HGEN 471: Meta-Analysis and Basics of Fine-mapping

Prof. Xin He

January 31, 2022

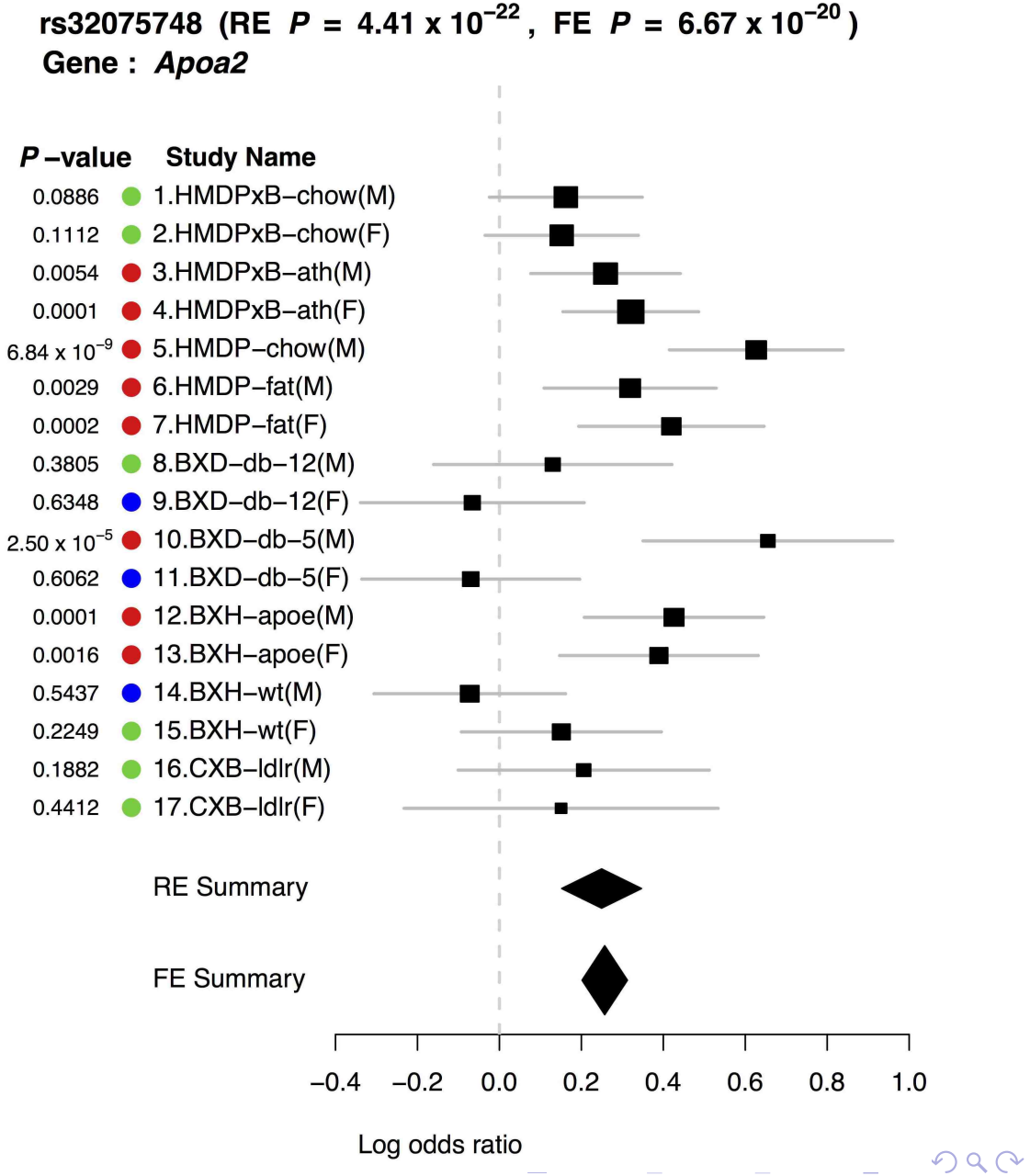
Lecture Objectives

- Learn general statistical methods for meta-analysis: effect size-based (fixed and random effects).
- Application of meta-analysis in GWAS.
- Learn the basic concepts for statistical fine-mapping analysis.

Meta-Analysis combines multiple studies to increase power

Most risk loci of common disease were discovered by large-scale meta-analysis of GWAS.

- Greatly increase the sample size.
- Only require summary statistics: much easier to share the data across studies.



Can we just count number of significant studies?

This is misleading:

- Counting loses quantitative information.
- All studies, regardless of sample sizes, contribute equally.
- Power not taken into account: when the power is 50%, only 50% studies will show significance even if there is an effect!

The GWAS example: vote counting suggests "conflicting" results, 9/17 studies have $p > 0.05$. Proper meta-analysis: $p = 6.7 \times 10^{-20}$.

Meta-analysis is often based on effect sizes, rather than p-values

Why not use p-values? Lose power information: all studies contribute equally.

Effect size: strength of relationship between two variables, β in $Y = X\beta + \epsilon$. Effect size in GWAS:

- Binary trait: log odds-ratio (or relative risk).
- Continuous trait: change of Y by one copy of alternative alleles.

Problems that meta-analysis addresses:

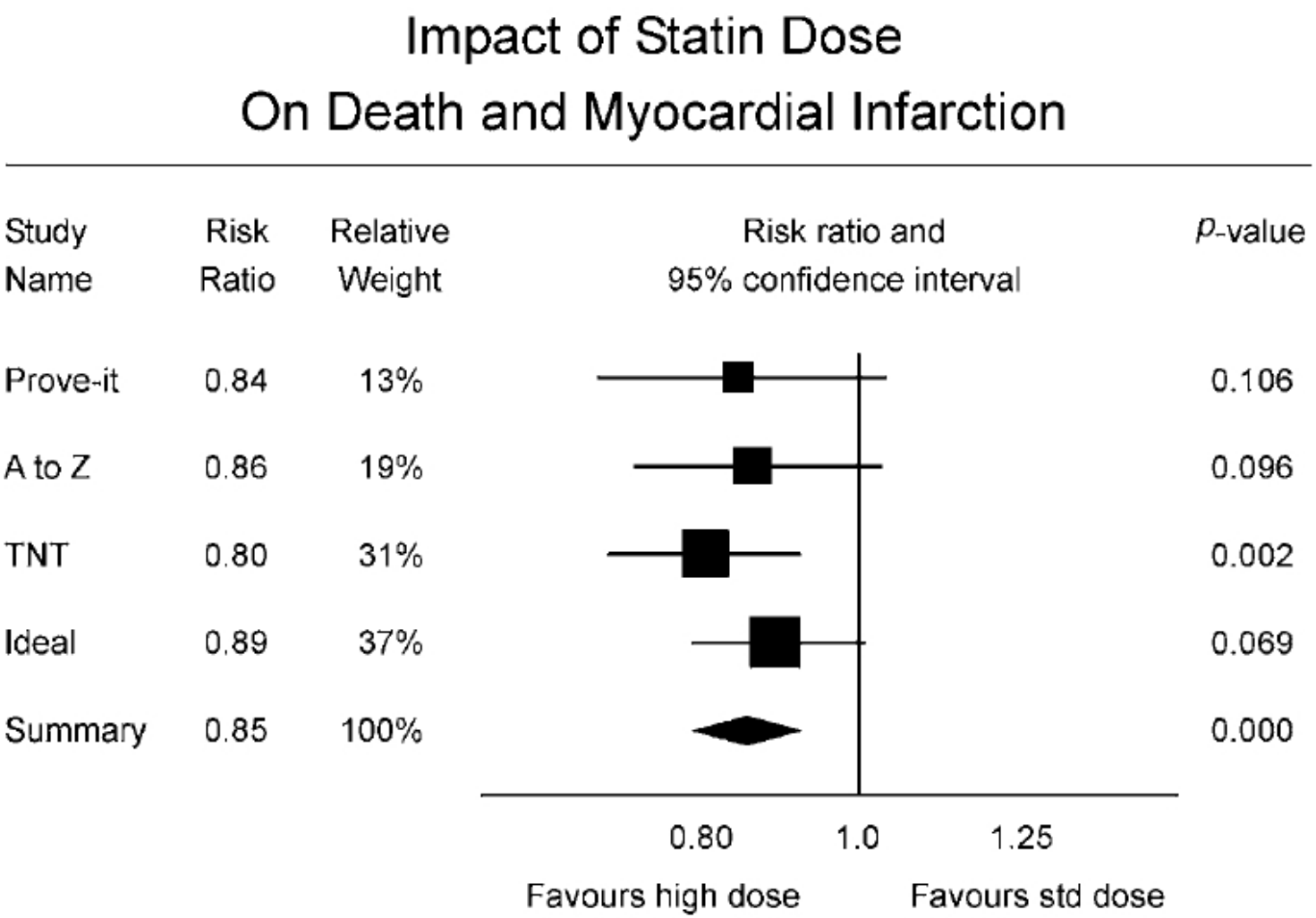
- What is the effect size? Is it significantly different from 0?
- Are the effect sizes consistent across multiple studies?

Procedure of effect size-based meta-analysis

- For each study, obtain the estimated effect size and its standard error.
- Assign study weights: based on the standard errors, which largely depends on the sample size.
- Summary effect is the weighted average of all effects.
- Determine the distribution of summary effect and its statistical singificance.

Two approaches for weighing and combining studies: fixed effects or random effects.

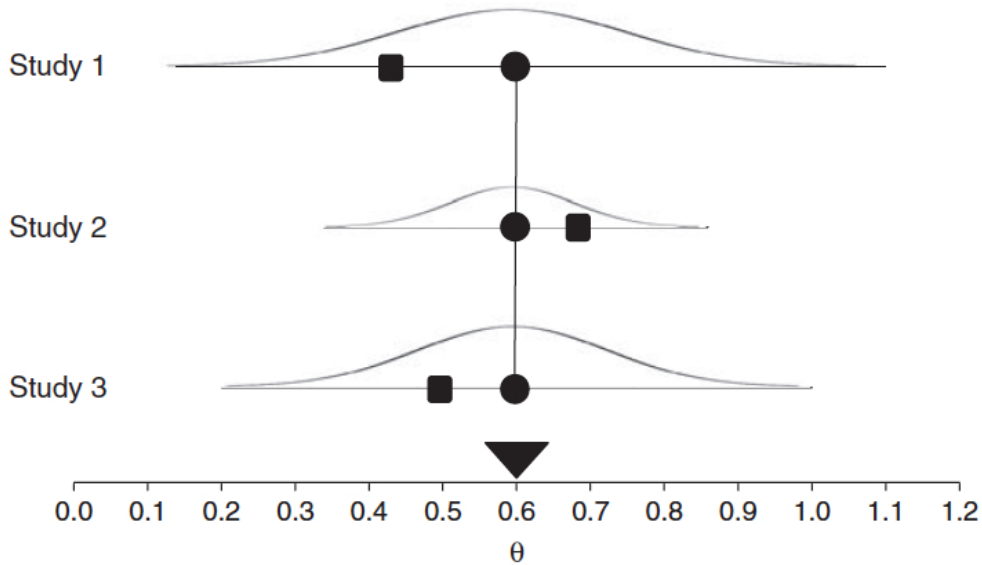
Meta-analysis results can be displayed with Forest Plot



Fix Effects (FE) meta-analysis estimates a single common effect

Assumption 1. All studies share a common true effect size.

Assumption 2. The observed effect in each study varies from the true effect because of sampling error.



	True effect	Observed effect
Study	●	■
Combined	▼	◆

The observed effect sizes follow normal distributions

Consider a simple regression: $Y = X\beta + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$.
The estimated effect of β, $\hat{\beta}$, follows the distribution:

$$\hat{\beta} \sim N(\beta, s^2), \qquad s^2 = \frac{\sigma^2}{n \cdot \hat{Var}(X)}$$

where n is sample size, and $\hat{Var}(X)$ is the sample variance of X.

The sampling variance of the estimated effect size (s^2) is proportional to the inverse of sample size (n).

Fix Effects (FE) model estimates the true effect size by Maximum Likelihood

Model: let Y_i be the observed effect size of study i, $1 \leq i \leq k$, and V_i its variance. The true effect size is μ. Our model is written as:

$$Y_i \sim N(\mu, V_i)$$

Fix Effects (FE) model estimates the true effect size by Maximum Likelihood

Model: let Y_i be the observed effect size of study i, $1 \leq i \leq k$, and V_i its variance. The true effect size is μ. Our model is written as:

$$Y_i \sim N(\mu, V_i)$$

The likelihood function:

$$L(\mu) = P(Y|\mu) = \underbrace{\prod_i P(Y_i|\mu)}_{\text{studies are independent}} \propto \exp\left[-\sum_i \frac{(Y_i - \mu)^2}{2V_i}\right]$$

Fix Effects (FE) model estimates the true effect size by Maximum Likelihood (II)

Maximize the likelihood function $L(\mu)$ now becomes:

$$\text{Minimize: } \sum_i \frac{(Y_i - \mu)^2}{2V_i}$$

Let $w_i = 1/V_i$, the MLE of the summary effect is:

$$\hat{\mu} = \frac{\sum_i w_i Y_i}{\sum_i w_i} = \sum_i \underbrace{\frac{w_i}{\sum_i w_i}}_{\text{weight of study } i} Y_i$$

Exercise: show this is the MLE.

FE meta-analysis weights studies by the inverse of sample sizes

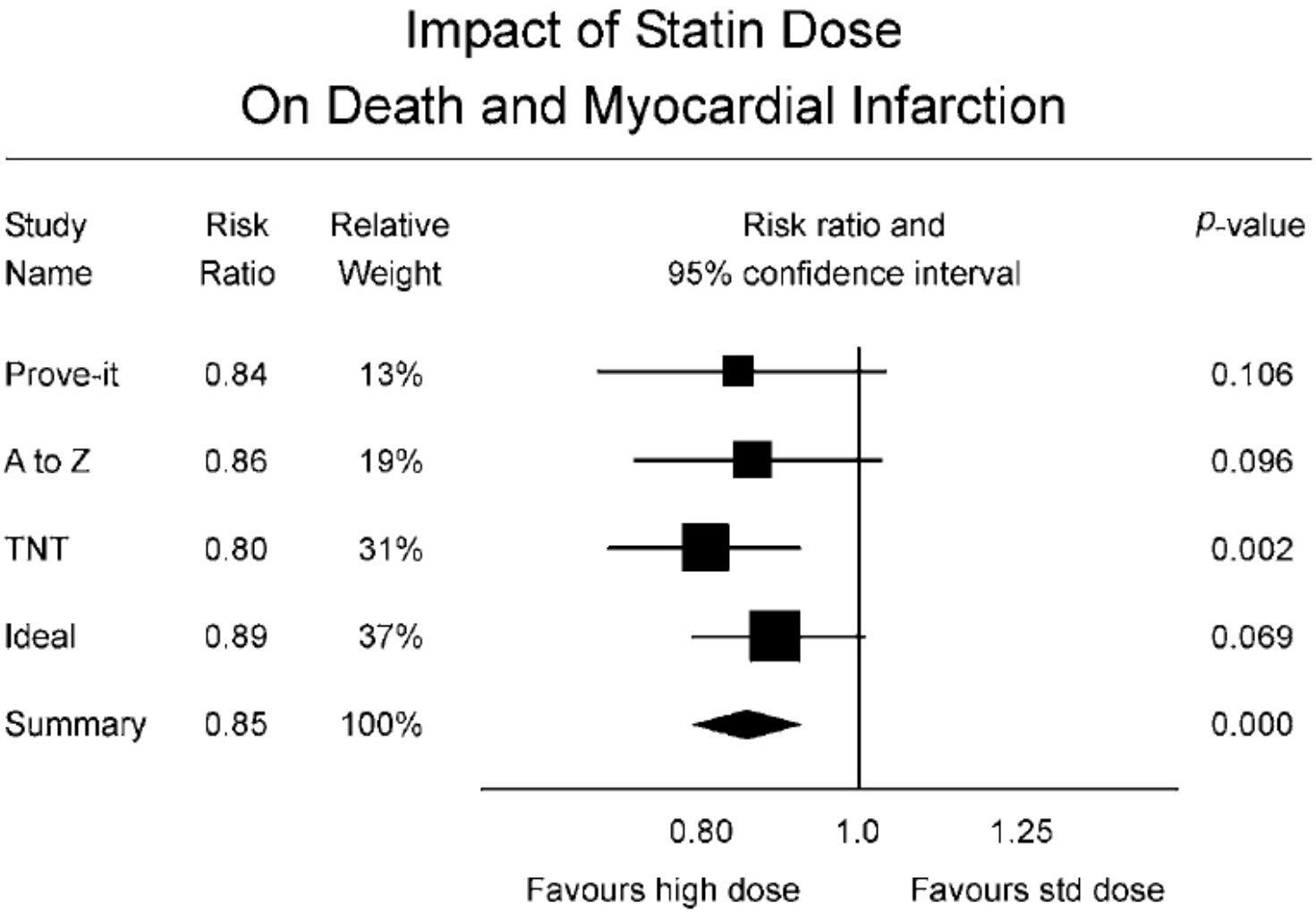
The weight of study i, w_i, is inversely proportional to sampling variance, and proportional to sample size N_i:

$$w_i = \frac{1}{V_i}, V_i \propto 1/N_i \Rightarrow w_i \propto N_i$$

This is called **Inverse Variance Weighting**.

The FE estimator $\hat{\mu} = \sum_i w_i Y_i / \sum_i w_i$, can also be used to derive the standard error of $\hat{\mu}$.
Hint: $\hat{\mu}$ is a linear combination of normally distributed random variables.

FE meta-analysis improves the statistical power: an example



Random effect models relaxes the assumption of identical effect sizes

Example 1. A pharamaceutial company tests the drug effect on 1000 patients with 10 studies. All studies were identical in patient pool, drug dosage, procedure, researchers and so on.

Example 2. 10 different studies were performed independently by ten different groups to assess the effect of a drug. The studies could be different in patient pool, procedure, etc.

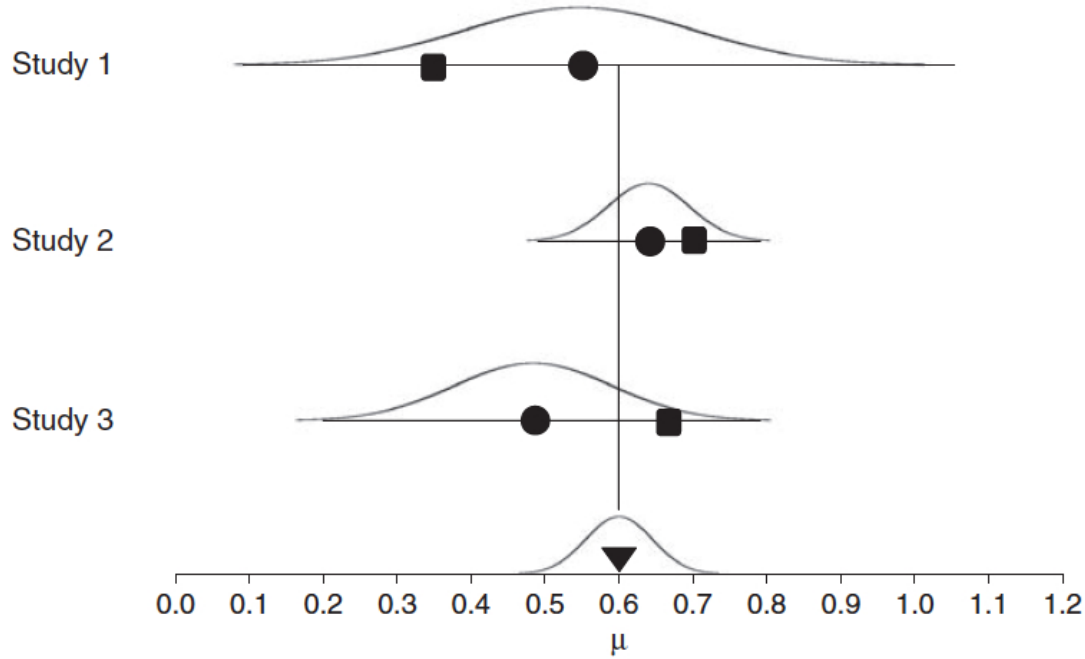
Random effect models relaxes the assumption of identical effect sizes

Example 1. A pharamaceutial company tests the drug effect on 1000 patients with 10 studies. All studies were identical in patient pool, drug dosage, procedure, researchers and so on.

Example 2. 10 different studies were performed independently by ten different groups to assess the effect of a drug. The studies could be different in patient pool, procedure, etc.

Fixed Effects model assumes *identical true effect sizes*, and **Random Effects Model** assumes the effects are similar but could be somewhat different.

Random Effects (RE) Model assumes effect sizes follow a common distribution



Assumption 1. True effect sizes of different studies (circles) can be different, but are sampled from a common distribution.

Assumption 2. The observed effect size (square) in each study varies from the true effect because of sampling error.

The estimated effect under RE Model is still a weighted average across studies

Model: let Y_i be the effect size of study i, and V_i its variance. Let μ_i be the true effect of study i. Our model:

$$Y_i|\mu_i \sim N(\mu_i, V_i) \qquad \mu_i \sim N(\mu, \tau^2)$$

The estimated effect under RE Model is still a weighted average across studies

Model: let Y_i be the effect size of study i, and V_i its variance. Let μ_i be the true effect of study i. Our model:

$$Y_i|\mu_i \sim N(\mu_i, V_i) \qquad \mu_i \sim N(\mu, \tau^2)$$

Define study weight as:

$$w_i^* = 1/(V_i + \tau^2)$$

The MLE of μ is given by:

$$\hat{\mu} = \frac{\sum_i w_i^* Y_i}{\sum_i w_i^*}$$

τ^2 can be obtain by MLE, or from other estimators. It is a measure of *heterogeneity* of studies.

Weights are more balanced under RE model

Comparison of weighting schemes:

$$w_i^{\text{fixed}} = 1/V_i \qquad w_i^{\text{random}} = 1/(V_i + \tau^2)$$

Weights are more balanced under RE model

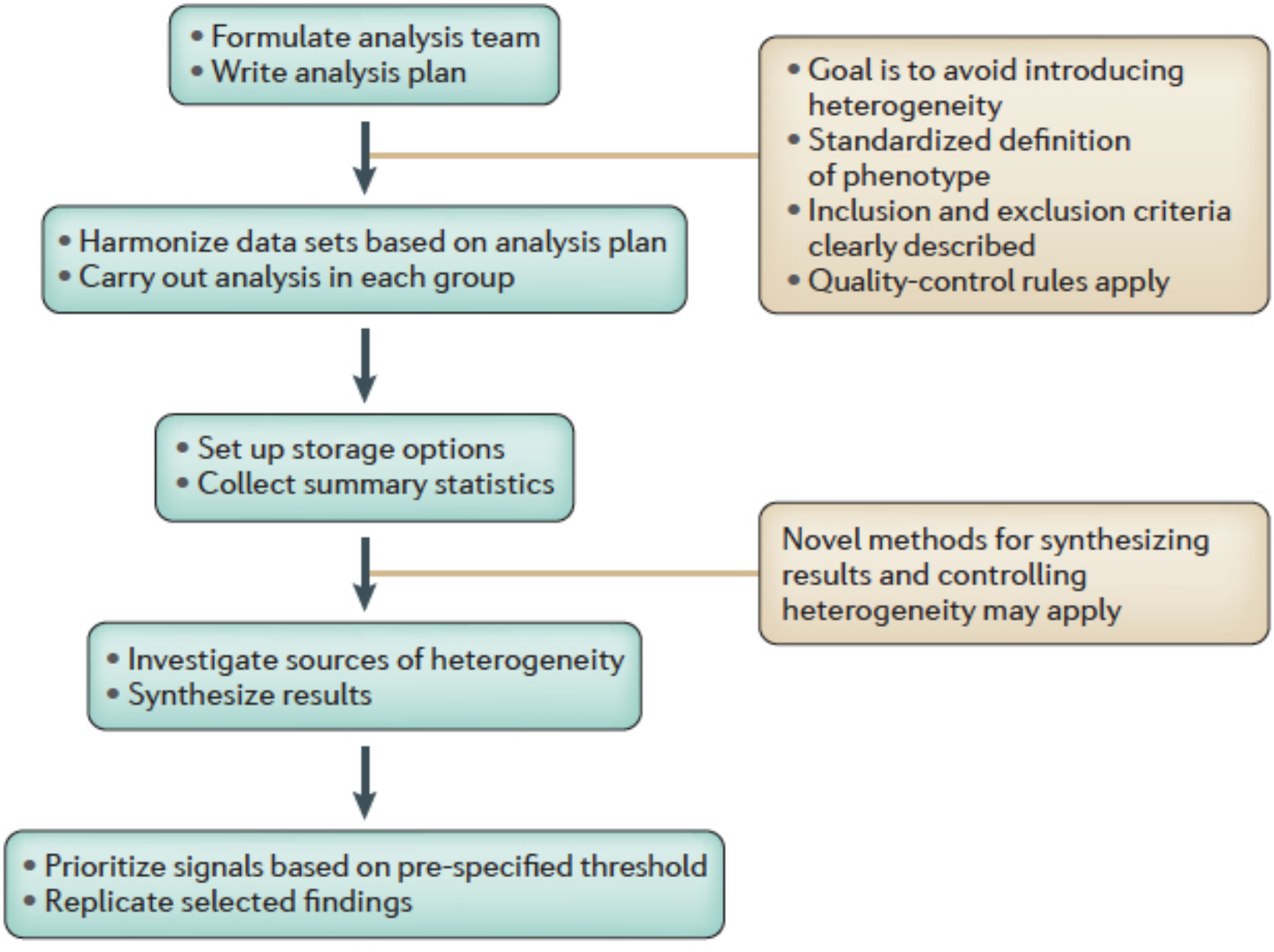
Comparison of weighting schemes:

$$w_i^{\text{fixed}} = 1/V_i \qquad w_i^{\text{random}} = 1/(V_i + \tau^2)$$

Under **fixed effects model**, the weight of a study is entirely determined by sample sizes, thus we can largely ignore information from smaller studies.

Under **random effects model**, each study provides some information of the population effect ($\mu_i \sim N(\mu, \tau^2)$), so smaller studies under the RE model make larger contributions than under the FE model.

GWAS meta-analysis workflow



GWAS Meta-Analysis often uses Fixed Effects model

Inverse variance weighting: $w_i = 1/V_i \propto N_i$.

Summary effect:

$$\hat{\mu} = \frac{\sum_i w_i Y_i}{\sum_i w_i}$$

Significance of summary effect: Z-score based on the standard error of the summary effect.

Most GWAS meta-analysis use Fixed Effect Model.

Consider RE model when the studies are very heterogeneous

Common sources of heterogeneity:

- Variation in phenotype definition. Ex. mental disorders are harder to define and standardized.
- Variation of ancestry across studies. The LD between the causal and tag SNPs may be different across different populations.
- Other difference between studies: environmental exposure, sex difference.

Consider RE model when the studies are very heterogeneous

Common sources of heterogeneity:

- Variation in phenotype definition. Ex. mental disorders are harder to define and standardized.
- Variation of ancestry across studies. The LD between the causal and tag SNPs may be different across different populations.
- Other difference between studies: environmental exposure, sex difference.

However, using RE model reduces the power compared with FE model, when there is not much heterogeneity.

Summary: Meta-analysis

- Meta-analysis uses estimated effect sizes, and is generally better than the use of p-values.
- FE model: inverse variance weighting.
- RE model: better capture heterogeneity of effect sizes.
- In GWAS: FE analysis is the most popular, but consider the RE model when the studies are heterogeneous.

Single causal variant can drive association signals in multiple SNPs

Causal variant or risk variant: the true disease-causing variant.

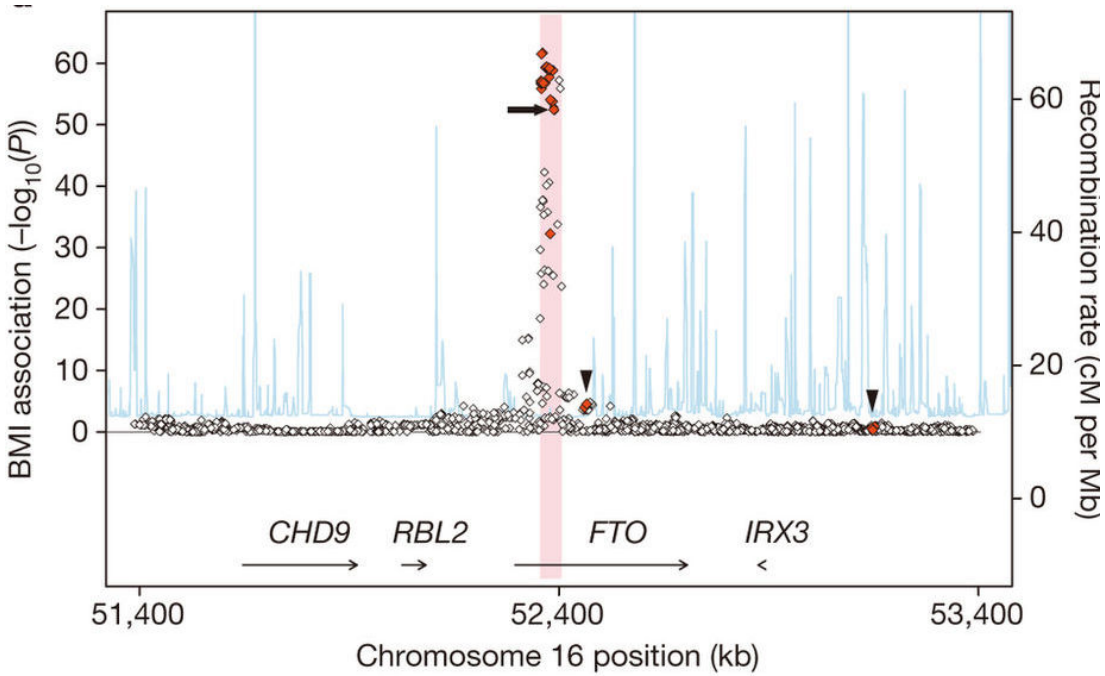


Figure: The FTO locus of obesity (BMI). Arrows shows the likely causal SNP: eQTL of a nearby gene IRX3 [Smemo et al, Nature, 2014]

Lead SNPs are often not causal SNPs

Because of sampling errors: a SNP in close LD with the causal SNP may have similar or even better 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-OR < 1.1.

Some regions may have multiple causal SNPs

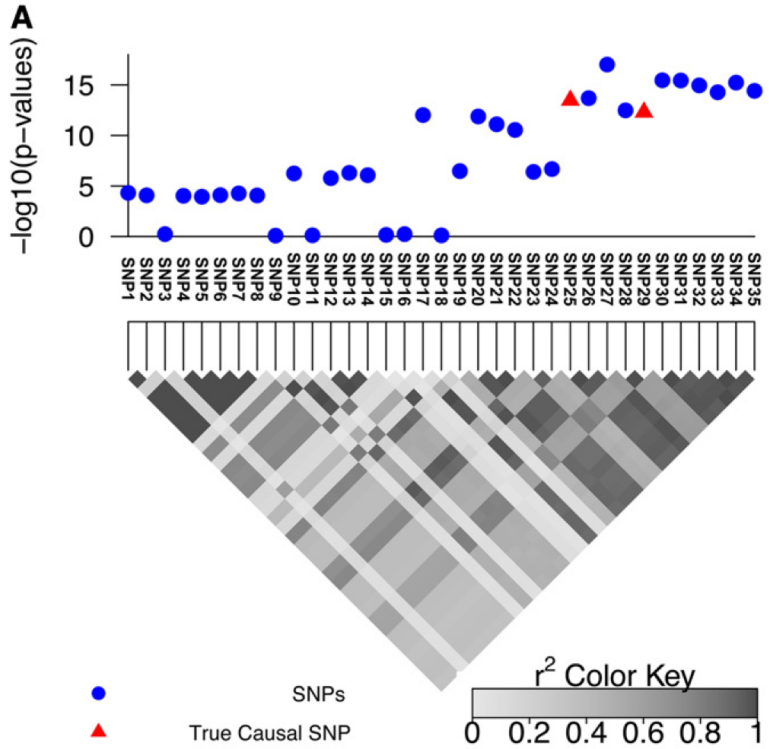
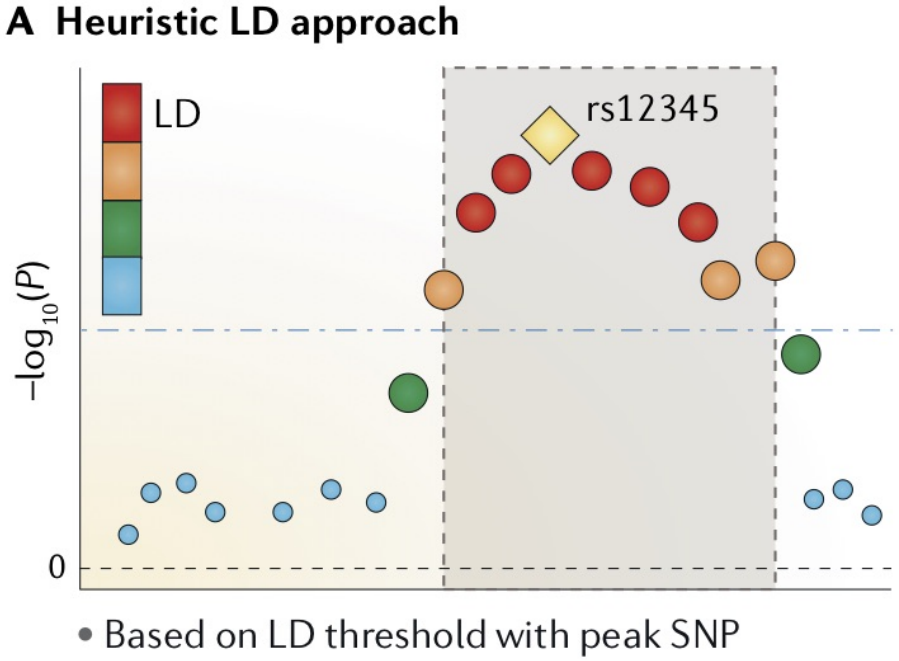


Figure: Simulated data with 2 causal variants (red) [Hormozdiari et al, Genetics, 2014]

Using the top SNP may get the wrong causal SNP, and miss additional signals(s).

Heuristic approach is to take lead SNPs and nearby ones



Limitation: the LD threshold is arbitrary; may have many nearby SNPs; not quantify the statistical evidence.

Conditional regression is often used to find multiple causal signals

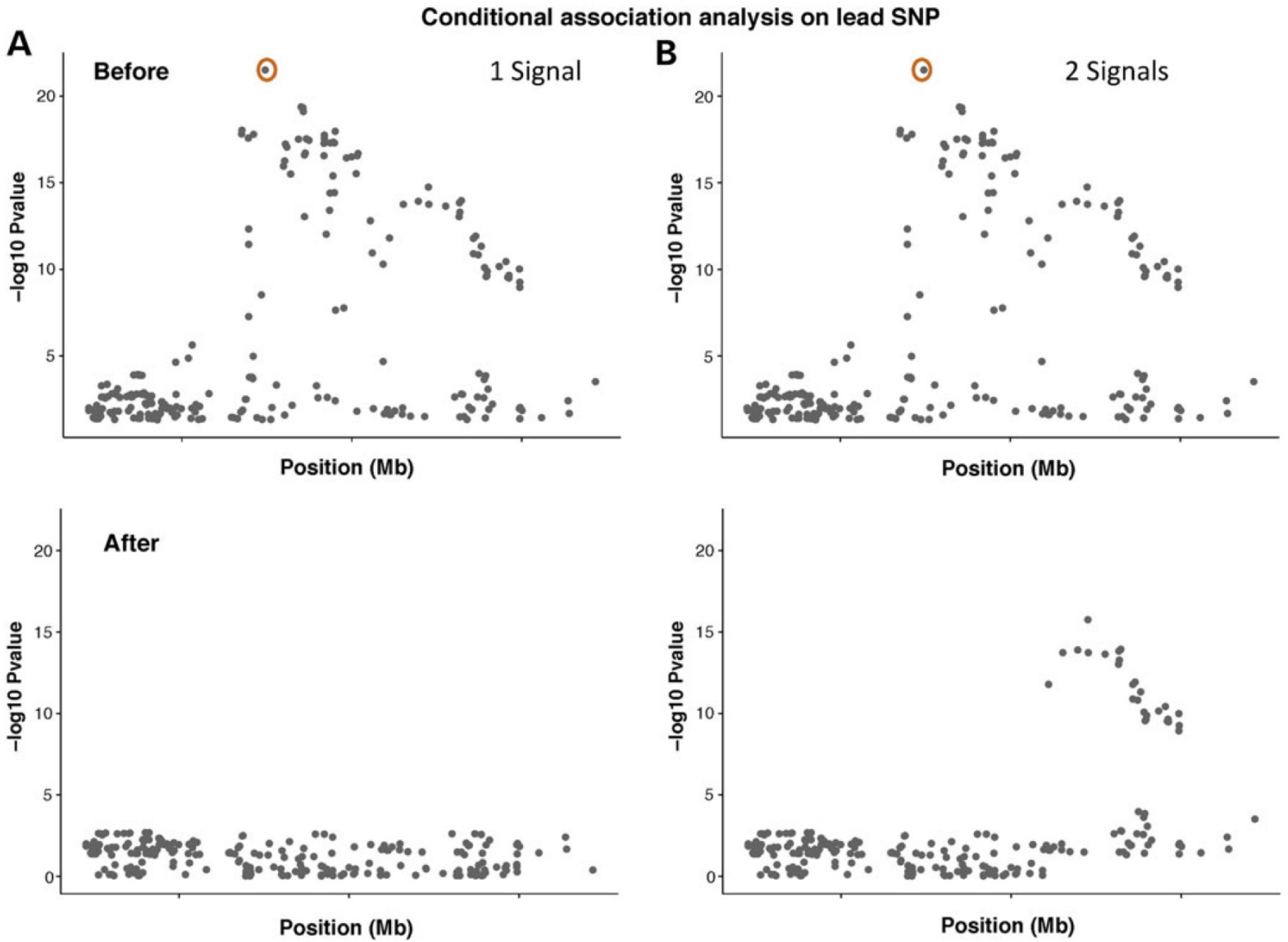
Regression analysis on a SNP j by conditioning on the lead SNP:

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

A SNP passing the threshold will be chosen.

Repeat this analysis: at each step, condition on all SNPs chosen at previous steps.

Example of conditional regression approach



Conditional regression approach cannot guarantee to find causal variants

- 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.

What if we apply conditional regression to the example in Slide 24?

Fine-mapping causal variants is a multiple regression problem

Regression model: let y_i be the phenotype of sample i, $1 \leq i \leq N$ and X_{ij} the genotype of SNP j. The phenotype is related to the genotypes by:

$$y_i = \sum_{j=1}^{p} X_{ij}\beta_j + \epsilon_i = \underbrace{X_i}_{\text{Genotype vector}} \beta + \epsilon_i, \ \epsilon_i \sim N(0, \sigma^2).$$

Our intuition is that for most SNPs, $\beta_j = 0$. So finding causal variants is equivalent to find one or few SNPs with $\beta_j \neq 0$.

Common regression methods are not ideal for fine-mapping

How do we infer β_j's?

- Standard multiple regression using least square estimator (MLE): difficult to apply because the number of variants are large and often highly correlated.
- Shrinkage methods (Lasso): cannot account for uncertainty. Ex. when two variants are highly correlated, Lasso will randomly choose one.

Introducing indicator variables helps identify causal variants

Indicator variables: let $\gamma_j \in \{0, 1\}$ be the indicator of whether SNP j is causal. $\gamma_j = 0$ implies $\beta_j = 0$. γ_j's are random variables, to be inferred from the data.

Causal configuration: $\gamma \in \{0, 1\}^p$ represents the status of all p SNPs.
Ex. a region with 4 SNPs, $\gamma = \{0, 1, 0, 0\}$ or $\{0, 0, 0, 1\}$.

Introducing indicator variables helps identify causal variants

Indicator variables: let $\gamma_j \in \{0, 1\}$ be the indicator of whether SNP j is causal. $\gamma_j = 0$ implies $\beta_j = 0$. γ_j's are random variables, to be inferred from the data.

Causal configuration: $\gamma \in \{0, 1\}^p$ represents the status of all p SNPs.
Ex. a region with 4 SNPs, $\gamma = \{0, 1, 0, 0\}$ or $\{0, 0, 0, 1\}$.

Problem: given data $D = \{X_i, y_i, 1 \leq i \leq N\}$, infer $P(\gamma|D)$, or $P(\gamma_j = 1|D)$ for each SNP j.

Obtaining the posterior distribution of γ

From the Bayes rule:

$$P(\gamma|D) = \frac{P(\gamma)P(D|\gamma)}{P(D)} \propto \underbrace{P(\gamma)}_{\text{Prior}} \cdot \underbrace{P(D|\gamma)}_{\text{Likelihood}}$$

Prior: each SNP has a small prior probability of being a causal variant, $\gamma_j \sim \text{Bernoulli}(\pi)$, then $P(\gamma) = \prod_j \pi^{\gamma_j}(1-\pi)^{1-\gamma_j}$.

Likelihood: assess how well the causal variant configuration explains the data.

Details of likelihood computation (advanced materials)

Our regression model is defined in terms of β, the effect sizes, so we need to relate the causal configuration γ to β. We assume:

$$\beta_j = 0 \text{ if } \gamma_j = 0; \qquad \beta_j \sim N(0, \gamma_\beta^2) \text{ if } \gamma_j = 1.$$

Details of likelihood computation (advanced materials)

Our regression model is defined in terms of β, the effect sizes, so we need to relate the causal configuration γ to β. We assume:

$$\beta_j = 0 \text{ if } \gamma_j = 0; \qquad \beta_j \sim N(0, \gamma_\beta^2) \text{ if } \gamma_j = 1.$$

Likelihood or model evidence of γ needs to marginalize β:

$$P(D|\gamma) = \int P(D|\beta) \cdot P(\beta|\gamma) d\beta$$

$P(D|\beta)$ is given by the standard linear regression

$$P(D|\beta) = P(\mathbf{Y}|\mathbf{X}, \beta) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_i (y_i - X_i\beta)^2\right]$$

Finding best configurations is computationally challenge

In a region with p SNPs, the number of configurations is 2^p.

Possible approaches:

- Enumeration of configurations: up to maximum number of causal SNPs. CAVIER.
- Stochastic search: FINEMAP.
- Variational Bayes: similar to conditional regression, but accounts for the uncertainty at each step. SuSiE.

Results of fine-mapping are summarized as posterior inclusion probabilities (PIPs)

PIP: the posterior probability that a SNP is causal, summing over all possible models. Often used to rank SNPs.

$$P(\gamma_j = 1|D) = \sum_{\gamma} \underbrace{[\gamma_j = 1]}_{\text{1 if } \gamma_j\text{=1;0 otherwise}} \cdot P(\gamma|D)$$

PIPs depend on both GWAS statistics and LD patterns:

- In a high-confidence region with a single causal variant, PIPs of all SNPs should sum to 1.
- A high-confidence region with a single causal SNP: suppose the causal SNP is in high LD with $k-1$ other SNPs, then PIP of each SNP $\approx 1/k$.

Bayesian credible set with a single causal variant

Credible set: the minimum set of variant that contains the causal variant with probability α (typically, 95%).

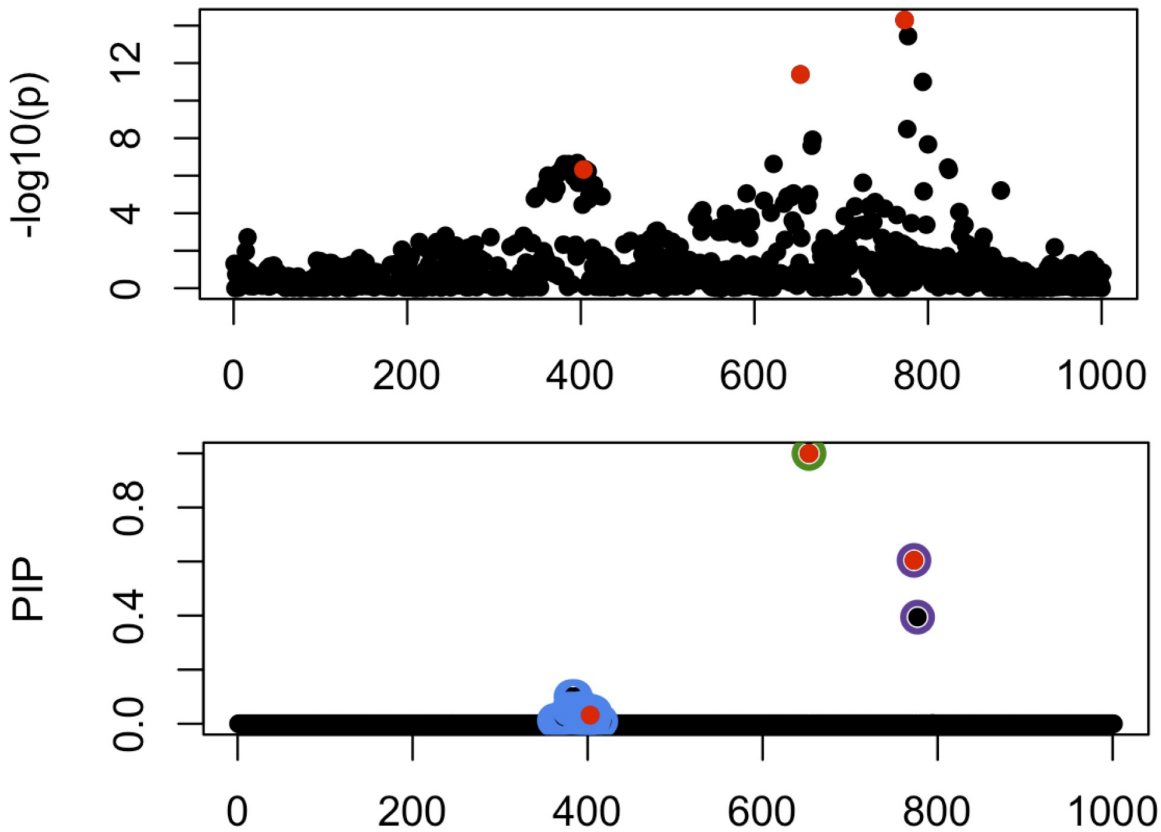
Define the "confidence level" of a variant set S as the total posterior probability of all models allowed under the set S. Ex. $S = \{A, B, C\}$. The "confidence level" of the set is simply:

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

If $\rho \geq \alpha$, and the set cannot be made smaller, it's a credible set.

Bayesian credible set with multiple causal variants

Different definitions have been used in literature. We use the one based on **SuSiE**: each causal variant has a credible set. So a locus may have multiple credible sets, each capturing one signal.



Summary: Fine-mapping

- Heuristic approach and conditional regression have limitations.
- Bayesian approach to fine-mapping: infer the posterior of causal configurations.
- Results of fine-mapping: PIPs and credible sets.

Recommended Readings

- Borenstein et al (2009) Introduction to Meta-Analysis, Wiley Press. Chapters 1-2, 10-13
- Evangelou & Ioannidis (2013) Meta-analysis methods for genome-wide association studies and beyond, *Nat Rev Genet*, 14(6):379
- Schaid et al (2018) From genome-wide associations to candidate causal variants by statistical fine-mapping, *Nat Rev Genet*