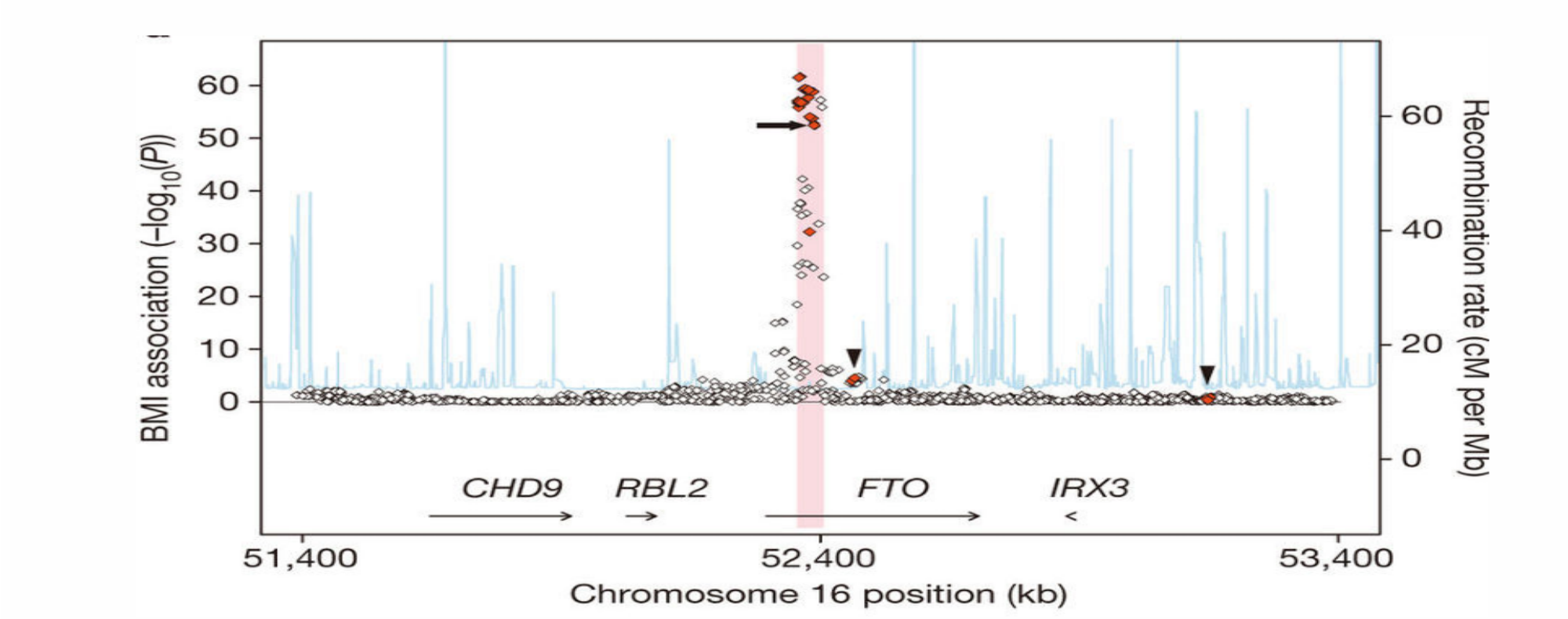
Finemapping and its application in Coloc

- A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping
- A Bayesian framework for multiple trait colocalization from summary association statistics
- A more accurate method for colocalisation analysis allowing for multiple causal variants

Single causal variant can drive association signals in multiple SNPs

- Causal variant: the variant with a true effect on the phenotype.



Lead SNPs are often not causal variants

- Because of sampling errors: a SNP in close LD with the causal SNP may have similar association statistics.
- Simulations with 1000 cases and 1000 controls: at effect size 1.1 and AF 5%, causal variant has 2.4% chance of being the lead SNP.
- The vast majority of variants discovered by GWAS have small effect sizes $\log -\mathrm{OR} < 1.1$.

Lead SNPs are often not causal variants

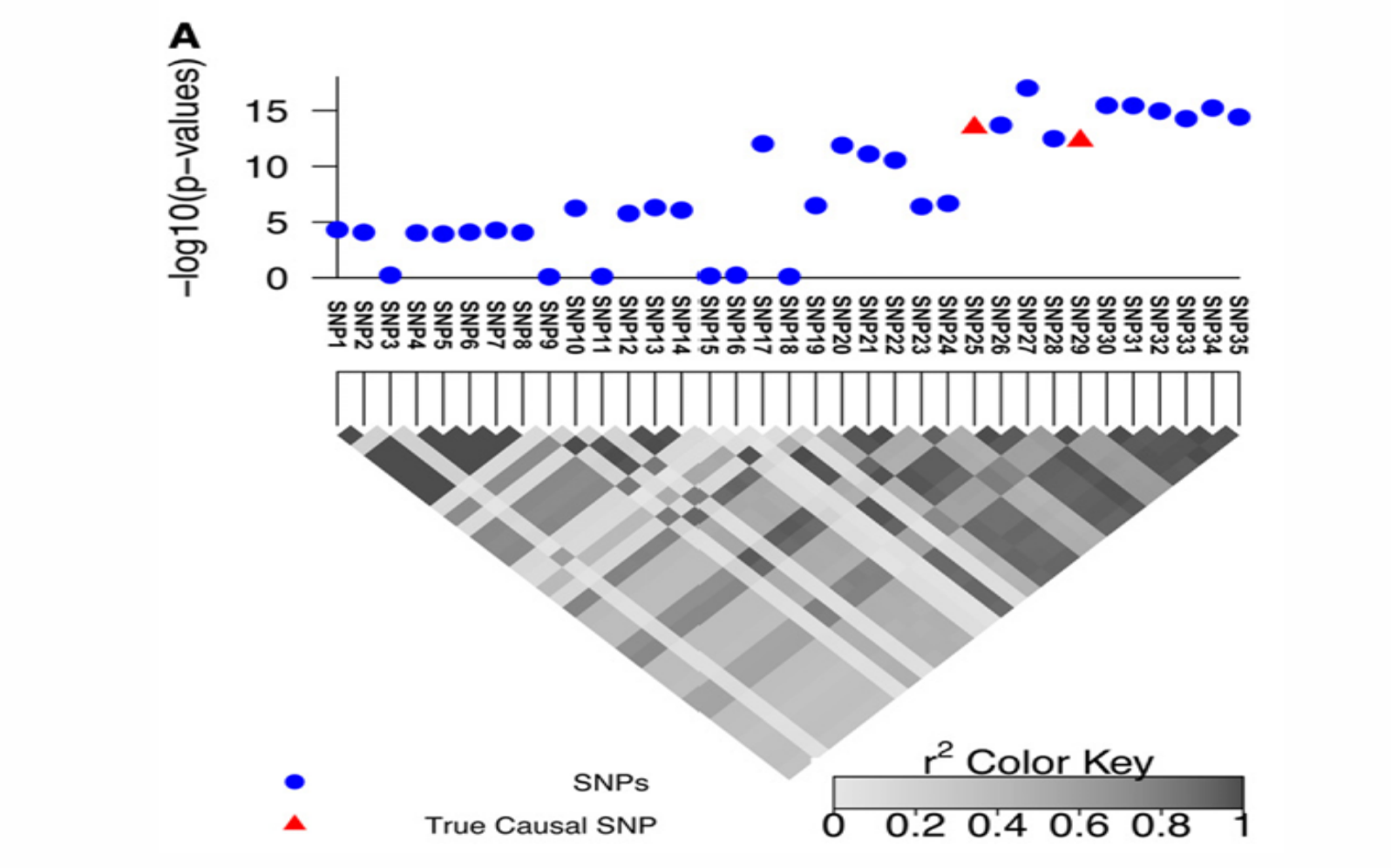


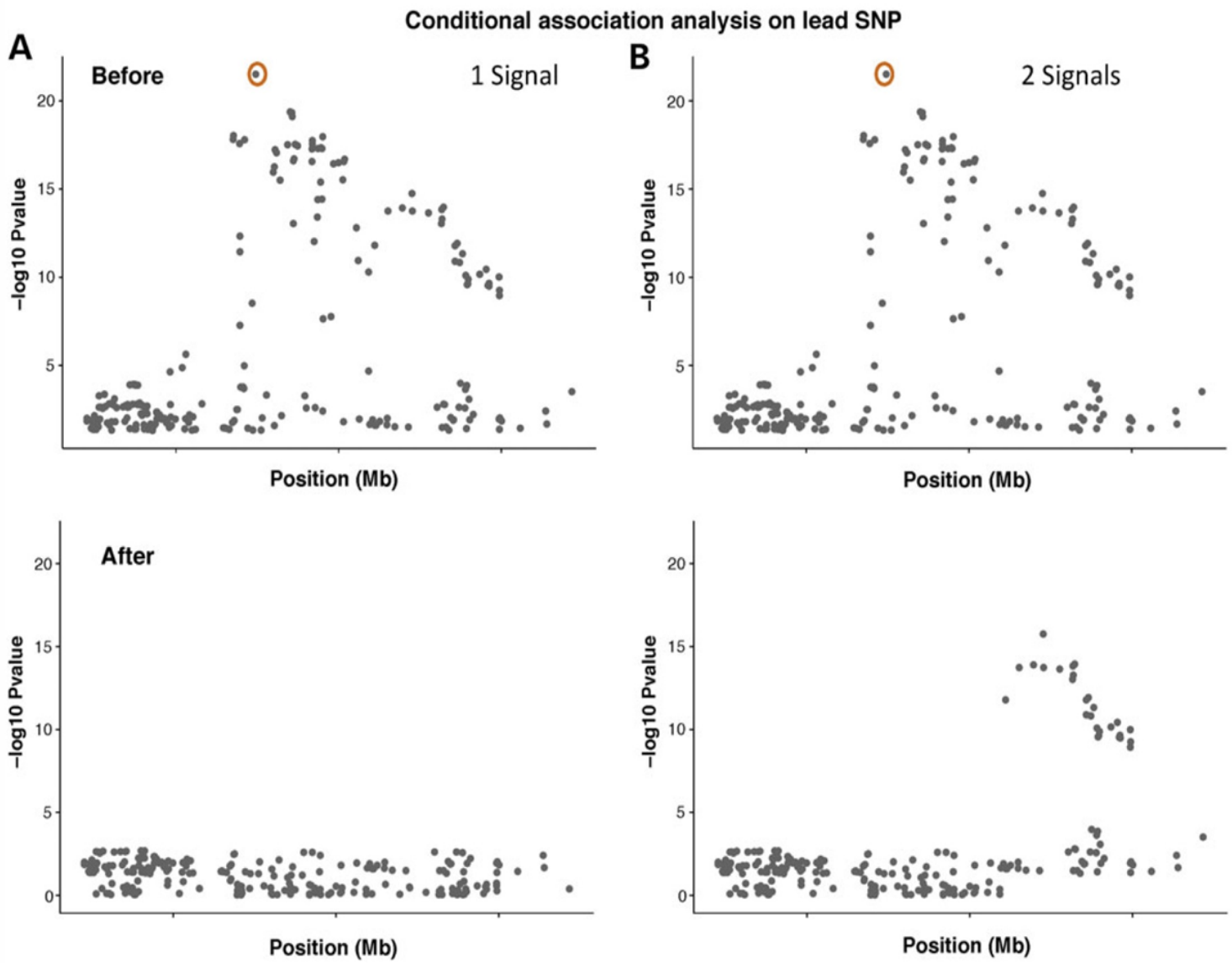
Figure: Simulated data with 2 causal variants (red) [Hormozdiari et al, Genetics, 2014]

Conditional Regression

- Regression analysis on a SNP j by conditioning on the lead SNP:

$$Y = G_{\text{lead}} \cdot \beta_{\text{lead}} + G_j \beta_j + \epsilon$$

- A SNP passing the threshold will be chosen.(pval<5*10^-8)
- Repeat this analysis: at each step, condition on all SNPs chosen at previous steps.


Conditional association analysis on lead SNP
A
Before
1 Signal
B
2 Signals
After
−log10 Pvalue
Position (Mb)
0
5
10
15
20

Conditional Regression

- Lead SNPs may not be causal: wrong decision at the beginning.
- It is unclear how to account for multiple testing and choose the threshold: at each step, many hypothesis are tested.
- Low power of detecting the secondary SNP.

Finemapping

- Finding a small number of variants that explain all the associations in a region.
- PIP: the posterior probability that a SNP is causal, summing over all possible configurations.

$$P\left(\gamma_j = 1 \mid D\right) = \sum_{\gamma} \underbrace{I\left(\gamma_j = 1\right)}_{1 \text{ if } \gamma_j=1;0 \text{ otherwise}} \cdot P(\gamma \mid D)$$

- Credible set: the minimum set of variant that contains the causal variant with probability α (typically, 95%).
  Define the "confidence level" of a variant set S. If there is a single causal variant, then it is simply the sum of PIP of all the variants in S. Ex. $S = \{A, B, C\}$, its confidence level is:

$$\rho = \mathrm{PIP}_A + \mathrm{PIP}_B + \mathrm{PIP}_C$$

SuSiE

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \quad \mathbf{e} \sim N_n\left(0, \sigma^2 I_n\right)$$

- variables j with non-zero effects $(b_j \neq 0)$ as 'effect variables'.
- Assume now that exactly two variables are effect variables-variables 1 and 4, and that these two effect variables are each completely correlated with another non-effect variable, say $x_1 = x_2$ and $x_3 = x_4$. Further suppose that no other pairs of variables are correlated. Here, because the effect variables are completely correlated with other variables, it is impossible to select the correct variables confidently, even when n is very large. However, given sufficient data it should be possible to conclude that there are (at least) two effect variables, and that

$$(b_1 \neq 0 \text{ or } b_2 \neq 0) \text{ and } (b_3 \neq 0 \text{ or } b_4 \neq 0)$$

SuSiE

‘single-effect regression’ (SER) model

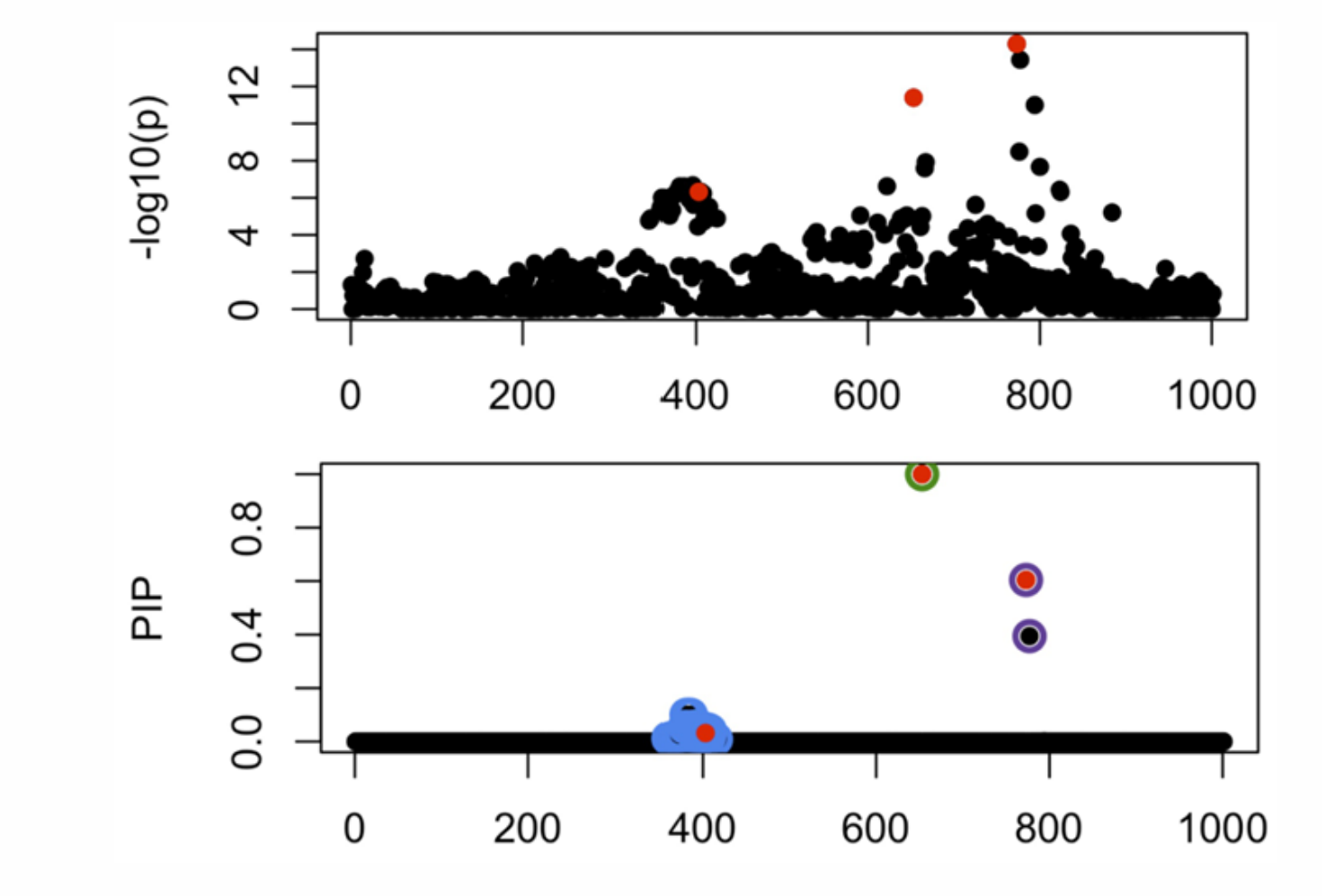
$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$
$$\mathbf{e} \sim N_n\left(0, \sigma^2 I_n\right)$$
$$\mathbf{b} = b\gamma$$
$$\gamma \sim \mathrm{Mult}(1, \boldsymbol{\pi})$$
$$b \sim N_1\left(0, \sigma_0^2\right)$$

Mult(m, π) denotes the multinomial distribution on class counts that is obtained when m samples are drawn with class probabilities given by π.
the prior variance of the non-zero effect, σ_0^2, and the prior inclusion probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)$, in which π_j gives the prior probability that variable j is the effect variable are given

SuSiE

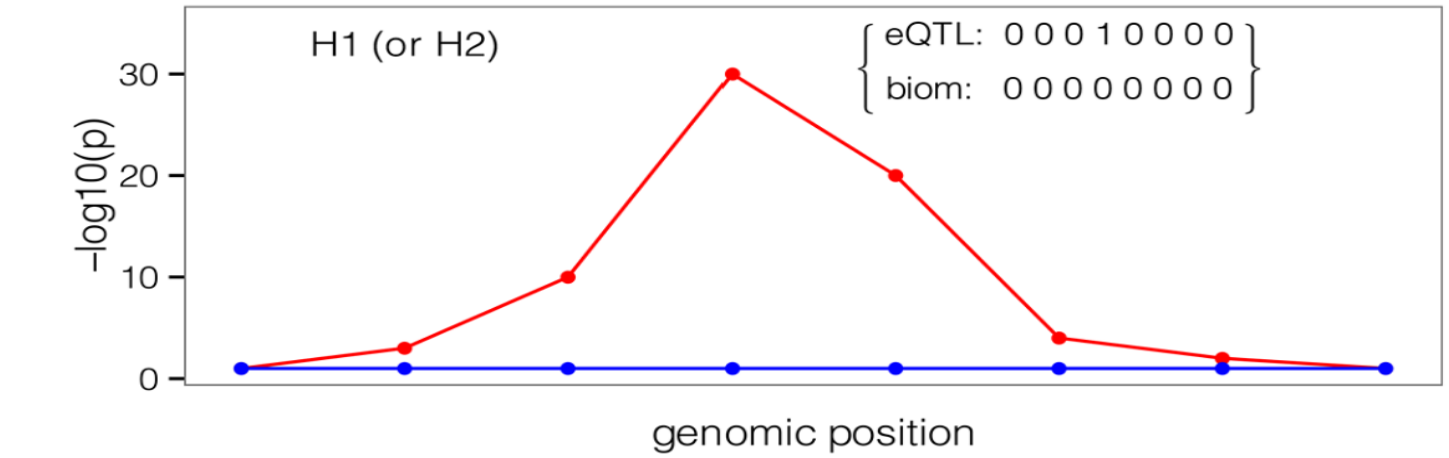
The key idea is simple: introduce multiple single-effect vectors $\mathbf{b}_1, \ldots, \mathbf{b}_L$ and construct the overall effect vector $\mathbf{b}$ as the sum of these single effects. We call this the 'sum of single effects' regression model SuSiE:

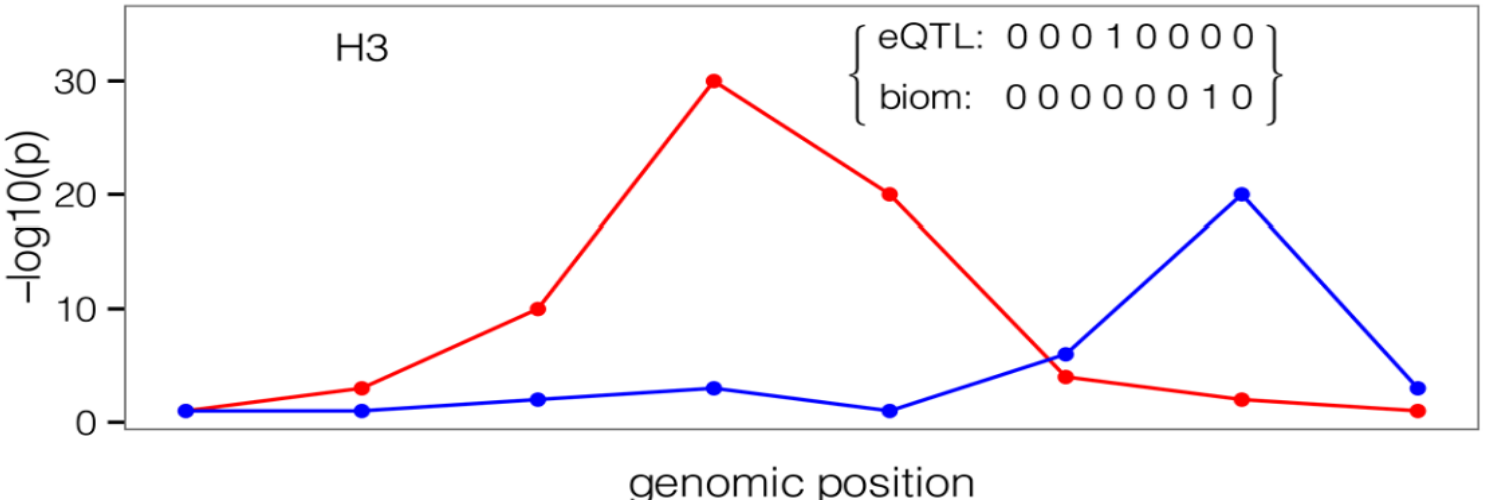
$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$
$$\mathbf{e} \sim N_n\left(0, \sigma^2 I_n\right)$$
$$\mathbf{b} = \sum_{l=1}^{L} \mathbf{b}_l$$
$$\mathbf{b}_l = \gamma_l b_l$$
$$\gamma_l \sim \mathrm{Mult}(1, \pi)$$
$$b_l \sim N_1\left(0, \sigma_{0l}^2\right)$$

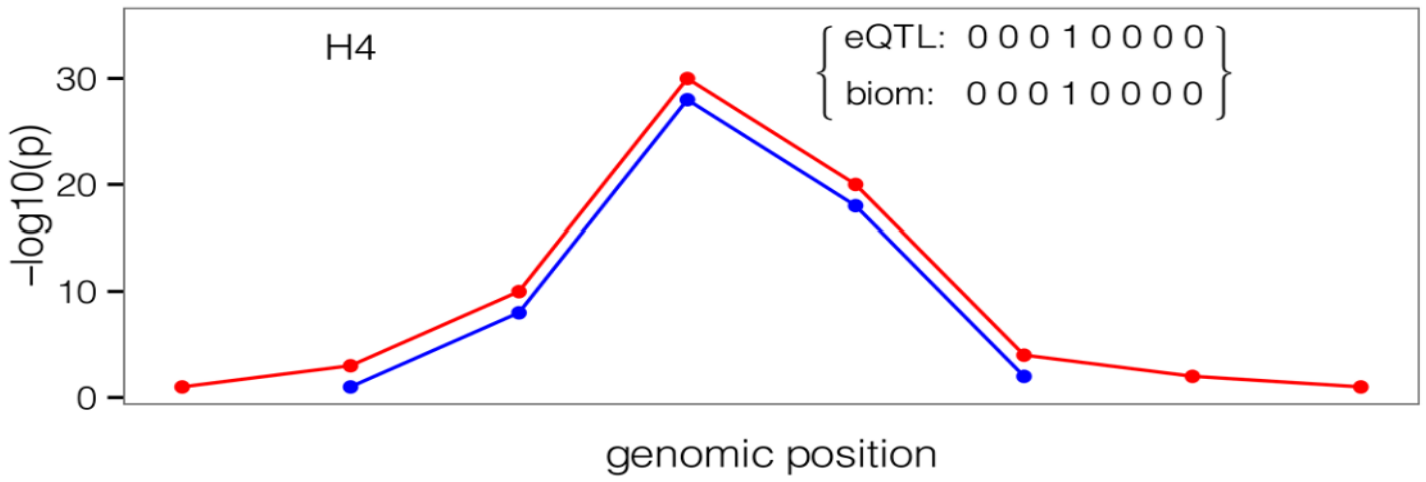


Fine-mapping (SuSiE) will uncover several credible sets , each capturing one causal variant.

COLOC


H1 (or H2)
eQTL: 0 0 0 1 0 0 0 0
biom: 0 0 0 0 0 0 0 0
30
20
10
0
−log10(p)
genomic position


H3
eQTL: 0 0 0 1 0 0 0 0
biom: 0 0 0 0 0 0 1 0
30
20
10
0
−log10(p)
genomic position



H_0 : no association with either trait in the region

H_1 : association with trait 1 only

H_2 : association with trait 2 only

H_3 : both traits are associated, but have different single causal variants

H_4 : both traits are associated and share the same single causal variant

Coloc

Assumption

- Firstly, that the causal variant is included in the set of Q variants, either directly typed or well imputed
- Secondly, that at most one association is present for each trait in the genomic region of interest.

Input

- Summary statistics
- Imputation

Coloc

$$
\begin{aligned}
PP4 &= P\left(H_4 \mid D\right) \\
&= \frac{P\left(H_4 \mid D\right)}{P\left(H_0 \mid D\right) + P\left(H_1 \mid D\right) + P\left(H_2 \mid D\right) + P\left(H_3 \mid D\right) + P\left(H_4 \mid D\right)} \\
&= \frac{\frac{P(H_4|D)}{P(H_0|D)}}{1 + \frac{P(H_1|D)}{P(H_0|D)} + \frac{P(H_2|D)}{P(H_0|D)} + \frac{P(H_3|D)}{P(H_0|D)} + \frac{P(H_4|D)}{P(H_0|D)}}
\end{aligned}
$$

- The ratios in the numerator and denominator are:

$$
\frac{P\left(H_h \mid D\right)}{P\left(H_0 \mid D\right)} = \sum_{S \in S_h} \frac{P(D \mid S)}{P\left(D \mid S_0\right)} \times \frac{P(S)}{P\left(S_0\right)}
$$

Coloc

ABF

- Asymptotically, that is as n increase, the MLE $\hat{\theta}$ has the normal distribution $N(\theta, V)$.
- Combining this "likelihood" with a normal prior, $N(0, W)$, on the log relative risk, θ, gives the asymptotic Bayes factor

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

where $z^2 = \hat{\theta}^2/V$ is the usual Wald statistic. High/low values of the asymptotic Bayes factor occur when z^2 is small/large and correspond to evidence for/against the null hypothesis.

Coloc

- If $S \in S_0$, then $\frac{P(S)}{P(S_0)} = \frac{p_0^Q}{p_0^Q} = 1$
- If $S \in S_1$, then $\frac{P(S)}{P(S_0)} = \frac{p_0^{Q-1}}{p_0^Q} \times p_1 = \frac{p_1}{p_0} \approx p_1$
- If $S \in S_2$, then $\frac{P(S)}{P(S_0)} = \frac{p_0^{Q-1}}{p_0^Q} \times p_2 = \frac{p_2}{p_0} \approx p_2$
- If $S \in S_3$, then $\frac{P(S)}{P(S_0)} = \frac{p_0^{Q-2}}{p_0^Q} \times p_1 \times p_2 = \frac{p_1}{p_0} \times \frac{p_2}{p_0} \approx p_1 \times p_2$
- If $S \in S_4$, then $\frac{P(S)}{P(S_0)} = \frac{p_0^{Q-1}}{p_0^Q} \times p_{12} = \frac{p_{12}}{p_0} \approx p_{12}$

Coloc

- $\frac{P(H_0|D)}{P(H_0|D)} = 1$
- $\frac{P(H_1|D)}{P(H_0|D)} = p_1 \times \sum_{j=1}^{Q} ABF_j^1$
- $\frac{P(H_2|D)}{P(H_0|D)} = p_2 \times \sum_{j=1}^{Q} ABF_j^2$
- $\frac{P(H_3|D)}{P(H_0|D)} = p_1 \times p_2 \times \sum_{j,k,j \neq k} ABF_j^1 ABF_k^2$
- $\frac{P(H_4|D)}{P(H_0|D)} = p_{12} \times \sum_{j=1}^{Q} ABF_j^1 \times ABF_j^2$

- Superscript 1,2 : trait1,2
- Subscript i,j : snp i,j

Coloc

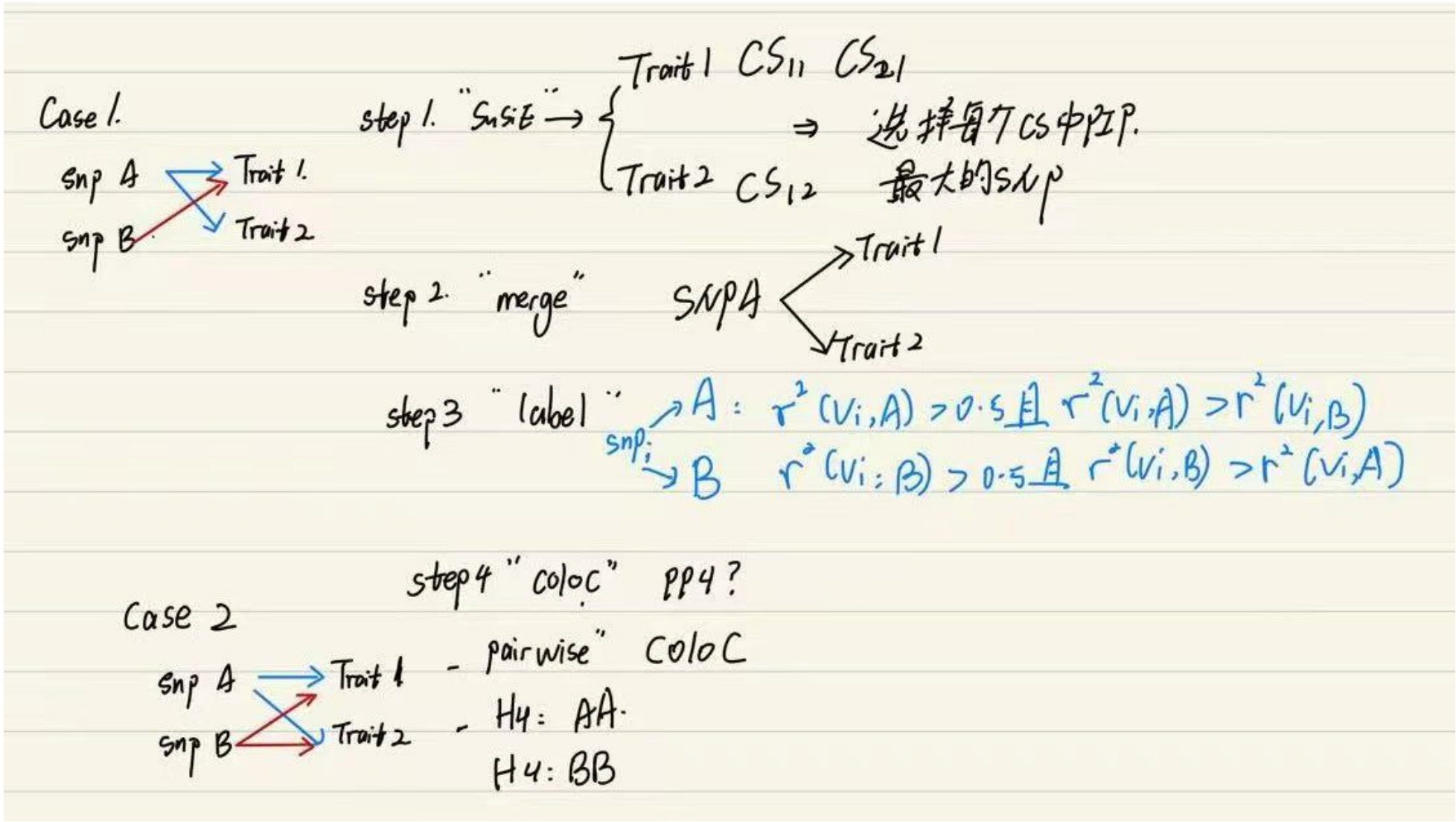
Coloc vs MR

- MR Asymmetric in the traits: one trait is the exposure, the other is the outcome
- Coloc Symmetric in the traits: the traits are treated equivalently in the analysis
- Coloc is often used as sensitivity analysis after MR

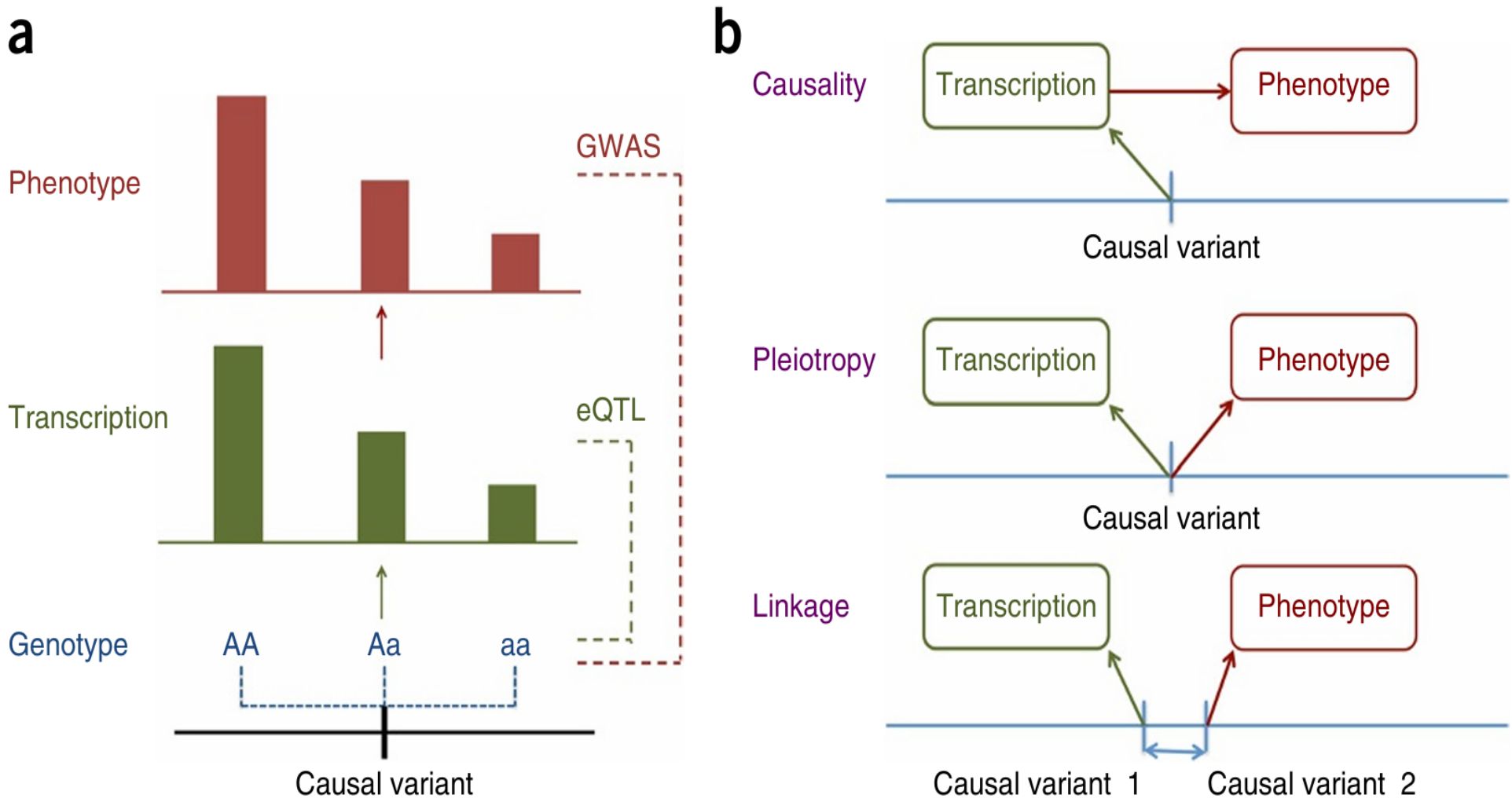
Advabtage & Disadvantage

- Robust
- Cumbersome if there are more than 2 traits（MOLOC）
- Difficult to scan the whole genomie

Coloc & SuSiE


Case 1.
snp A
snp B
Trait 1.
Trait 2
step 1. "SuSiE" → { Trait 1 CS11 CS21 ⇒ 选择每个CS中的P. 最大的SNP
Trait 2 CS12
step 2. "merge"
SNP A
Trait 1
Trait 2
step 3 "label"
snp: A : r²(Vi,A) > 0.5 且 r²(Vi,A) > r²(Vi,B)
B r²(Vi,B) > 0.5 且 r²(Vi,B) > r²(Vi,A)
step 4 "coloc" PP4?
Case 2
snp A
snp B
Trait 1
Trait 2
- pairwise" coloC
- H4: AA.
H4: BB

SMR & HEIDI


a
Phenotype
GWAS
Transcription
eQTL
Genotype
AA
Aa
aa
Causal variant
b
Causality
Transcription
Phenotype
Causal variant
Pleiotropy
Transcription
Phenotype
Causal variant
Linkage
Transcription
Phenotype
Causal variant 1
Causal variant 2

HEIDI

- Analogue to Cochran Q test in MR_IVW
- Distinguish functional association from linkage.
  If a trait and gene expression are affected by the same causal variant (pleiotropy), then b_{xy} calculated using any SNP in LD with the causal variant is identical. This is because, under Hardy-Weinberg equilibrium, for any SNP i

$$b_{xy(i)} = \frac{b_{zy(i)}}{\beta_{zx(i)}} = \frac{b_{zy(0)} r_{0i} \sqrt{h_0/h_i}}{\beta_{zx(0)} r_{0i} \sqrt{h_0/h_i}} = \frac{b_{zy(0)}}{\beta_{zx(0)}} = b_{xy(0)}$$

HEIDI

- Under the null hypothesis that there is no heterogeneity, that is, where $\mathbf{d} = 0$, we have a vector of standard normal variables $\mathbf{z}_d = \left\{ z_{d(1)}, z_{d(2)}, \ldots, z_{d(m)} \right\}$ with $z_{d(i)} = \hat{d}_i / \sqrt{\mathbf{var}\left(\hat{d}_i\right)}$ and $\mathbf{z}_d \sim \mathbf{MVN}(0, \mathbf{R})$, where $\mathbf{R}$ is the correlation matrix with the ij th element being $r\left(z_{d(i)}, z_{d(j)}\right) = \mathbf{cov}\left(\hat{d}_i, \hat{d}_j\right) / \sqrt{\mathbf{var}\left(\hat{d}_i\right) \mathbf{var}\left(\hat{d}_j\right)}$
- HEIDI test statistic: $T_{\text{HEIDI}} = \mathbf{z}_d \mathbf{z}_d^T$, that is, $T_{\text{HEIDI}} = \sum_i^m z_{d(i)}^2$

Multi-HEIDI

$$\widehat{d}_{x_{ij}(s)} = \hat{b}_{x_i(s)} - \hat{b}_{x_i(\text{top } i)}$$

- where the subscript "i" represents the i th molecular phenotype, s represents an instrument from a set of x SNPs present in all the xQTL datasets and passing an xQTL p value threshold, and the subscript "top;" represents the lead xSNP for molecular phenotype i. The x SNPs are required to be instrument SNPs (i.e., $P_{XQTL} < 1.6 \times 10^{-3}$) for all tested molecular phenotypes.
- The null hypothesis of the multi-exposure HEIDI test is that all $d_{x_{iy}(s)} = 0$ and the alternative hypothesis of the multi-exposure HEIDI test is that any $d_{x_{iy}(s)} \neq 0$. If $\widehat{d}_{x_{iy}(s)}$ is computed for two xSNPs

Remark

- HEIDI+COJO
  （部分）解决了这个问题，但是整体而言，这个方法相较coloc对LD是比较敏感的，且在有多个exposure时表现较差
- COLOC+SuSiE
  利用了finemapping的结果，但是对于哪些snp是某个casual snp的“LD代理”的做法比较草率,PIP最大的也未必是causal variant
  $\mathbf{A} r^2\left(v_i, A\right)>0.5 \wedge r^2\left(v_i, A\right)>r^2\left(v_i, B\right)$
  $\mathbf{B} r^2\left(v_i, B\right)>0.5 \wedge r^2\left(v_i, B\right)>r^2\left(v_i, A\right)$

Remark

- HEIDI
- Low power in detecting linkage when two varients are in high LD
- Singe casual variant assumption

- Multi-HEIDI
- Reject "pleiotropy" when two casual varients are in LD
- Heterogeneity in different QTLs datasets