources

Sources	Purpose
CADD, SIFT, PolyPhen	Predicts the deleteriousness of individual variants
regulomeDB, ROADMAP ENCODE	Evaluates the regulatory potential of individual (non-coding) variants

Functional prior improves PIP

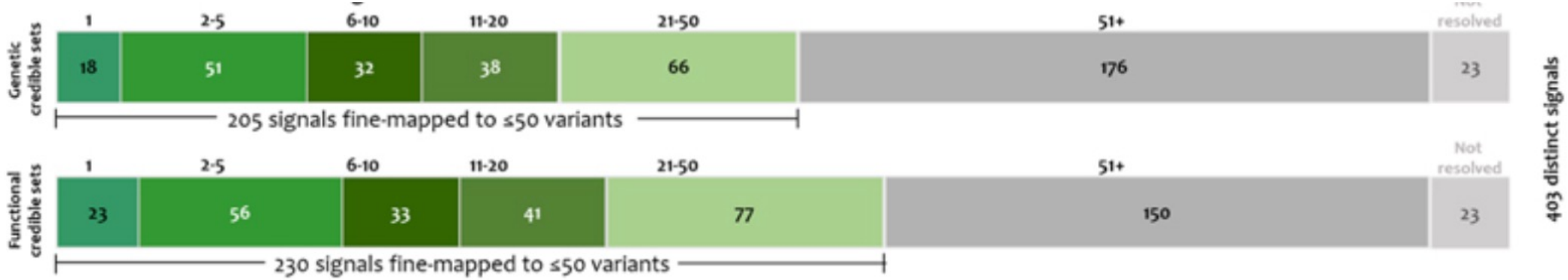


Analysis from Mahajan et al (2018): GWAS analysis with approximately 75K Type 2 diabetes cases and 900K controls

Colors signify SNPs different functional category (near TSS, promoter, enhancer etc.) in pancreatic islet

Image from: Mahajan et al (2018); *Nat. Gen.*

Functional priors improve CS resolution



Incorporating functional annotations as priors can improve resolution of credible sets: Smaller credible sets

Provides much granular information in comparison to noninformative prior

Image from: Mahajan et al (2018); *Nat. Gen.*

Improvement in PIP across traits

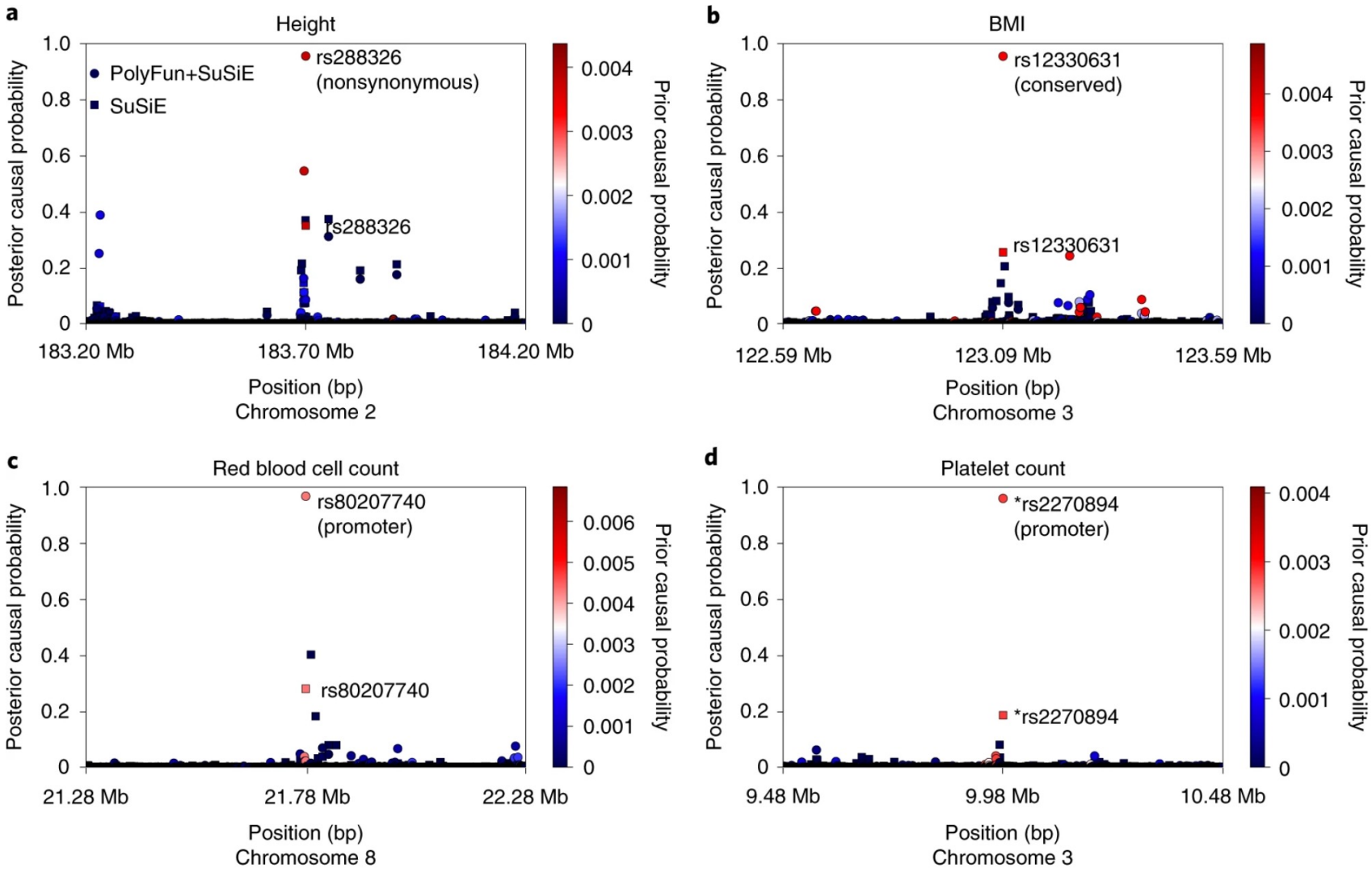


Image from: Weissbrod et al (2021); *Nat. Gen.*

Potential problems

- **Genomic annotations can be specific** to tissue, cell type and cell developmental stages
- Incorporating **misinformed prior** might blur inference and introduce noise
- Incorporating **too many annotations** might not be beneficial always
- Choice of the prior causal probability (PolyFun addresses this)

Fine mapping: Using summary statistics

Using summary statistics

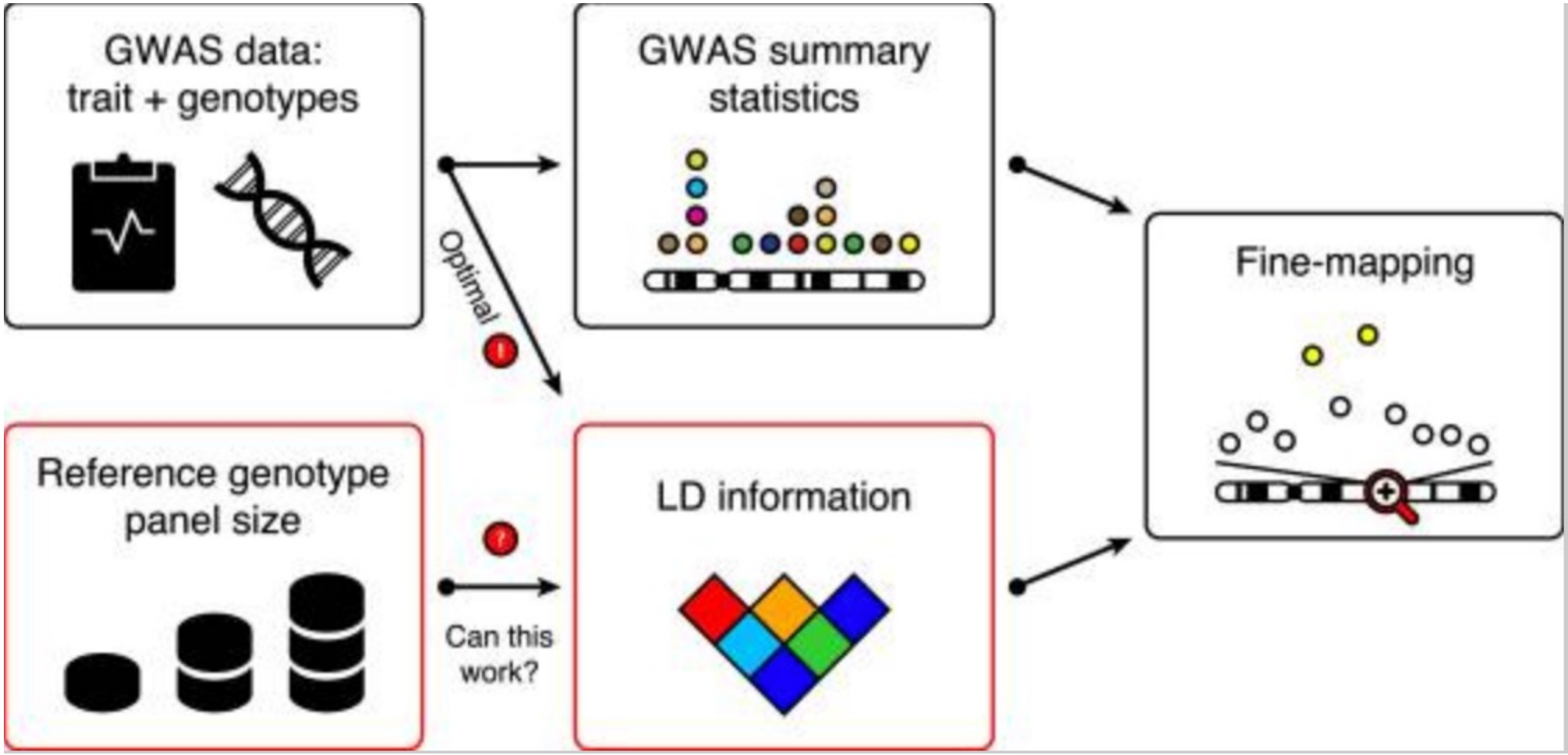


Image from: Benner et al (2017); *AJHG.*

Typical use of Fine mapping software

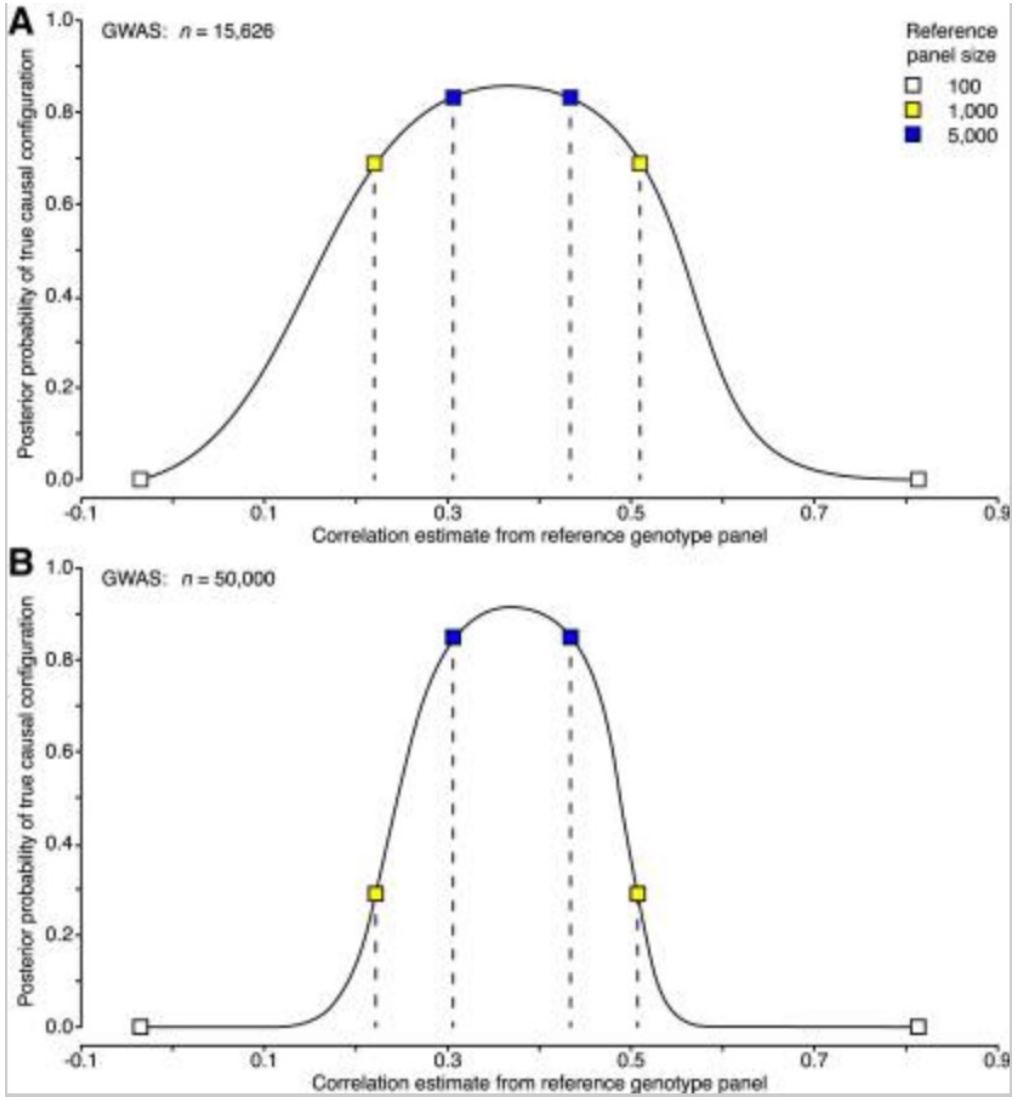
Input

- GWAS Summary statistics
  - Effect sizes
  - Standard errors
  - Minor allele frequencies
- LD estimated from publicly available reference panels

Output

- Posterior probability of each variant to be causal
- Credible set
- Posterior probability of different number of causal variants (if you want to choose empirically)

How big a reference panel to use



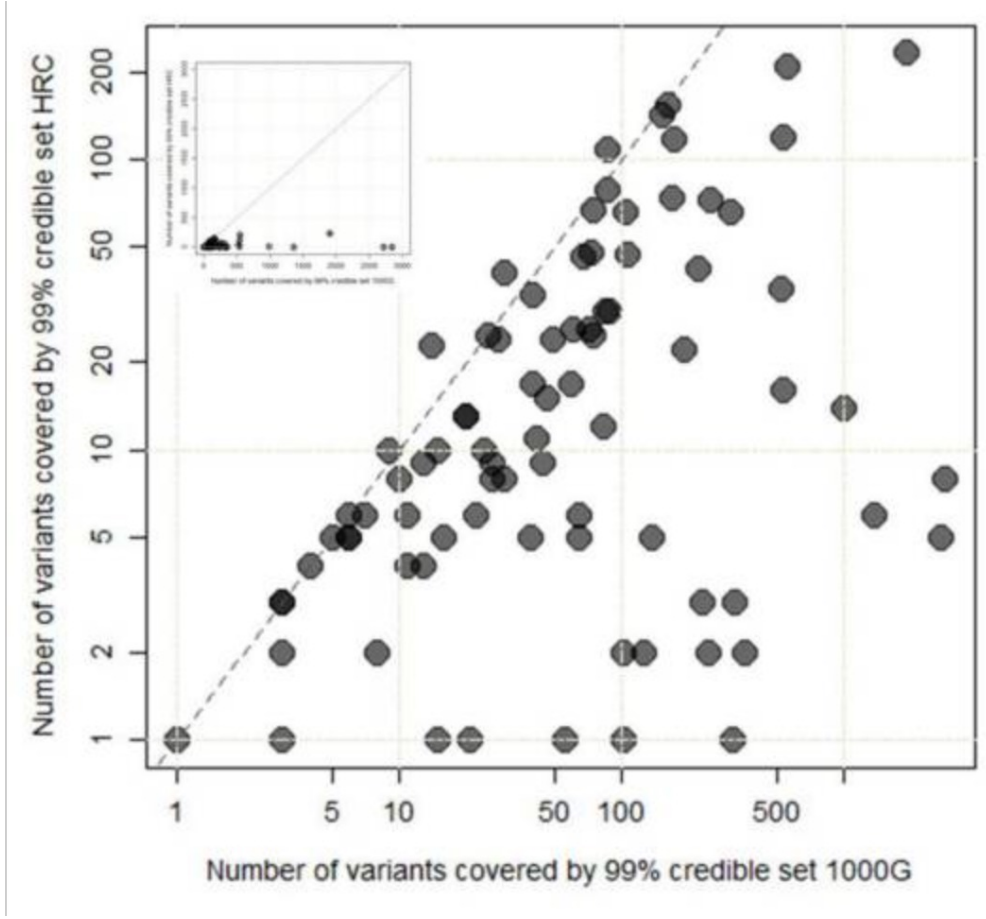
Reference panel size must scale with the sample size of GWAS

Populations must **match closely in terms of ancestral background**

Simulation with 2 variants: 1 causal & 1 non-causal with correlation 0.37

Image from: Benner et al (2018); *AJHG.*

Comparison of reference panels



1000G unrelated Europeans (N = 498) vs Haplotype Reference Consortium (N = 60,000)

Precise estimates of LD improves fine mapping resolution

Image from: Mahajan et al (2018); *Nat. Gen.*

Summary of Fine mapping

- Identifies putative **statistically causal variants** that impacts phenotypic variation
- **Bayesian methods** are usually preferred for their ease of interpretation
- Requires only **GWAS summary statistics and LD** estimated from reference panel
- **Functional genomic annotations** can be incorporated to make results biologically plausible
- Of the shelf software are readily available: **PAINTOR, CAVIAR & SuSiE** are the most well used
- Needs thorough benchmarking and evaluation for performance
- **Topics not covered**: Trans-ancestry fine mapping

Software

PAINTOR: Needs Z-values and LD estimates (C++ based)

https://github.com/gkichaev/PAINTOR_V3.0

Outputs posterior probability for each variant & credible set for given number of causal variants.

SuSiE: Needs Z-values and LD estimates (R based)

https://github.com/stephenslab/susieR

Outputs posterior probability for each variant & independent credible sets (each credible set contains at most one independent causal signal)

Take a look at their vignettes to learn more about how to use them

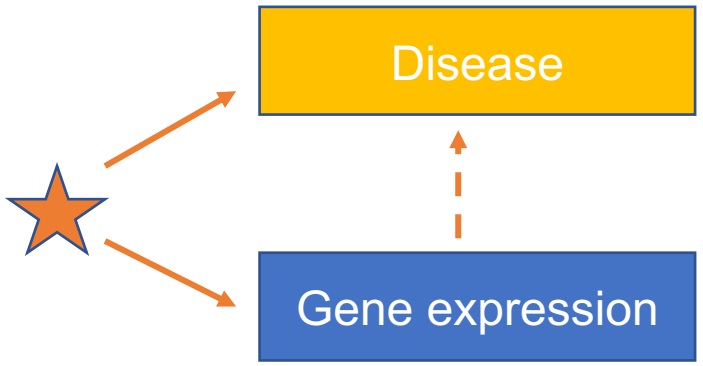
Colocalization

Colocalization: Motivation

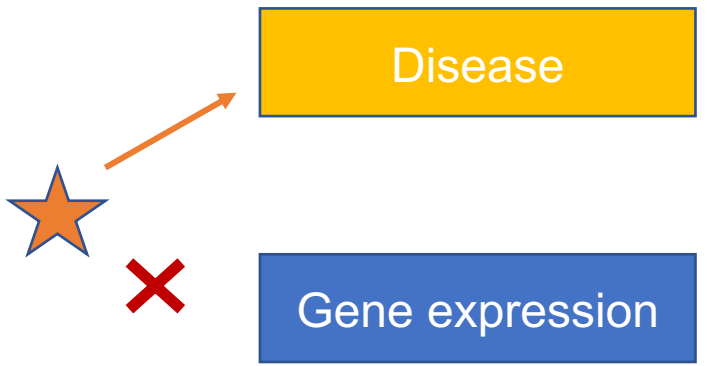
- Does the same “causal” variant affect more than one trait?

- Helps in interpretation if one of the traits is a molecular phenotype (like gene expression or protein level etc.)

High confidence & interpretable scenario



Low confidence & non-interpretable scenario



Setup

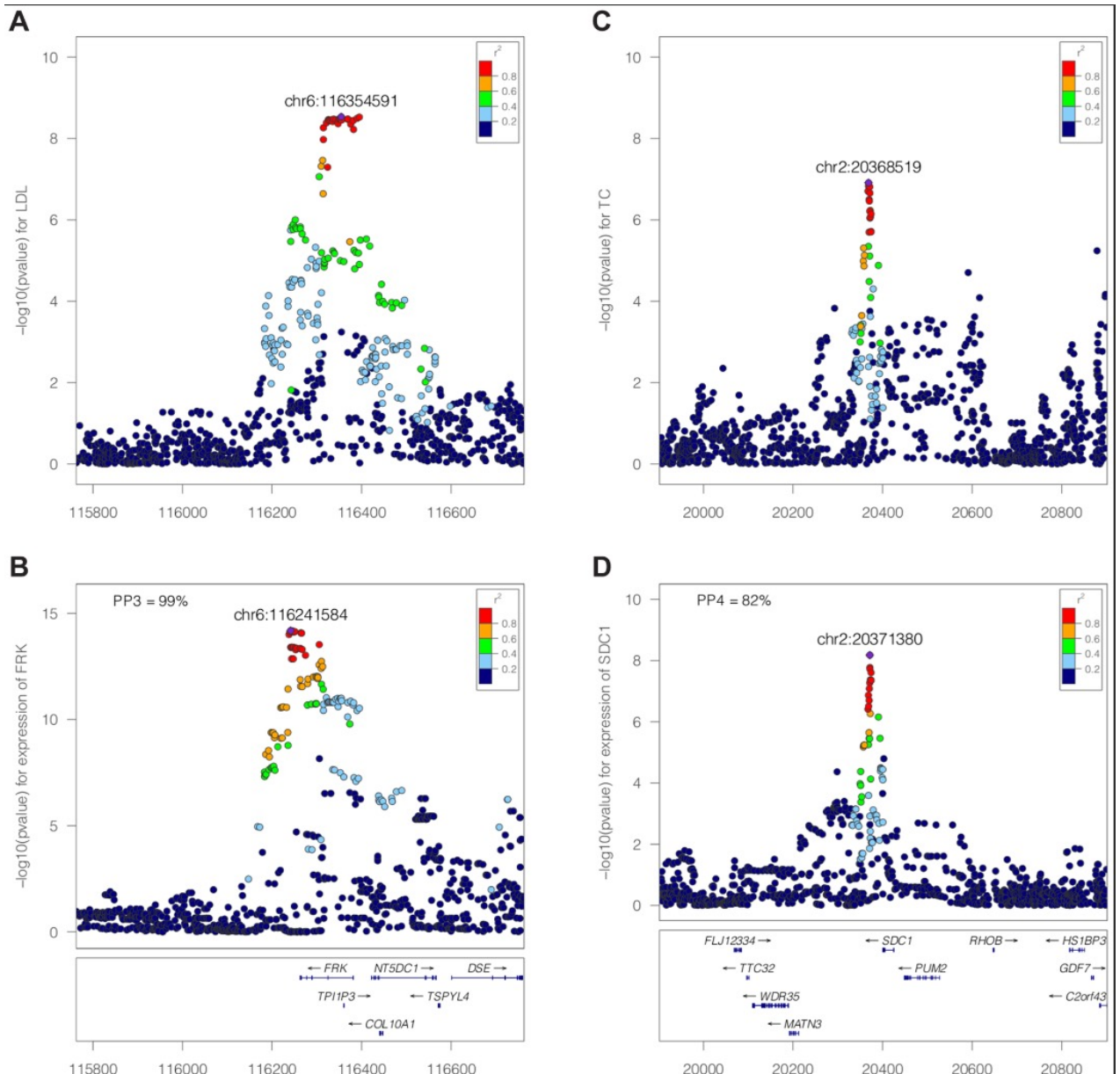
- Region of “m” SNPs
- Two traits: Marginal association test summary statistics
  - Trait 1: Disease outcome
  - Trait 2: Gene expression
- At most one causal variant per trait
- Partition the possible causal scenarios as:
  - **H0**: None of the variants are causal to any trait
  - **H1**: 1 causal variant for trait 1 & no causal variant for trait 2
  - **H2**: No causal variant for trait 1 & 1 causal variant for trait 2
  - **H3**: 1 causal variant for trait 1 & 1 causal variant for trait 2 but they are different
  - **H4**: 1 causal variant for trait 1 & 1 causal variant for trait 2 and they are the same (shared)

Input and output

- **Region-based** posterior probability for each of the causal scenarios
  - Most popularly used
  - Does not require identification of causal variant
  - Implemented in coloc R package

- **Variant level posterior probability** of colocalization (H4)
  - Depends on inference of causal variants
  - Implemented in eCAVIAR

Example



A-B: Colocalization of LDL with gene expression of *FRK* gene in Liver

C-D: Colocalization of Total Cholesterol with gene expression of *SDC1* in Liver

Image from: Giambartolomei et al (2018); *Plos. Gen.*

Pros & Cons

- In its original form, colocalization is a **region-based inference** and hence aggregates results across multiple causal scenarios.
- Can also get **posterior probability that a particular variant is the shared causal variant** across traits
- Similar to fine mapping, can be performed with **summary statistics**.
- Usually, the **space of causal scenarios is complicated** when multiple causal variants are present in either trait
- **Computation scales exponentially** with the number of traits (HyPrcolor addresses this)
- Somewhat **sensitive to prior values**

Multiple causal signals

- **Can partition the locus into subregions** containing at most one causal variant using conditional regression (or SuSiE) for each trait
- **SuSiE outputs independent credible sets** which contain at most one causal variant
- **Use all possible pairs of SuSiE CS** (one from each trait) to perform colocalization

Software

coloc R package: Needs GWAS summary statistics (for usual use)

To incorporate multiple causal signals, authors recommend a pipeline using SuSiE. Here we need LD estimates as well

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

Outputs: PPH0, …, PPH4

eCAVIAR: Needs Z-values and LD estimates for both the traits

https://github.com/fhormoz/caviar

Outputs: CLPP (colocalization posterior probability) for each variant (posterior prob. that the variant is a shared causal variant)

Not restricted to single causal variant setup

Questions?

NATIONAL CANCER INSTITUTE

www.cancer.gov

www.cancer.gov/espanol