Genome-wide Association Studies

BIOS 770

02/19/2025

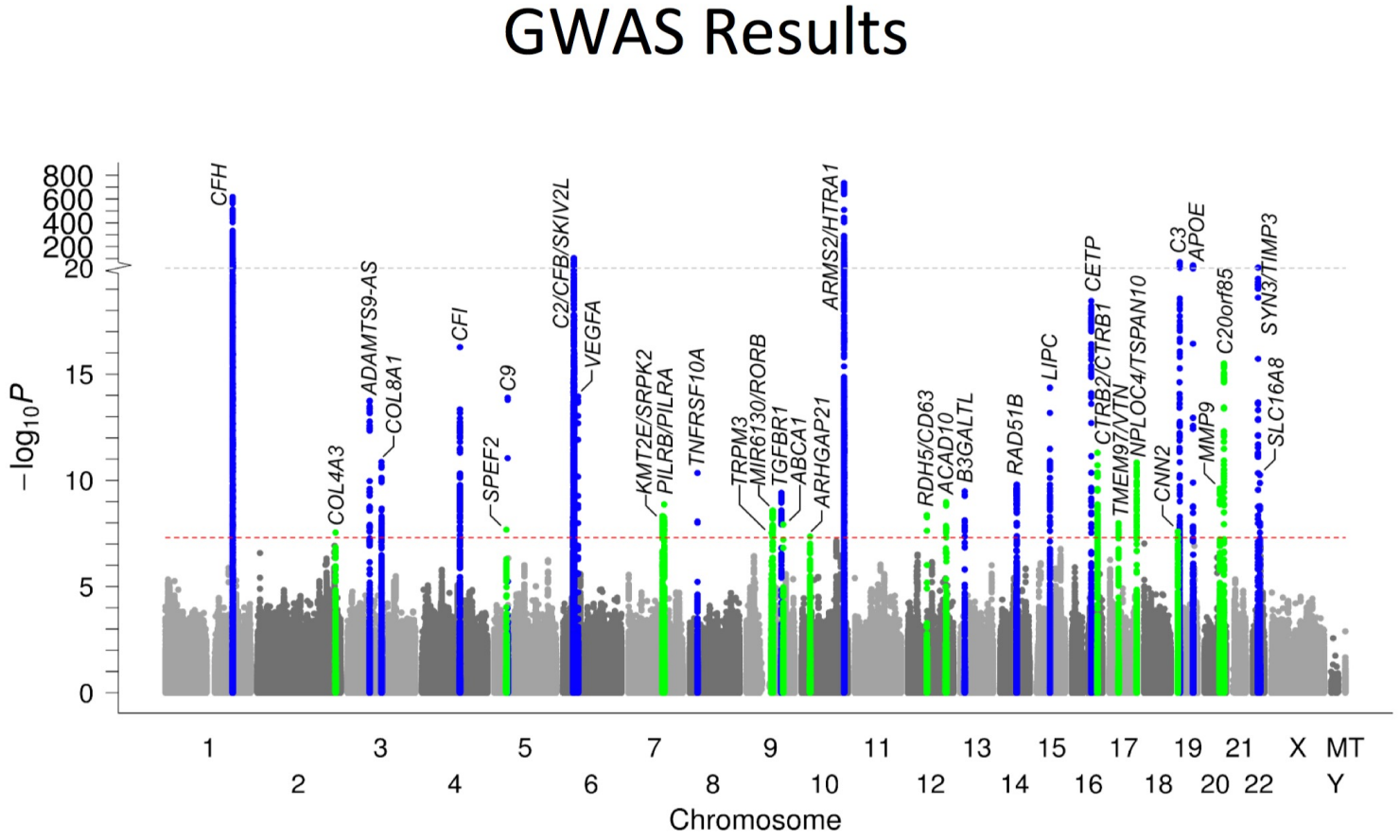
Jingjing Yang (jingjing.yang@emory.edu)

Outline

- Fine-map GWAS Results
  - Conditional Analysis
  - Bayesian Methods: FINEMAP
  - Based on the “Sum of Single Effects” model: SuSiE

- Bayesian Functional GWAS

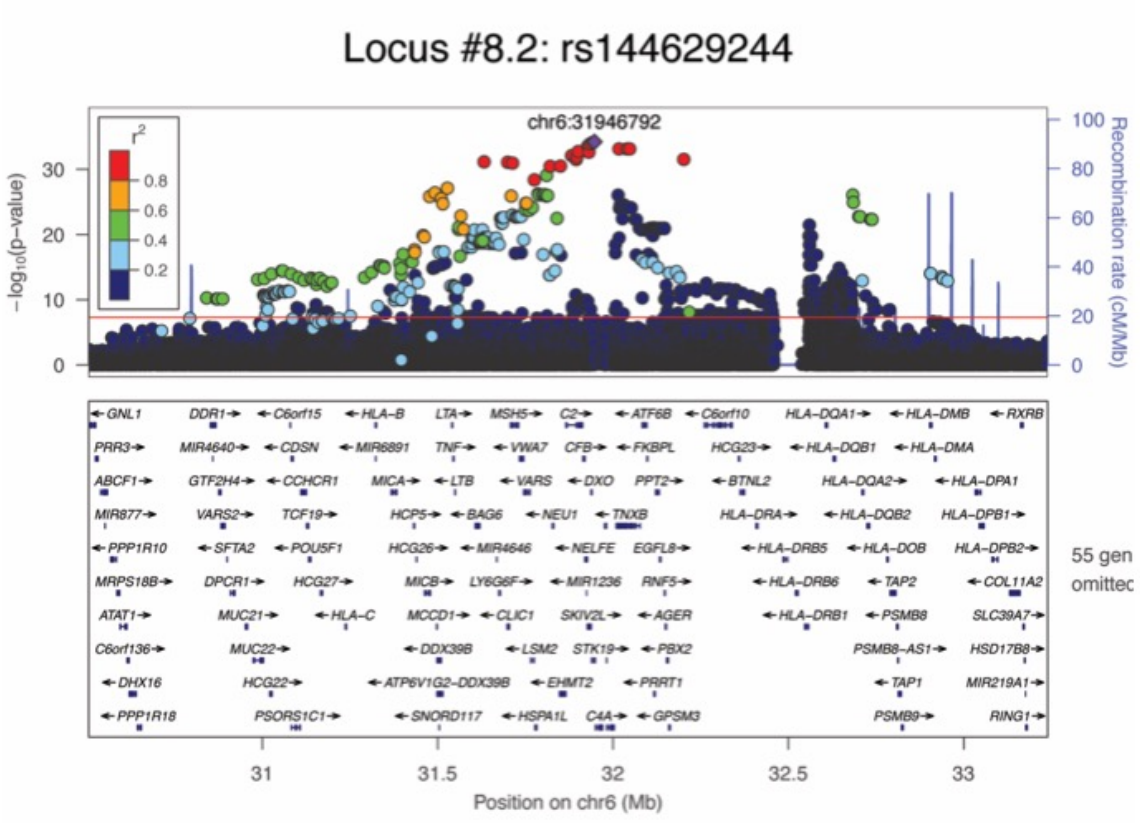
Fine-mapping GWAS Results



18 known AMD loci and **16 novel AMD loci**

Visualize GWAS Loci by Locus Zoom Plot

– Zoom into the peak region with gene annotations
– Visualize r^2 between the specified significant (purple diamond) signal and its neighbor SNPs
– Visualize recombination rate



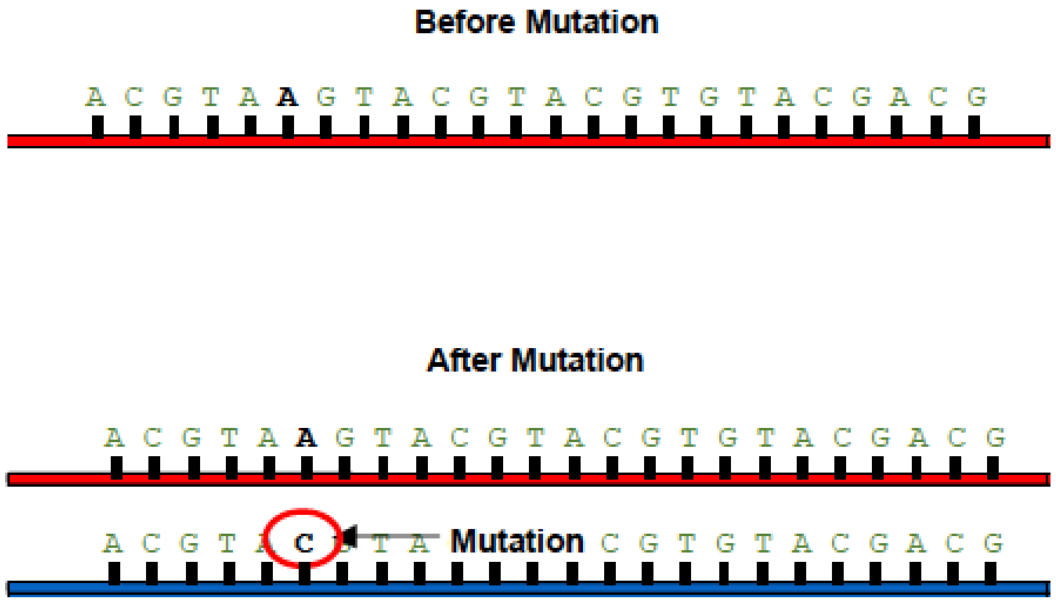
Fritsche L.G. et al. Nat Genet, 2016.

Linkage Disequilibrium (LD)

- Definition in Population Genetics
  - **Linkage Disequilibrium (LD)** is the **non-random association** of alleles at different loci in a given population.
- Why nearby markers are likely to be correlated?
- The origin of LD?

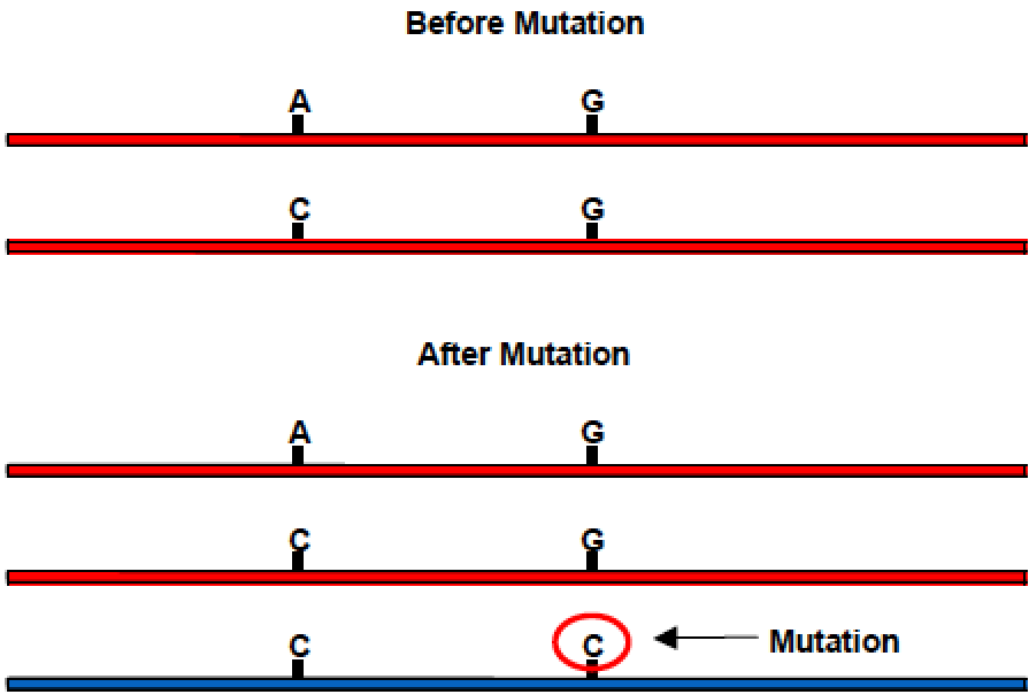
Linkage Disequilibrium (LD)

- Consider the history of two neighboring single nucleotide polymorphism (SNP)
- SNPs exist today arose through ancient mutation events…



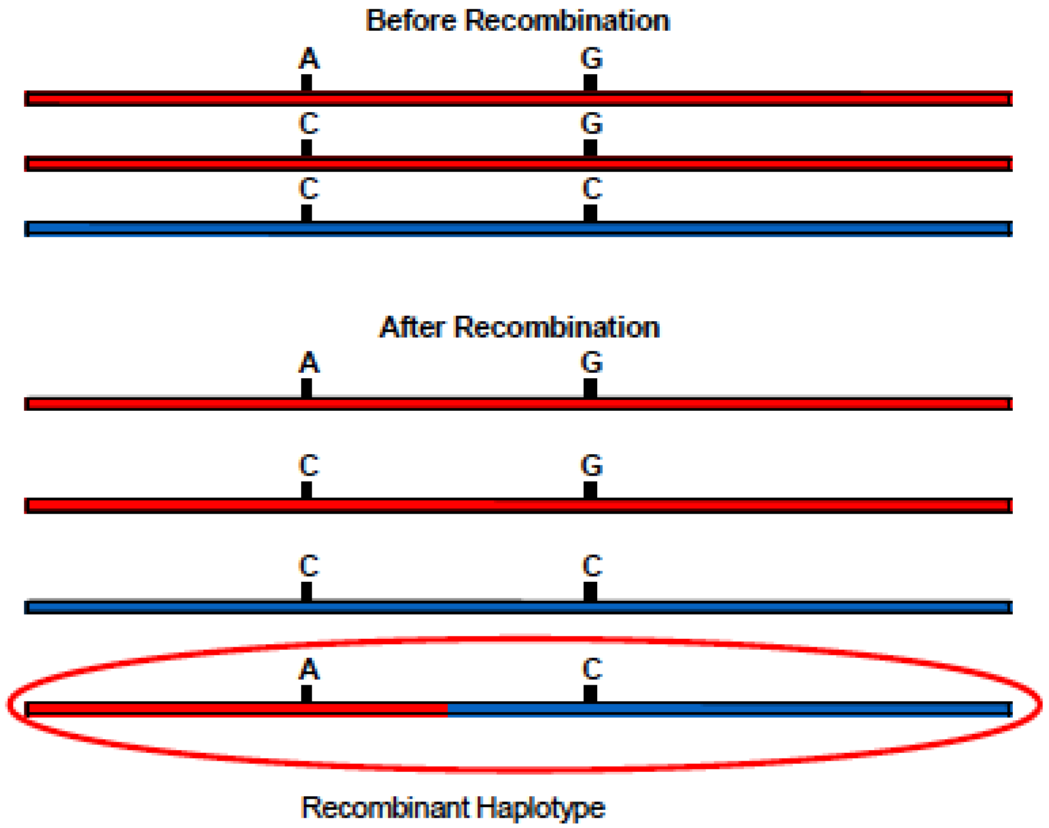
Linkage Disequilibrium (LD)

- One SNP arose first and then the other …



Linkage Disequilibrium (LD)

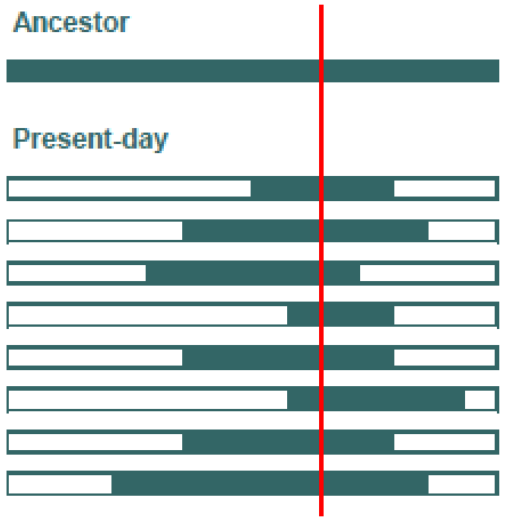
- Recombination generates new arrangements for the ancestral alleles



Linkage Disequilibrium (LD)

- Chromosomes are mosaics

- Extent and conservation of mosaic pieces depends on
  - Recombination rate
  - Mutation rate
  - Population size
  - Natural selection



- Combinations of alleles at very close markers reflect ancestral haplotypes

Quantify Linkage Disequilibrium (LD)

- **LD** is defined as the **difference** between the **observed frequency** of a particular combination of alleles at two loci and the **frequency expected** for random association.
- Allele frequency
  - $P_A, P_a, P_A + P_a = 1$
  - $P_B, P_b, P_B + P_b = 1$
  - $P_{AB} = P_A P_B$
    if and only if alleles A, B are independent
- Minor Allele Frequency (MAF)

		Locus B		Totals
		B	b	
Locus A	A	p_{AB}	p_{Ab}	p_A
	a	p_{aB}	p_{ab}	p_a
Totals		p_B	p_b	1.0

Linkage Equilibrium
Expected for Distant Loci

$$p_{AB} = p_A p_B$$

$$p_{Ab} = p_A p_b = p_A(1 - p_B)$$

$$p_{aB} = p_a p_B = (1 - p_A) p_B$$

$$p_{ab} = p_a p_b = (1 - p_A)(1 - p_B)$$

Linkage Disequilibrium
Expected for Nearby Loci

$$p_{AB} \neq p_A p_B$$

$$p_{Ab} \neq p_A p_b = p_A(1 - p_B)$$

$$p_{aB} \neq p_a p_B = (1 - p_A) p_B$$

$$p_{ab} \neq p_a p_b = (1 - p_A)(1 - p_B)$$

Disequilibrium Coefficient D_{AB}

$$D_{AB} = p_{AB} - p_A p_B$$

$$p_{AB} = p_A p_B + D_{AB}$$

$$p_{Ab} = p_A p_b - D_{AB}$$

$$p_{aB} = p_a p_B - D_{AB}$$

$$p_{ab} = p_a p_b + D_{AB}$$

D_{AB} is hard to interpret

- Sign is arbitrary …
  - A common convention is to set…
    - *A, B* as the common alleles
    - *a, b* as the rare allele
- Range depends on allele frequencies
  - Hard to compare between markers
- Can you see why the range of D_{AB} depends on allele frequencies?

Boundaries for D_{AB}

- By using the fact that $p_{AB} = P(AB)$ must be less than both $p_A = P(A)$ and $p_B = P(B)$, and that allele frequencies cannot be negative, the following relations can be obtained:
  - $0 \leqslant p_{AB} = p_A p_B + D_{AB} \leqslant p_A, p_B$
  - $0 \leqslant p_{aB} = p_a p_B - D_{AB} \leqslant p_a, p_B$
  - $0 \leqslant p_{Ab} = p_A p_b - D_{AB} \leqslant p_A, p_b$
  - $0 \leqslant p_{ab} = p_a p_b + D_{AB} \leqslant p_a, p_b$
- These inequalities lead to bounds for D_{AB} :

$$-p_A p_B, -p_a p_b \leqslant D_{AB} \leqslant p_a p_B, p_A p_b$$

Normalized Linkage Disequilibrium Coefficient

- The possible values of D depend on allele frequencies. This makes D difficult to interpret. For reporting purposes, the normalized linkage disequilibrium coefficient D' is often used.

$$D'_{AB} = \begin{cases} \frac{D_{AB}}{max(-p_A p_B, -p_a p_b)} & \text{if } D_{AB} < 0 \\ \frac{D_{AB}}{min(p_a p_B, p_A p_b)} & \text{if } D_{AB} > 0 \end{cases} \quad (1)$$

Estimate D_{AB}

- Suppose we have the N haplotypes for two loci on a chromosomes that have been sampled from a population of interest. The data might be arranged in a table such as:

	B	b	Total
A	n_{AB}	n_{Ab}	n_A
a	n_{aB}	n_{ab}	n_a
	n_B	n_b	N

- We would like to estimate D_{AB} from the data. The maximum likelihood estimate of D_{AB} is

$$\hat{D}_{AB} = \hat{p}_{AB} - \hat{p}_A \hat{p}_B$$

where $\hat{p}_{AB} = \frac{n_{AB}}{N}$, $\hat{p}_A = \frac{n_A}{N}$, and $\hat{p}_B = \frac{n_B}{N}$

- So the population frequencies are estimated by the sample frequencies

Measuring LD with r^2

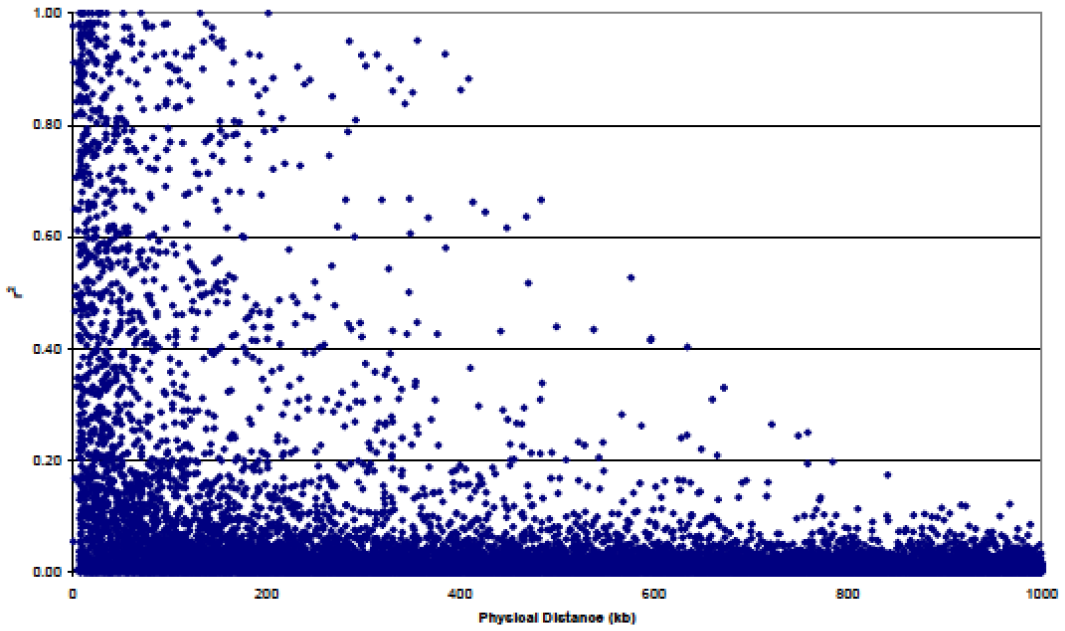
- Define a random variable X_A to be the number allele A at the first marker, with values 0, 1, 2;
- Define a random variable X_B to be the number of allele B at the second marker, with values 0, 1, 2;
- X_A follows a Binomial(2, p_A); X_B follows a Bionomial(2, p_B) distribution.
- Correlation between these two random variables is given by

$$r_{AB} = \frac{Cov(X_A, X_B)}{\sqrt{Var(X_A)Var(X_B)}} = \frac{D_{AB}}{\sqrt{p_A(1-p_A)p_B(1-p_B)}}$$

$$r_{AB}^2 = \frac{D_{AB}^2}{p_A(1-p_A)p_B(1-p_B)} = \frac{D_{AB}^2}{p_A(p_a)p_B(p_b)}$$

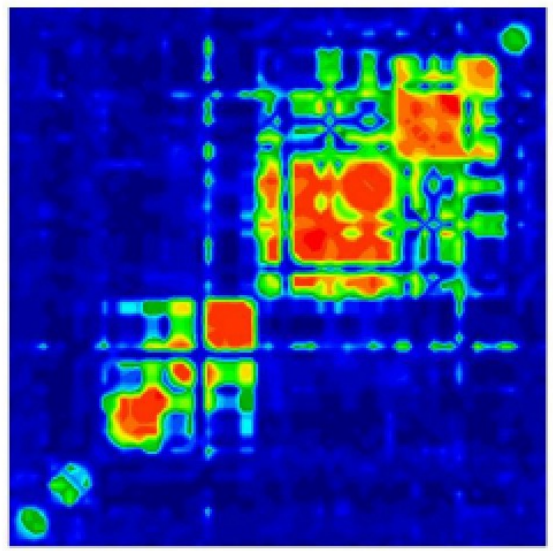
Properties for r^2

- Ranges between 0 and 1
  - 1 means two markers provide identical information, referred to as Perfect LD
  - 0 means two markers are in Perfect Equilibrium
- Raw r^2 from CHR22

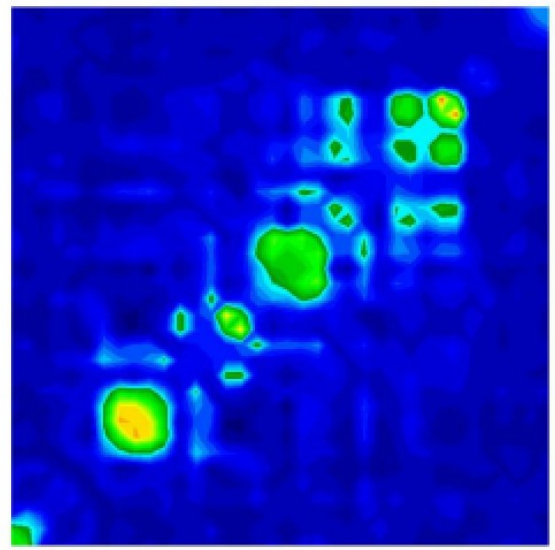


Dawson et al, *Nature*, 2002

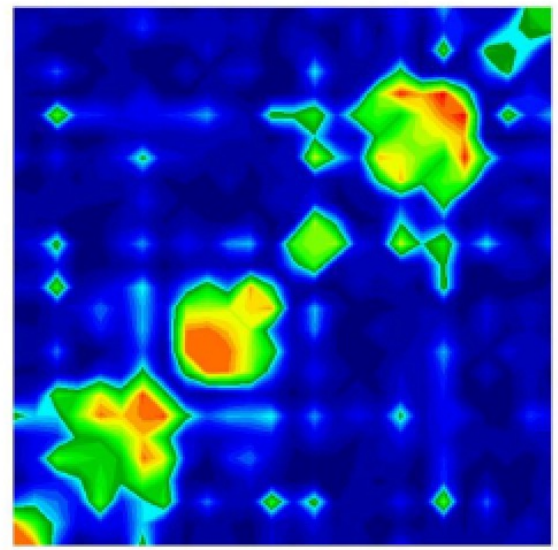
Linkage Disequilibrium in Three Regions



2q13
(63 markers)



13q13
(38 markers)

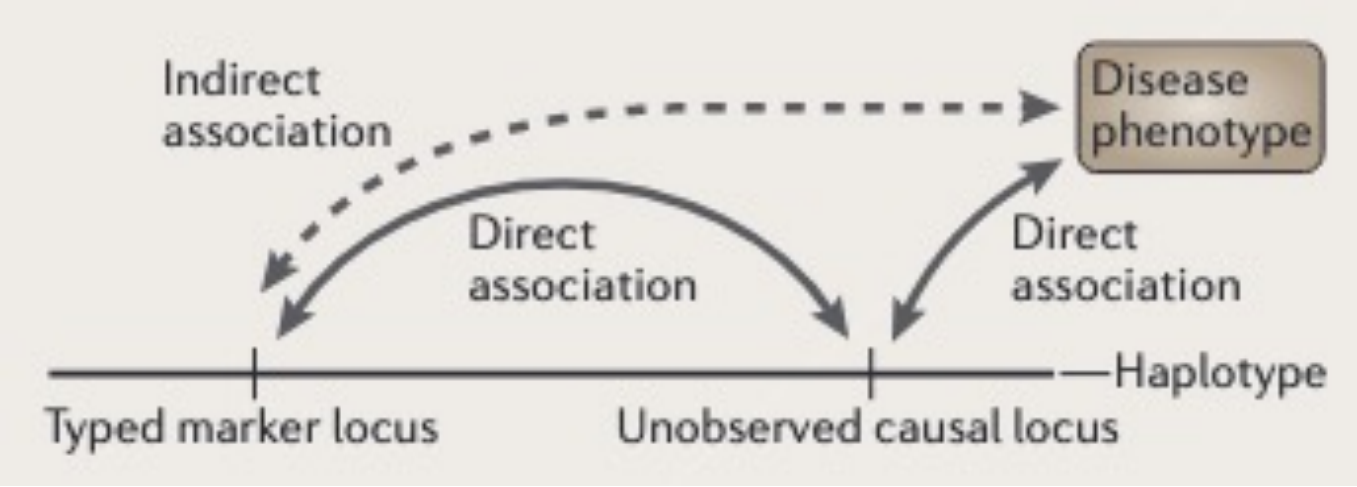


14q11
(26 markers)

Abecasis et al, *Am J Hum Genet*, 2001

Why LD is Important for Association Studies?

- Hypothesis: SNPs in strong LD with disease variant are good proxies for disease variant



Balding, 2006

- If testing (unobservable) disease variant for association would yield chi-squared statistic X^2, testing variant in LD yields r^2X^2 (useful for meta-analysis)

Fine-mapping GWAS Results

- Hypothesis:
  - Only a small number of genetic variants (dozens or hundreds vs. millions) would be true causal variants
- Problem:
  - Most significant GWAS signals, i.e., significant SNPs, are located in non-coding regions
  - All SNPs in LD (i.e., highly correlated) with the nearby most significant GWAS signal are likely to be tested with significant p-values
- Fine-mapping:
  - Pinpointing potential true causal SNPs (true biological molecular mechanisms) from all SNPs that are in LD
  - Conducted per risk locus with significant GWAS signals (region, e.g., +-5KB)

Fine-mapping GWAS Results

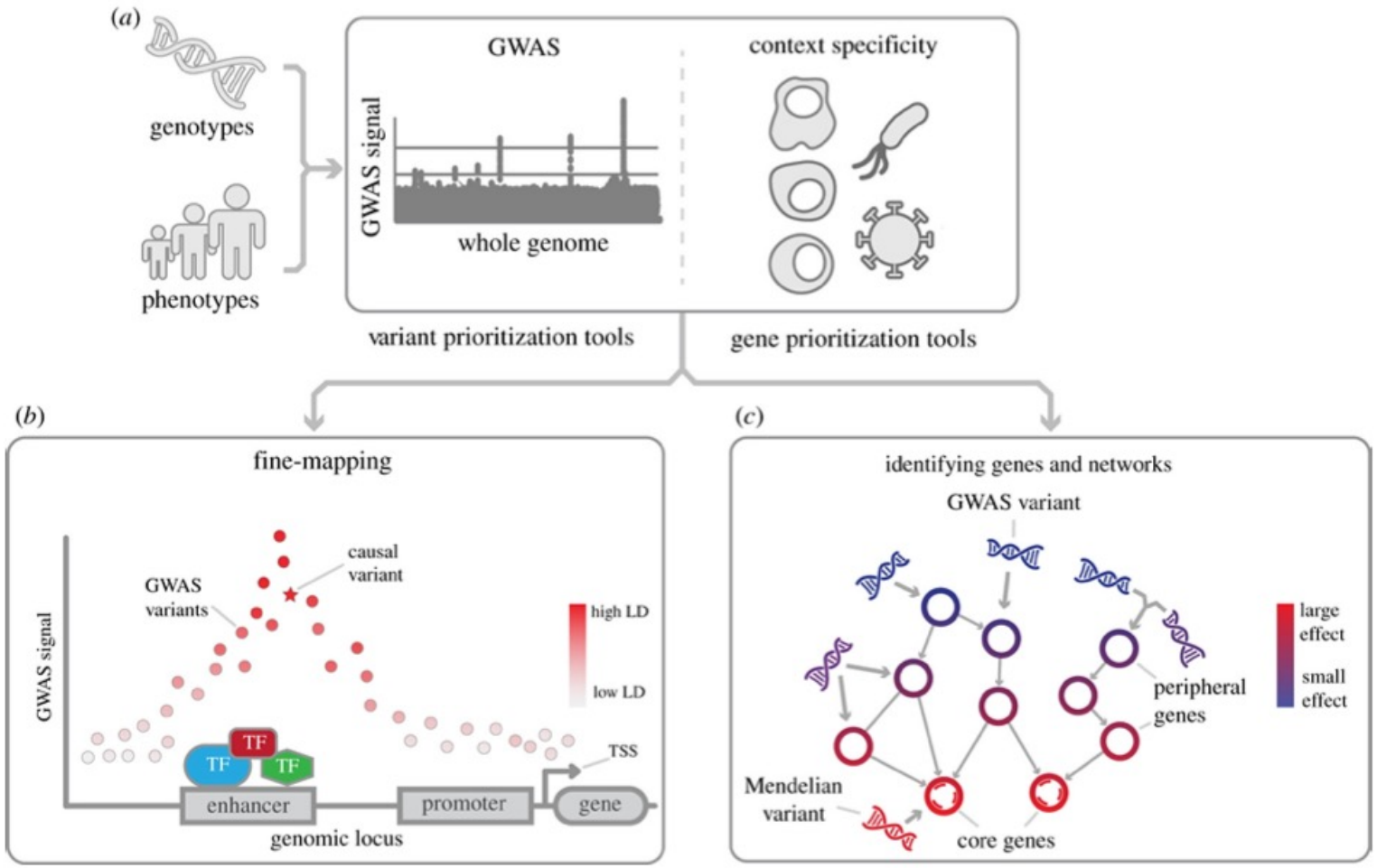


Figure: Broekema *et al.* (2020) Open Biol.

Fine-mapping: Conditional Analysis

Sequential Forward Selection

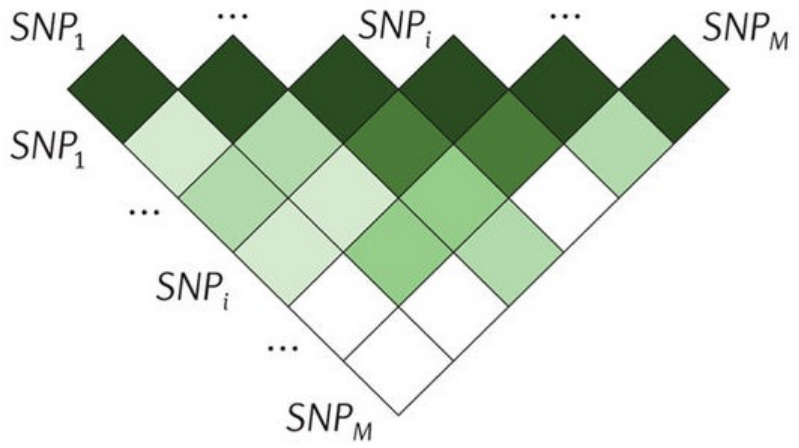
Aim: Within each region of interest, identify all statistically independent variants

1. Select variant with smallest P value ($P < 5x10^{-8}$), write into results file

2. Conduct region-wide association analysis conditioning on variants in results file

3. From the results of 2., if smallest $P < 5x10^{-8}$, select variant write into results file; otherwise stop

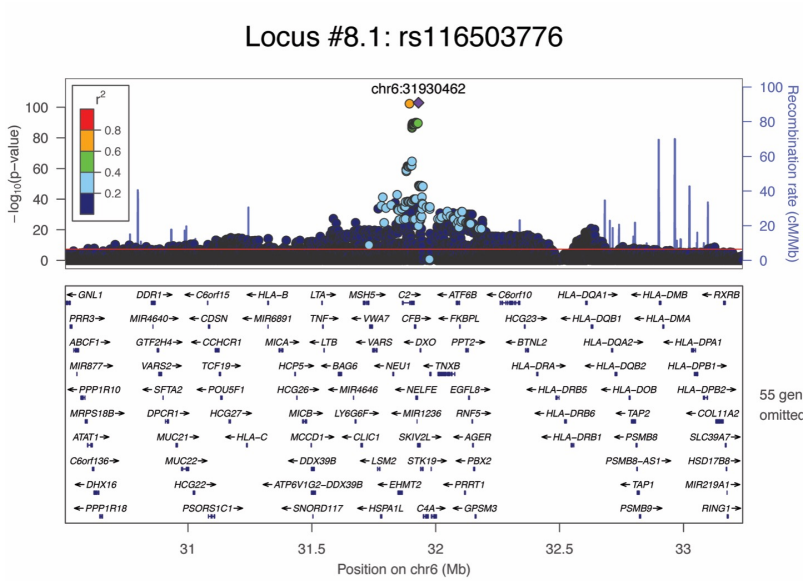
4. Repeat 2. and 3.



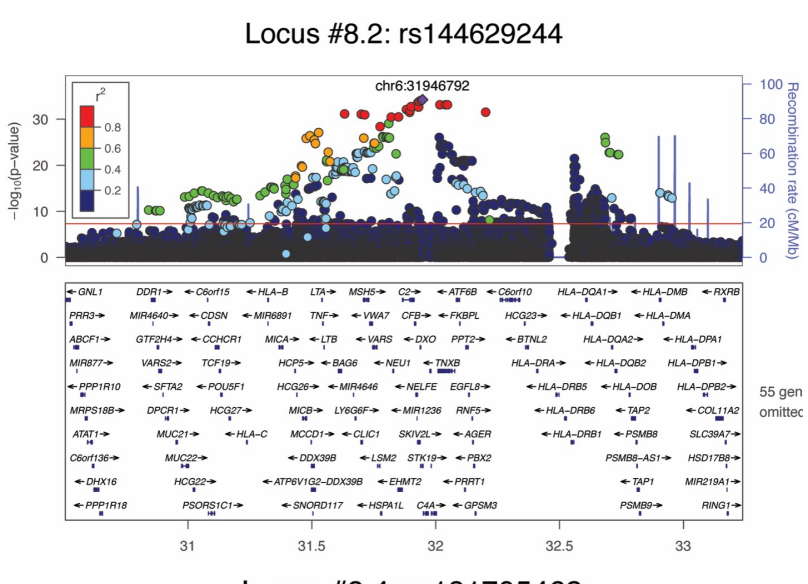
Fritsche L and Pasaniuc B and Price AL, Nat. Rev. 2017

Conditional Analysis

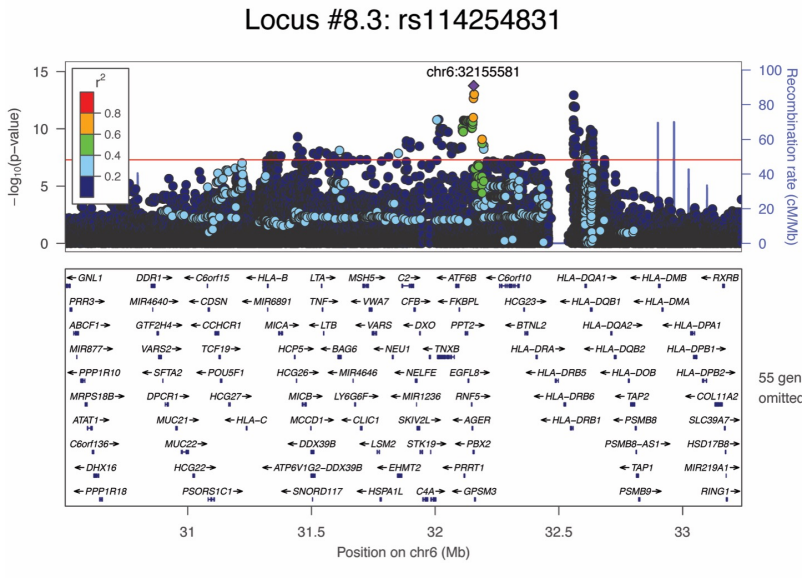
Locus #8.1: rs116503776



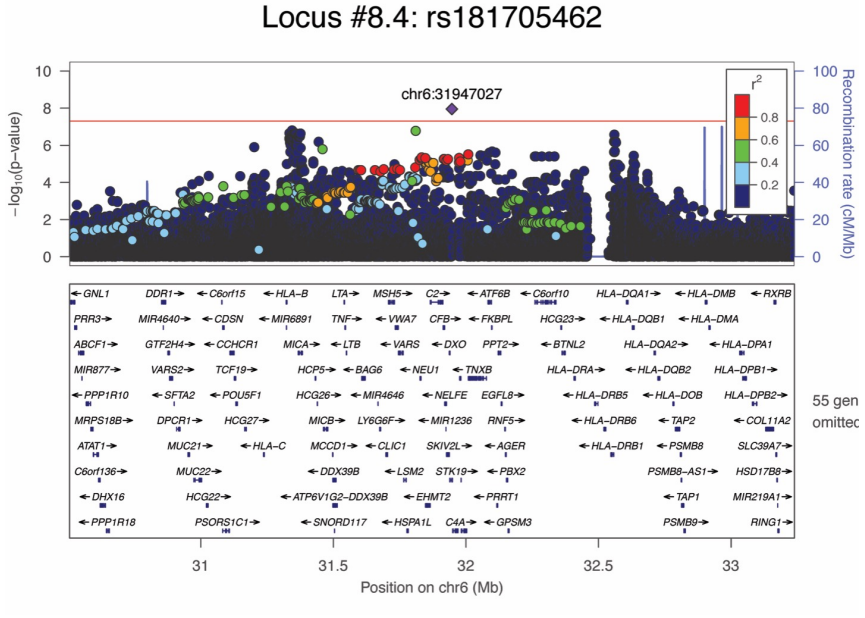
Locus #8.2: rs144629244



Locus #8.3: rs114254831



Locus #8.4: rs181705462



COnditional and **JO**int Analysis (**COJO**) by GCTA

- Implements stepwise variable selection requiring only GWAS summary statistics
  1. Start with a model with the most significant SNP in the single-SNP meta-analysis across the whole genome with P value below a cutoff P value, such as 5×10^{-8} .
  2. For the ith step, calculate the P values of all the remaining SNPs conditional on the SNP(s) that have already been selected in the model. To avoid problems due to collinearity, if the squared multiple correlation between a SNP to be tested and the selected SNP(s) is larger than a cutoff value, such as 0.9, the conditional P value for that SNP will be set to 1.
  3. Select the SNP with minimum conditional P value that is lower than the cutoff P value. However, if adding the new SNP causes new collinearity problems between any of the selected SNPs and the others, we drop the new SNP and repeat this process.
  4. Fit all the selected SNPs jointly in a model and drop the SNP with the largest P value that is greater than the cutoff P value.
  5. Repeat (2), (3) and (4) until no SNPs can be added or removed from the model.
- GCTA software implementation (Jian Yang et al. 2011).

Conditional Analysis

- Informative about the number of complementary sources of association signals within the region
- Fails to provide probabilistic measures of causality for individual variants
- Not accounting for functional annotations (i.e., biological functions) of SNPs

Account for LD in GWAS by Multivariate Regression Model

- Consider the multivariate regression model with all genome-wide variants in the genotype matrix X

$$y = X\beta + \epsilon, \qquad \epsilon \sim N(0, \sigma^2 I)$$

- Variable selection is needed

- Memory and computation issues for ~10 Million SNPs

LASSO : Least Absolute Shrinkage and Selection Operator

- Lasso-penalized least squares objective function

$$Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right|.$$

- Tunning penalty parameter λ (S. Yang et. al., Bioinformatics, 2020).
- Solve for estimates of genetic effect sizes: β
- R function “glmnet()”

Bayesian Variable Selection Regression (BVSR)

- Consider the multivariate regression model with a point-and-spike prior on β_l

$$y = X\beta + \epsilon, \qquad \epsilon \sim N(0, \tau^{-1}I)$$
$$\beta_l \sim \pi N\left(0, \tau^{-1}\sigma_\beta^2\right) + (1-\pi)\delta_0(\beta_l), l = 1, \dots, m$$
$$\sigma_\beta^2 \sim InverseGamma(k_1, k_2), \pi \sim Beta(a, b)$$
$$\tau \sim Gamma(k_3, k_4),$$

- Assume an indicator vector γ, that is,

$$\gamma_l \sim Bernoulli(\pi)$$

- Inference goal: estimate $\beta_l, E[\gamma_l], \sigma_\beta^2, \pi$
- Approach: Monte Carlo Markov Chain (MCMC) (BIMBAM: Guan and Stephens, 2011)
- Convergence would be an issue for studying ~10 Millions SNPs

Parameters of Interest

- π : Genome-wide causal probability for a SNP
- β_l: Genetic effect size of SNP l
- $E[\gamma_l]$: Posterior inclusion probability (PIP), aka, posterior causal probability
- 95% credible set S: Pr(effect variable in S) $\geq$ 95%

BVSR posterior

Assess **combinations** of variables

SNPs	1	2	3	4	5	⋯	Probability
model configurations	1	0	1	0	0	⋯	0.25
	1	0	0	1	0	⋯	0.25
	0	1	1	0	0	⋯	0.25
	0	1	0	1	0	⋯	0.25

- $\text{PIP}_j := Pr(z_j \text{ is non-zero})$

$$\text{PIP} = (0.5, 0.5, 0.5, 0.5, 0, \cdots)$$

BVSR quantifies uncertainty in variable selection.

Assessing multi-effects configurations

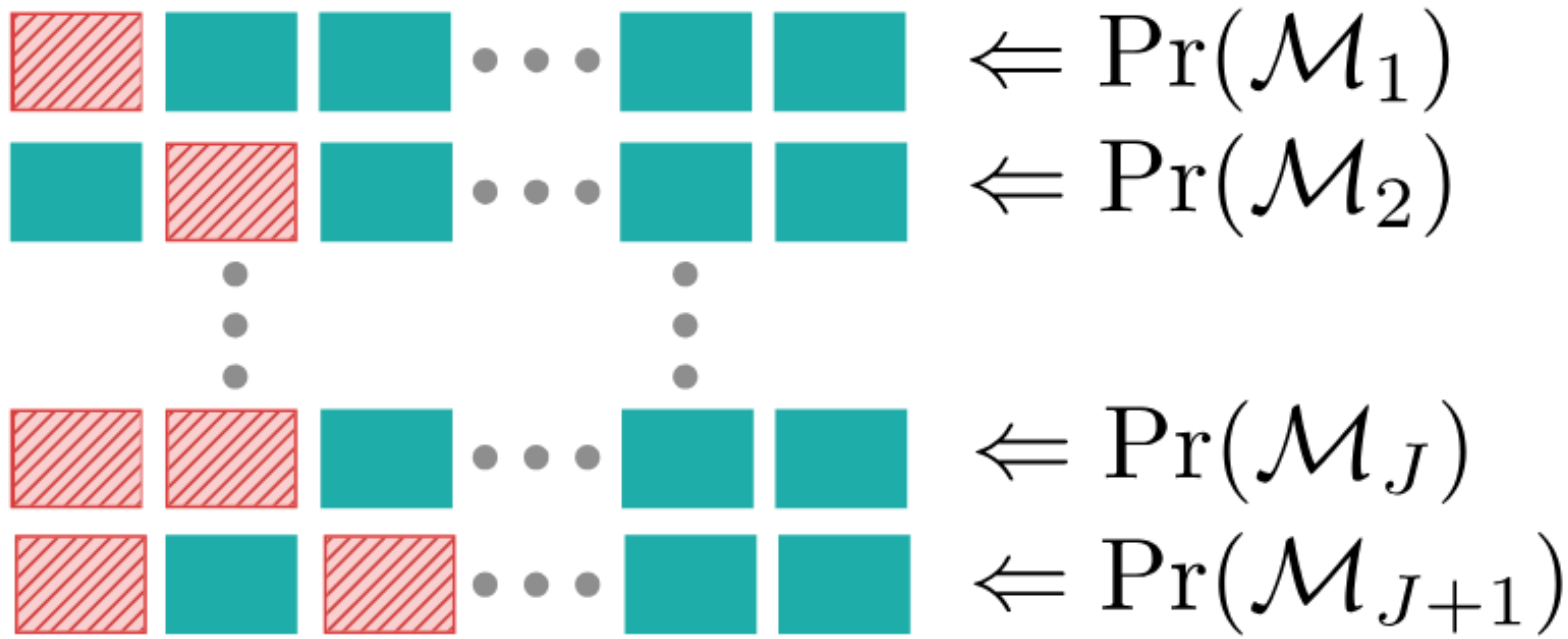
$L = 1$

$\Leftarrow \Pr(\mathcal{M}_1)$

$\Leftarrow \Pr(\mathcal{M}_2)$

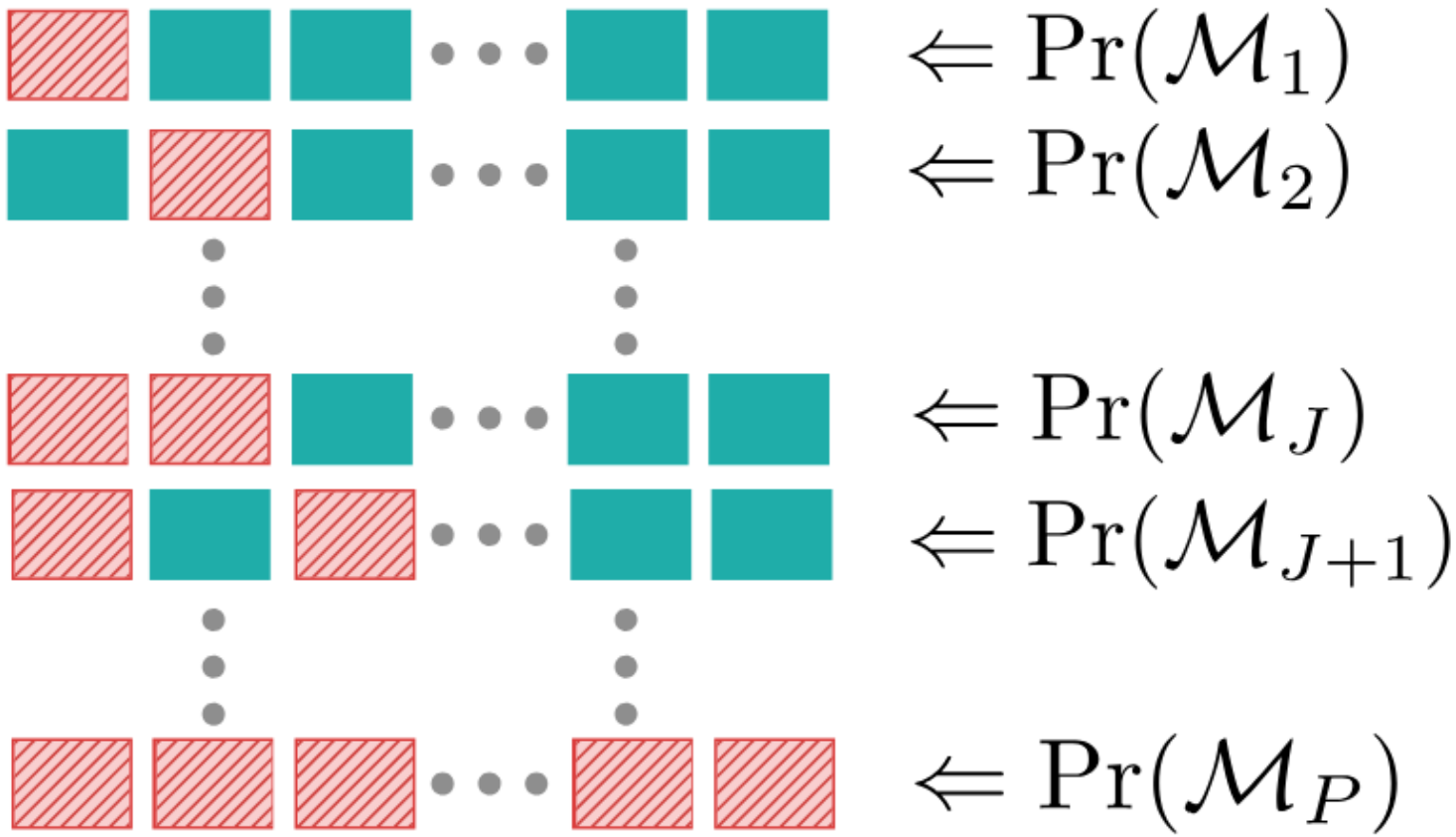
Assessing multi-effects configurations

$L = 2$



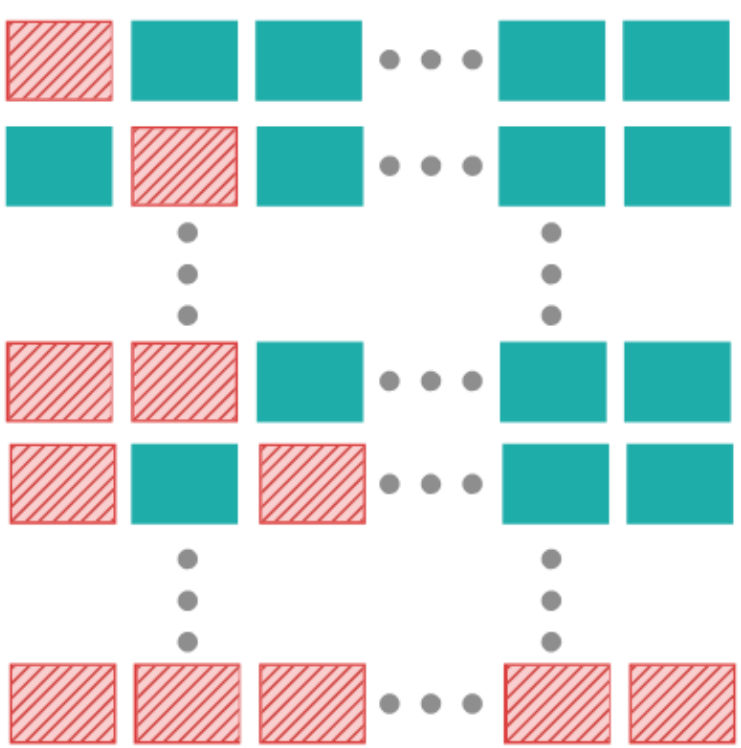
Assessing multi-effects configurations

$L = P$



Assessing multi-effects configurations

$L = P$

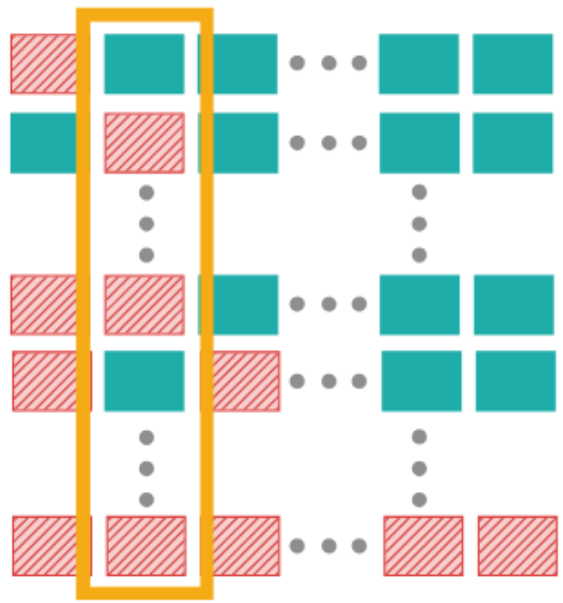


$$\Rightarrow \mathrm{PIP}(\mathcal{M}_1)$$
$$\Rightarrow \mathrm{PIP}(\mathcal{M}_2)$$
$$\Rightarrow \mathrm{PIP}(\mathcal{M}_J)$$
$$\Rightarrow \mathrm{PIP}(\mathcal{M}_{J+1})$$
$$\Rightarrow \mathrm{PIP}(\mathcal{M}_P)$$

PIP: Posterior Inclusion Probability, aka, Posterior Causal Probability.

Assessing multi-effects configurations

Marginal associations

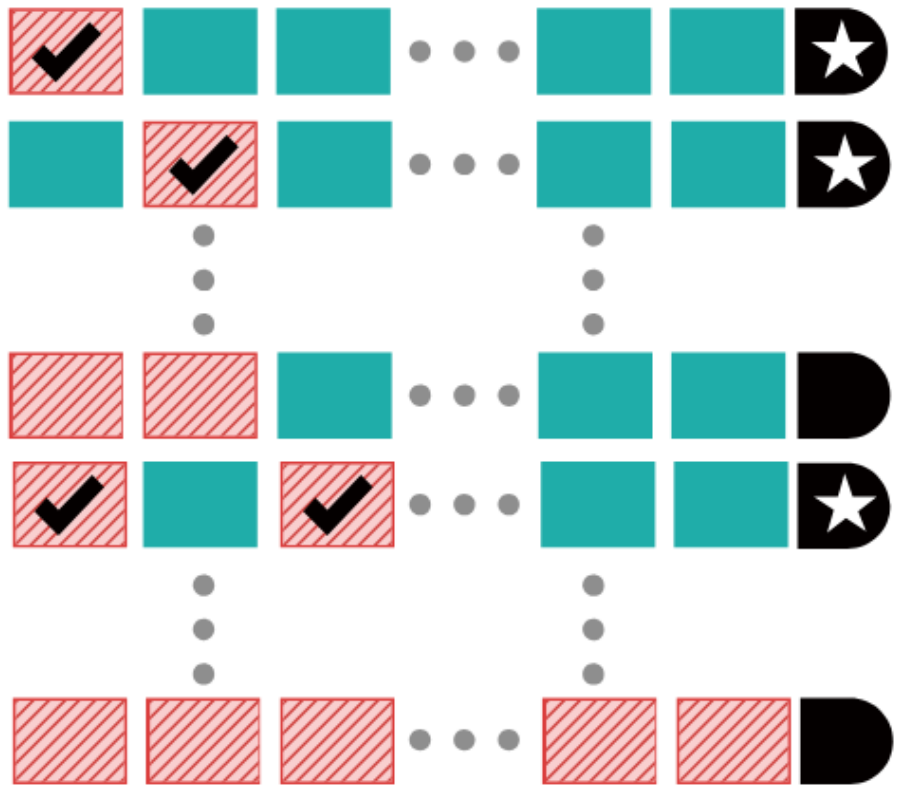


$\mathrm{PIP}(\mathcal{M}_1)$
$\mathrm{PIP}(\mathcal{M}_2)$
$\mathrm{PIP}(\mathcal{M}_J)$
$\mathrm{PIP}(\mathcal{M}_{J+1})$
$\mathrm{PIP}(\mathcal{M}_P)$

$$\mathrm{PIP}_2 = \mathrm{PIP}(\mathcal{M}_2) + \mathrm{PIP}(\mathcal{M}_J) + \mathrm{PIP}(\mathcal{M}_P)$$

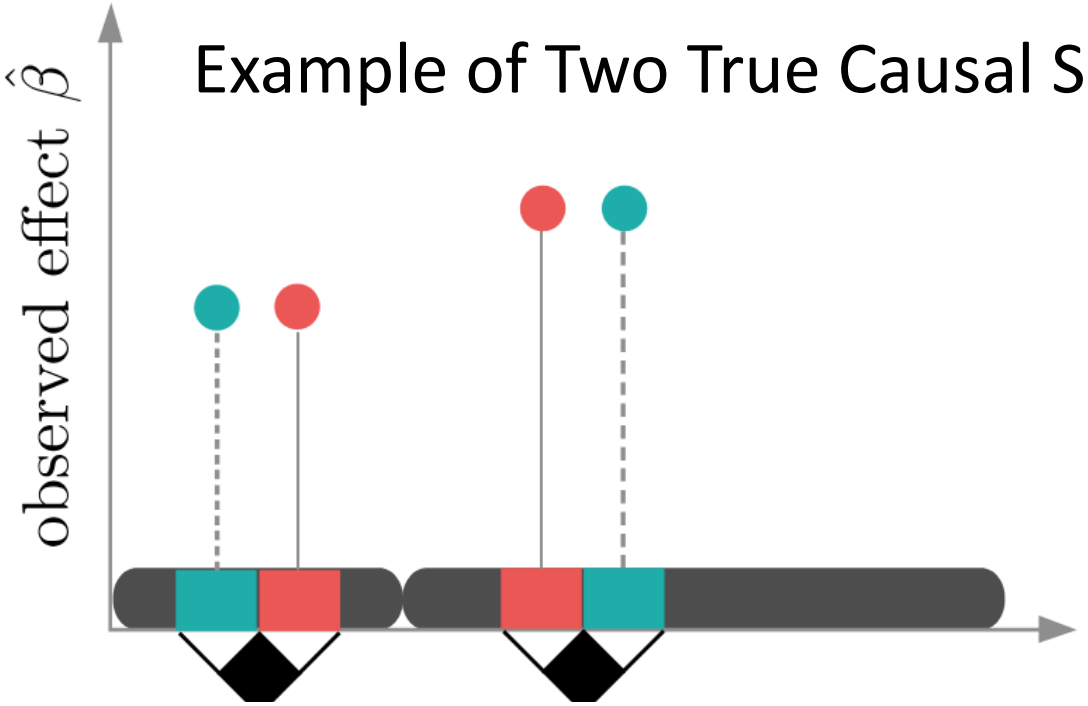
Assessing multi-effects configurations

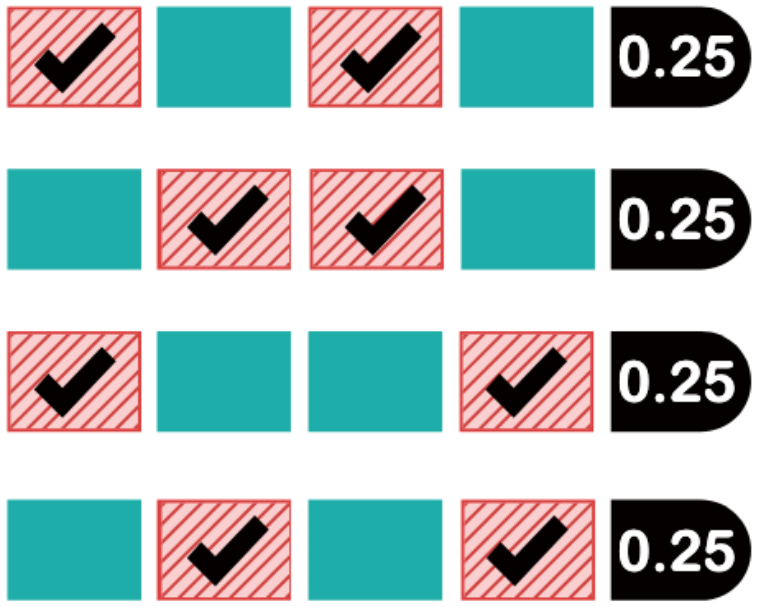
The 95% (smallest) Credible Set

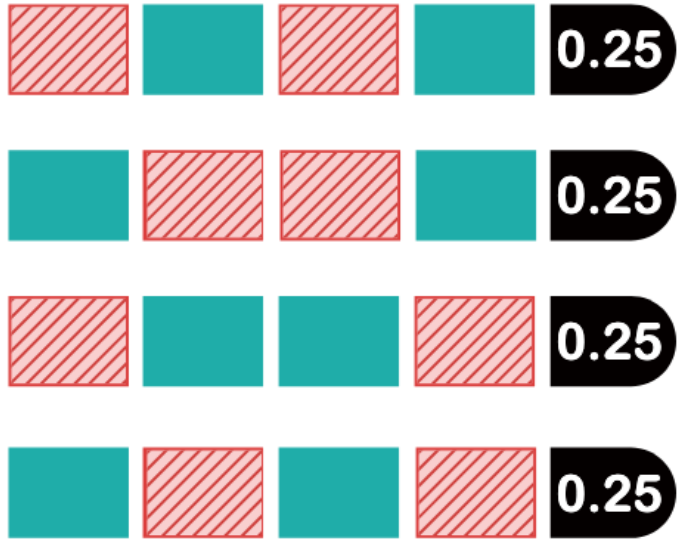


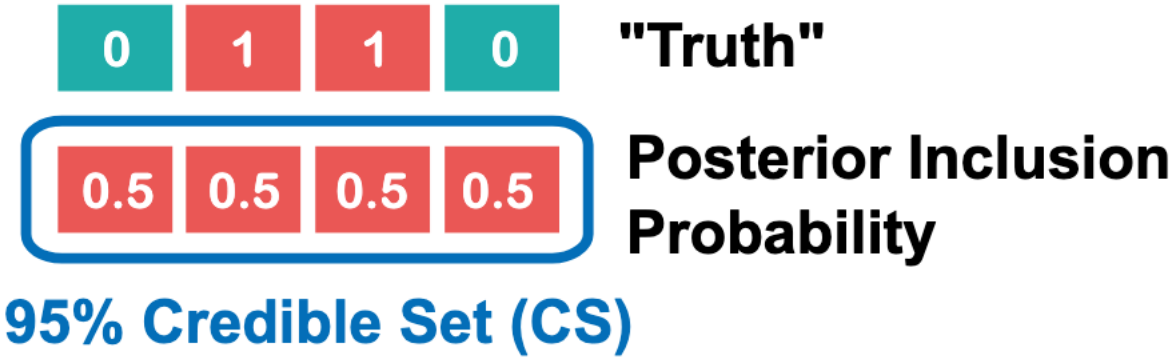
$$\alpha_1 = 0.70$$
$$\alpha_2 = 0.15$$
$$\alpha_3 = 0.02$$
$$\alpha_4 = 0.10$$
$$\alpha_5 = 0.00$$

Example of Two True Causal SNPs, each has one nearby SNP with LD=1









Bayesian Method for Fine-mapping

- Existing methods/tools assuming the same BVSR framework:
  - CAVIAR (Hormozdiari et al., 2014) : Enumeration
  - FINEMAP (Benner et al., 2016) : Stochastic search
  - DAP-G (Wen et al., 2016): Deterministic approximation

- Requires only GWAS Summary Statistics and Reference LD

- Provide probabilistic measures of causality for individual variants

Bayesian Fine-mapping using GWAS Summary Statistics

- Likelihood based on Multiple Linear Regression Model

$$y = X\beta + \epsilon, \qquad \epsilon \sim N(0, \sigma^2 I)$$

- MLE estimates of β depends on column-standardized (mean 0, standard deviation 1) X, y only through SNP correlation (LD) matrix R and single-SNP Z-score test statistic $\hat{z}$:

$$\hat{\beta} = (X^T X)^{-1} X^T y = n^{-\frac{1}{2}} \sigma R^{-1} \hat{z}$$

$$R = n^{-1}(X^T X), \qquad \hat{z} = \frac{X^T y}{\sqrt{n}\sigma}$$

$$Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1} = n^{-1} \sigma R^{-1}$$

$$E[\hat{\beta}] = \beta$$

Bayesian Fine-mapping using GWAS Summary Statistics

- Single-SNP Z-score test statistic $\hat{z}$ can be obtained from GWAS summary statistics
- SNP correlation matrix R can be approximated from a reference panel with the same ethnicity
- The likelihood function for β can be approximated by

$$\hat{\beta} \sim MVN(\beta, Var(\hat{\beta}))$$

- Use a Bayesian approach with a prior distribution to account for sparsity among causal effects

Priors for β with a binary indicator vector γ

- Assume an indicator vector $\gamma : \gamma_l = 1$ if the lth variant has non-zero causal effect $\beta_l \neq 0$; $\gamma_l = 0$ if $\beta_l \neq 0$.
- For non-zero effect sizes, the likelihood is given by

$$\beta|\gamma \sim MVN(0, s_\beta^2 \sigma^2 \Delta_\gamma)$$

Δ_r: Diagonal matrix with γ on the diagonal

- σ^2 can be taken as 1 for quantitative traits or $1/(\varphi(1-\varphi))$ with φ denoting the proportion of cases among n individuals
  - Assuming standardized phenotype vector
  - Assuming no other confounding covariates
- Taking $s_\beta^2 = 0.05^2$ means with 95% probability a causal SNP explains less than 1% of the phenotype variation (FINEMAP)

Prior of binary indicator vector γ with respect to the number of assumed true causal SNPs

- $p_k = \Pr(\#\ of\ k\ causal\ SNPs)\ , k = 1, \dots, K;\ \ K \ll m$ total number SNPs
- $p_0 = 0$, assuming there is at least one causal SNP for the fine-mapped region
- Assume the same probability for each configuration with k causal SNPs (FINEMAP)

$$p(\gamma) = p_k / \binom{m}{k},\ \sum_{l=1}^{m} \gamma_l = k$$

Likelihood function of indicator vector γ by integrating out β

- Posterior distribution of the indicator vector γ infers the posterior causal probability per SNP : $\mathrm{P}(\gamma|y,X)$
- Likelihood function of indicator vector γ by integrating out β:

$$\mathrm{L}(\gamma) = \mathrm{P}(y|\gamma,X) = \int P(y\,|\beta,X)P(\beta|\gamma)d\beta$$
$$= N\left(\hat{\beta}\middle|0,\sigma^2(nR)^{-1} + s_\beta^2\sigma^2\Delta_\gamma\right)$$
$$= N\left(\hat{z}\middle|0,R + R\Sigma_\gamma R\right),\Sigma_\gamma = ns_\beta^2\Delta_\gamma$$

Δ_r: Diagonal matrix with γ on the diagonal

- <u>The likelihood function $L(\gamma)$ need to be evaluated per γ</u>
- Computational efficiency is needed because of all $\sum_{k-1}^{K}\binom{m}{k}$ causal configurations

<u>$\hat{z}$: single-SNP Z-score test statistic; R: SNP correlation (LD) matrix</u>

Evaluate likelihood function $L(\gamma)$ by FINEMAP

- Partition $\hat{z}$ into components for the Causal SNPs $\hat{z}_C$ and Non-causal SNPs $\hat{z}_N$
- Partition R, Σ_γ, and $R\Sigma_\gamma R$

$$R = \left[\begin{array}{c|c} R_{CC} & R_{CN} \\ \hline R_{NC} & R_{NN} \end{array}\right] \qquad \Sigma_\gamma = \begin{bmatrix} \Sigma_{CC} & 0 \\ 0 & 0 \end{bmatrix}$$

$$R + R\Sigma_\gamma R = \left[\begin{array}{c|c} R_{CC} + R_{CC}\Sigma_{CC}R_{CC} & R_{CN} + R_{CC}\Sigma_{CC}R_{CN} \\ \hline R_{NC} + R_{NC}\Sigma_{CC}R_{CC} & R_{NN} + R_{NC}\Sigma_{CC}R_{CN} \end{array}\right]$$

- Use the properties of conditional multivariate normal distribution

$$\mathbb{E}[\hat{z}_N|\hat{z}_C] = R_{NC}R_{CC}^{-1}\hat{z}_C$$

$$\mathbb{V}[\hat{z}_N|\hat{z}_C] = R_{NN} - R_{NC}R_{CC}^{-1}R_{CN}$$

Evaluate likelihood function $L(\gamma)$ by FINEMAP

- Rewrite the marginal likelihood function $L(\gamma)$:

$$L(\gamma) = P(\hat{z}|\gamma, R, \Sigma_\gamma) = N(\hat{z}|0, R + R\Sigma_\gamma R) = P(\hat{z}_N|\gamma, \hat{z}_C, R, \Sigma_\gamma)P(\hat{z}_C|\gamma, R_{CC}, \Sigma_{CC})$$
$$= N(\hat{z}_C|0, R_{CC} + R_{CC}\Sigma_{CC}R_{CC})N(\hat{z}_N|E[\hat{z}_N|\hat{z}_C], Var(\hat{z}_N|\hat{z}_C))$$

NULL: $L(\gamma = 0) = P(\hat{z}|\gamma = 0, R) = N(\hat{z}_C|0, R_{CC})N(\hat{z}_N|E[\hat{z}_N|\hat{z}_C], Var(\hat{z}_N|\hat{z}_C))$

$$\mathcal{N}(\hat{\boldsymbol{z}}|\boldsymbol{0}, \boldsymbol{R} + \boldsymbol{R}\boldsymbol{\Sigma}_\gamma\boldsymbol{R}) = \mathcal{N}(\hat{\boldsymbol{z}}_C|\boldsymbol{0}, \boldsymbol{R}_{CC} + \boldsymbol{R}_{CC}\boldsymbol{\Sigma}_{CC}\boldsymbol{R}_{CC}) \times$$
$$\mathcal{N}(\hat{\boldsymbol{z}}_N|\mathbb{E}[\hat{\boldsymbol{z}}_N|\hat{\boldsymbol{z}}_C], \mathbb{V}[\hat{\boldsymbol{z}}_N|\hat{\boldsymbol{z}}_C])$$
$$= \mathcal{N}(\hat{\boldsymbol{z}}_C|\boldsymbol{0}, \boldsymbol{R}_{CC} + \boldsymbol{R}_{CC}\boldsymbol{\Sigma}_{CC}\boldsymbol{R}_{CC}) \times \frac{\mathcal{N}(\hat{\boldsymbol{z}}|\boldsymbol{0}, \boldsymbol{R})}{\mathcal{N}(\hat{\boldsymbol{z}}_C|\boldsymbol{0}, \boldsymbol{R}_{CC})}$$

- Bayes factor for assessing the evidence with a given γ against the null model using only causal SNPs (<u>calculation only involves causal SNPs</u>).

Bayes factor (BF) is a likelihood ratio of the marginal likelihood of two competing hypotheses

$$\mathrm{BF}(\boldsymbol{\gamma}:\mathrm{NULL}) = \frac{\mathcal{N}(\hat{\boldsymbol{z}}|\boldsymbol{0}, \boldsymbol{R} + \boldsymbol{R}\boldsymbol{\Sigma}_\gamma\boldsymbol{R})}{\mathcal{N}(\hat{\boldsymbol{z}}|\boldsymbol{0}, \boldsymbol{R})}$$
$$= \frac{\mathcal{N}(\hat{\boldsymbol{z}}_C|\boldsymbol{0}, \boldsymbol{R}_{CC} + \boldsymbol{R}_{CC}\boldsymbol{\Sigma}_{CC}\boldsymbol{R}_{CC})}{\mathcal{N}(\hat{\boldsymbol{z}}_C|\boldsymbol{0}, \boldsymbol{R}_{CC})}$$

Posterior for γ

- Unnormalized posterior probability

$$\mathrm{P}(\gamma|y,X) = BF(\gamma: NULL) * \left(p_k / \binom{m}{k}\right)$$

- Can be normalized over all $\sum_{k-1}^{K} \binom{m}{k}$ causal configurations
- A Shotgun Stochastic Search (SSS) algorithm (Hans et al. 2007) was used by FINAMAP to rapidly evaluate many configurations and is designed to discover especially those with highest posterior probability

Shotgun Stochastic Search (SSS) algorithm

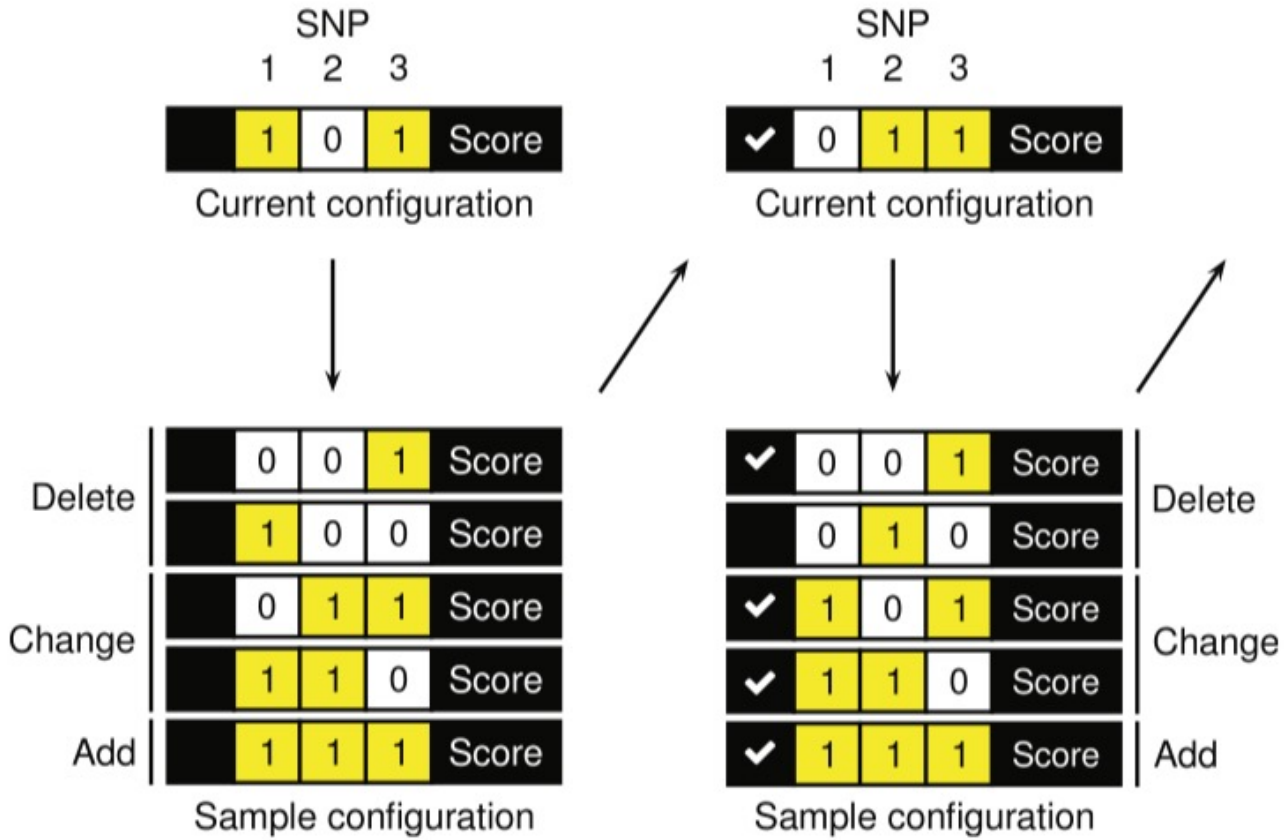


Fig. 2. Shotgun stochastic search rapidly identifies configurations of causal SNPs with high posterior probability. In each iteration, the neighborhood of the current causal configuration is defined by configurations that result from deleting, changing or adding a causal SNP (1) from the current configuration. The next iteration starts by sampling a new causal configuration from the neighborhood based on the scores normalized within the neighborhood. The unnormalized posterior probabilities remain fixed throughout the algorithm and can thus be memorized (✔) to avoid recomputation when already-evaluated configurations appear in another neighborhood

Single-SNP Bayes factor

- Marginal posterior probability that the lth SNP is causal, i.e., single-SNP inclusion probability:

$$p(\gamma_\ell = 1|\boldsymbol{y}, \boldsymbol{X}) = \sum_{\boldsymbol{\gamma}\in\Gamma^*} 1(\gamma_\ell = 1)p(\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{X}).$$

- Single-SNP Bayes factor

$$\mathrm{BF}(\gamma_\ell = 1 : \gamma_\ell = 0) = \frac{p(\gamma_\ell = 1|\boldsymbol{y}, \boldsymbol{X})}{p(\gamma_\ell = 0|\boldsymbol{y}, \boldsymbol{X})} \Big/ \frac{p(\gamma_\ell = 1)}{p(\gamma_\ell = 0)},$$

where the prior probability of the ℓth SNP being causal is

$$p(\gamma_\ell = 1) = \sum_{k=1}^{K} \left(\frac{k}{m}\right) p_k.$$

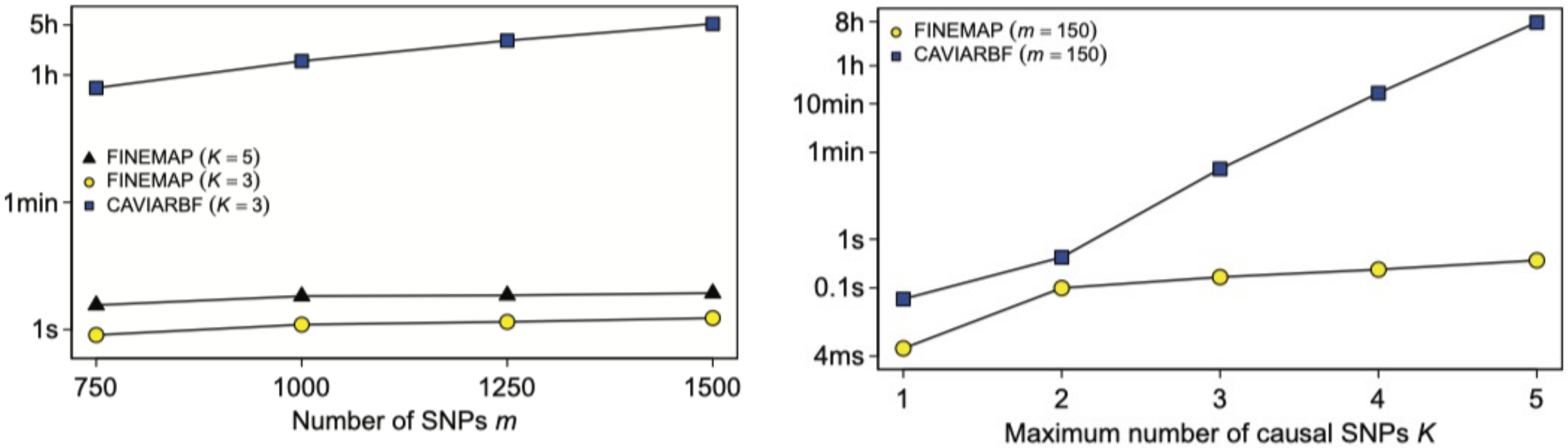


Fig. 3. Processing time of one locus with FINEMAP and CAVIARBF on $\log_{10}$ scale. Top panel: Scenario A with increasing number of SNPs allowing $K = 3$ or $K = 5$ causal SNPs. Bottom panel: Scenario B with 150 SNPs considering causal configurations with different maximum numbers of SNPs. All processing times are averaged over 500 datasets using one core of a Intel Haswell E5-2690v3 processor running at 2.6 GHz

Fine-mapping accuracy

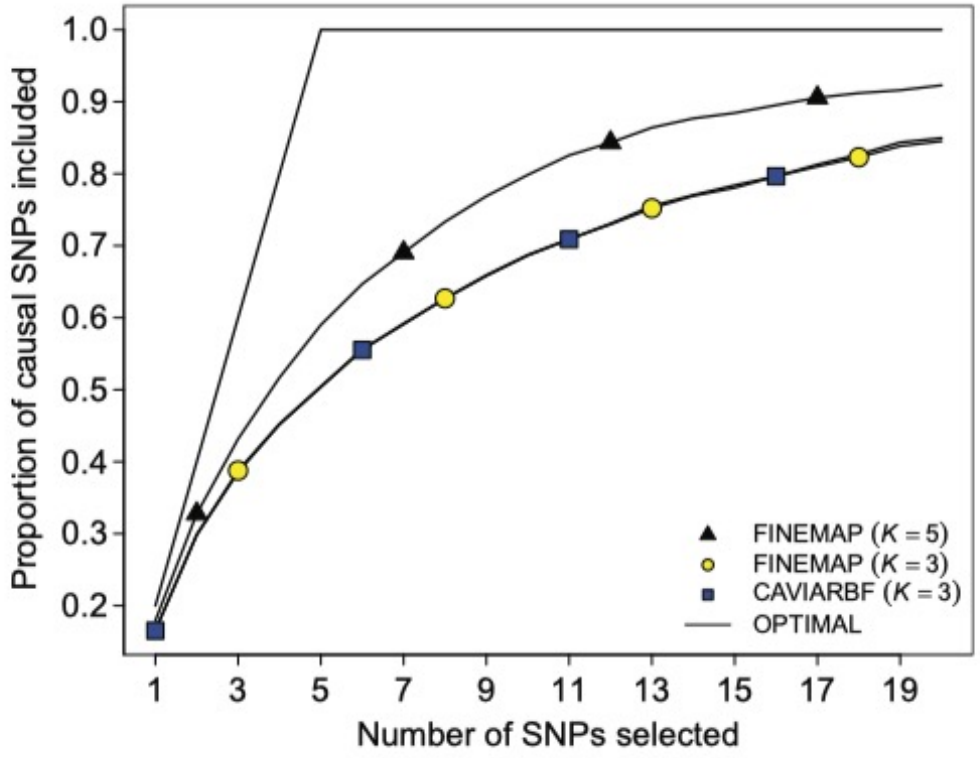


Fig. 5. Fine-mapping accuracy of FINEMAP and CAVIARBF on data with five causal SNPs, allowing either $K = 3$ or $K = 5$ causal SNPs. The proportion of causal SNPs included is plotted against the number of top SNPs selected on the basis of ranked single-SNP inclusion probabilities. Proportions are averaged over 500 datasets with 1500 SNPs. Case $K = 5$ is computationally intractable for CAVIARBF

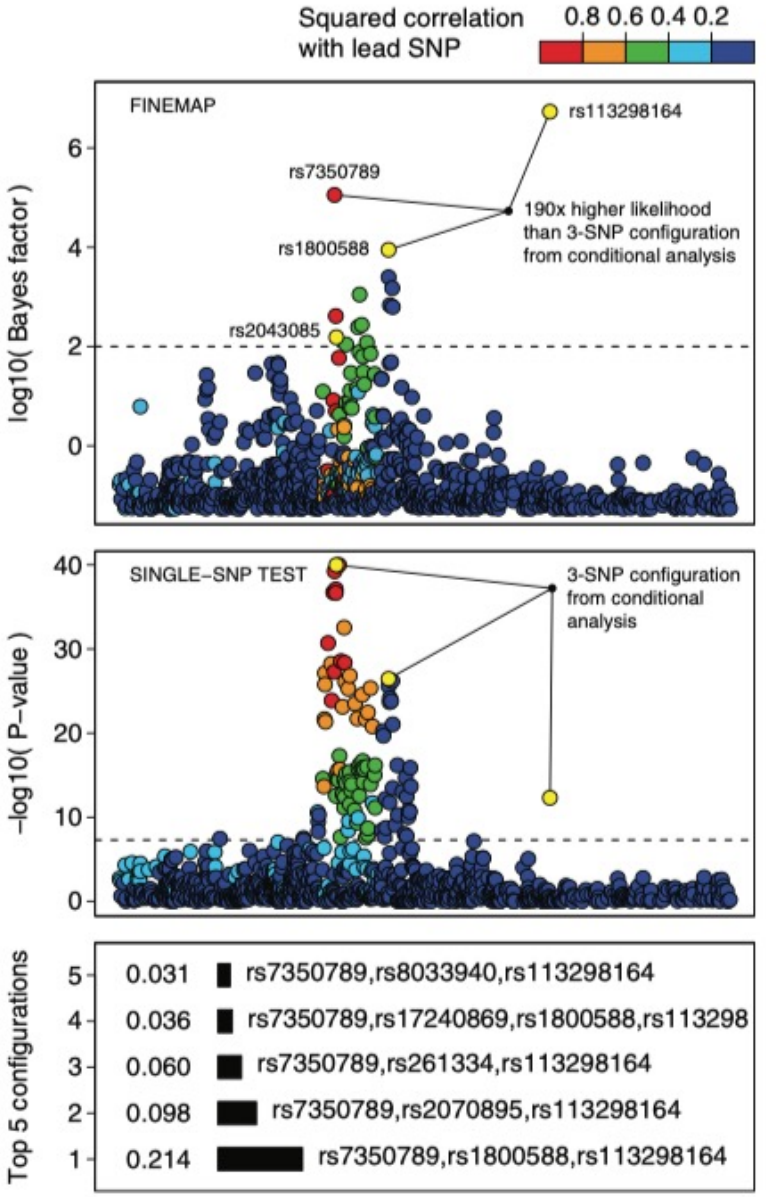


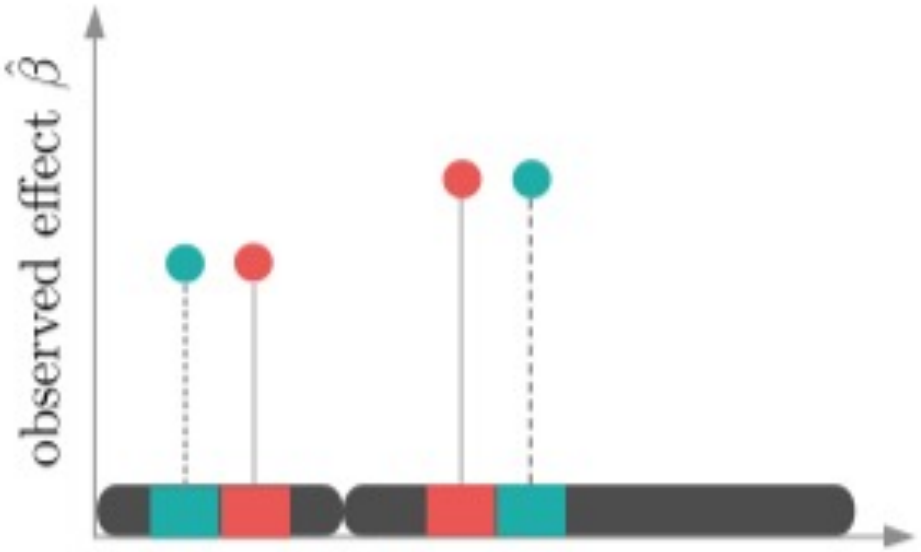
Fig. 7. Fine-mapping of 15q21/*LIPC* region associated with high-density lipoprotein cholesterol. Independent association signals in conditional analysis are highlighted by ○. Dashed lines correspond respectively to a single-SNP Bayes factor of 100 and *P*-value of 5×10^{-8}. Squared correlations are shown with respect to rs2043085

Limitation of conventional BVSR inference

0.5	0.5	0.5	0.5

- There are 2 signals expected $(0.5 + 0.5 + 0.5 + 0.5)$
- But **which two?** Any two?
- 95% certainty that **all** effect variables are captured?
- We need to quantify this better!

Example



Quantifying uncertainty in variable selection

Consider a sparse regression example

$$\boldsymbol{y} = \sum_{j=1}^{p} \boldsymbol{x}_j b_j + \boldsymbol{e} \quad \boldsymbol{e} \sim N(0, \sigma^2 I_n), \tag{1}$$

where $\boldsymbol{x}_1 = \boldsymbol{x}_2, \boldsymbol{x}_3 = \boldsymbol{x}_4, b_1 \neq 0, b_4 \neq 0, b_{j \notin \{1,4\}} = 0.$

We are interested in making the following statement,

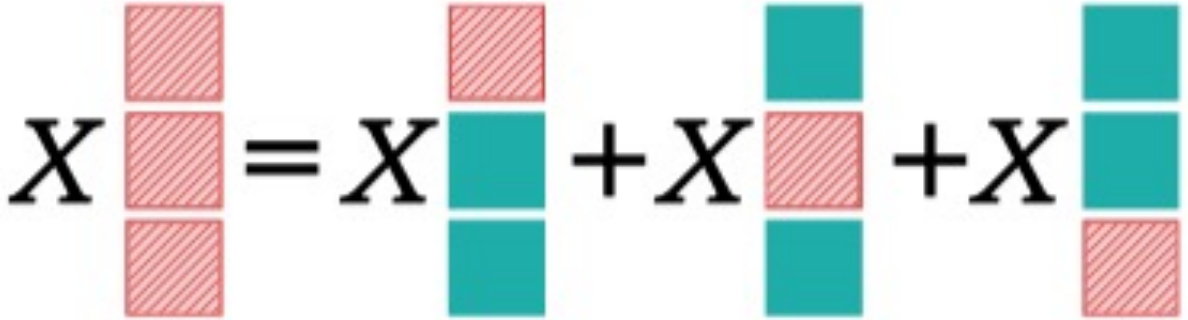
$$(b_1 \neq 0 \text{ or } b_2 \neq 0) \text{ AND } (b_3 \neq 0 \text{ or } b_4 \neq 0).$$

1. There are two independent variables with non-zero effect
2. $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ (and $\boldsymbol{x}_3$ and $\boldsymbol{x}_4$) are too similar to distinguish
3. yet they can be prioritized relative to each other

The Sum of Single Effects model (*SuSiE*)

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$$

$$\boldsymbol{b} = \sum_{l=1}^{L} \boldsymbol{b}_l$$



A variational approximation to posterior under *SuSiE*

$$q(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_L) = \prod_l q_l(\boldsymbol{b}_l)$$

- $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_L$ are treated as **independent** *a posteriori*.
- **Do not** assume q_l factorizes over the elements of $\boldsymbol{b}_l$.

G. Wang et. al. JRSS B. 2020.

Single-Effect Regression (SER) Model

Specifically, we consider the following SER model, with hyperparameters for the residual variance, σ^2, 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:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}, \tag{2.4}$$

$$\mathbf{e} \sim N_n(0, \sigma^2 I_n), \tag{2.5}$$

$$\mathbf{b} = b\boldsymbol{\gamma}, \tag{2.6}$$

$$\boldsymbol{\gamma} \sim \text{Mult}(1, \boldsymbol{\pi}), \tag{2.7}$$

$$b \sim N_1(0, \sigma_0^2). \tag{2.8}$$

Here, $\mathbf{y}$ is the n-vector of response data, $\mathbf{X} = (x_1, \ldots, x_p)$ is an $n \times p$ matrix containing n observations of p explanatory variables, b is a scalar representing the 'single effect', $\boldsymbol{\gamma} \in \{0, 1\}^p$ is a p-vector of indicator variables, $\mathbf{b}$ is the p-vector of regression coefficients, $\mathbf{e}$ is an n-vector of independent error terms and $\text{Mult}(m, \boldsymbol{\pi})$ denotes the multinomial distribution on class counts that is obtained when m samples are drawn with class probabilities given by $\boldsymbol{\pi}$. We assume that $\mathbf{y}$ and the columns of $\mathbf{X}$ have been centred to have mean 0, which avoids the need for an intercept term (Chipman *et al.*, 2001).

G. Wang et. al. JRSS B. 2020.

Posterior under SER Model

Given σ^2 and σ_0^2, the posterior distribution on $\mathbf{b} = \boldsymbol{\gamma} b$ is easily computed:

$$\boldsymbol{\gamma} | \mathbf{X}, \mathbf{y}, \sigma^2, \sigma_0^2 \sim \text{Mult}(1, \boldsymbol{\alpha}), \tag{2.9}$$

$$b | \mathbf{X}, \mathbf{y}, \sigma^2, \sigma_0^2, \gamma_j = 1 \sim N(\mu_{1j}, \sigma_{1j}^2), \tag{2.10}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$ is the vector of PIPs, with $\alpha_j := \Pr(\gamma_j = 1 | \mathbf{X}, \mathbf{y}, \sigma^2, \sigma_0^2)$, and μ_{1j} and σ_{1j}^2 are the posterior mean and variance of b given $\gamma_j = 1$. Calculating these quantities simply involves performing the p univariate regressions of $\mathbf{y}$ on columns x_j of $\mathbf{X}$, for $j = 1, \ldots, p$, as shown in the on-line appendix A. From $\boldsymbol{\alpha}$, it is also straightforward to compute a level ρ credible set (definition 1), $\text{CS}(\boldsymbol{\alpha}; \rho)$, as described in Maller *et al.* (2012), and detailed in appendix A. In brief, this involves sorting variables by decreasing α_j and then including variables in the credible set until their cumulative probability exceeds ρ.

For later convenience, we introduce a function, SER, that returns the posterior distribution for $\mathbf{b}$ under the SER model. Since this posterior distribution is uniquely determined by the values of $\boldsymbol{\alpha}$, $\boldsymbol{\mu}_1 := (\mu_{11}, \ldots, \mu_{1p})$ and $\boldsymbol{\sigma}_1^2 := (\sigma_{11}^2, \ldots, \sigma_{1p}^2)$ in distributions (2.9)–(2.10), we can write

$$\text{SER}(\mathbf{X}, \mathbf{y}; \sigma^2, \sigma_0^2) := (\boldsymbol{\alpha}, \boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2). \tag{2.11}$$

G. Wang et. al. JRSS B. 2020.

Sum of Single Effect regression, *SuSiE*

"single effect": $\boldsymbol{b}_l$'s

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$$
$$\boldsymbol{e} \sim N(0, \sigma^2 I_n)$$
$$\boldsymbol{b} = \sum_{l=1}^{L} \boldsymbol{b}_l$$
$$\boldsymbol{b}_l = \gamma_l \beta_l$$
$$\gamma_l \sim \text{Mult}(1, \boldsymbol{\pi})$$
$$\beta_l \sim N(0, \sigma_{0_l}^2)$$
$$\sigma_{0_l}^2 \geq 0$$

A mean-field approximation

$$q(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_L) = \prod_l q_l(\boldsymbol{b}_l)$$

- $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_L$ are treated as **independent** *a posteriori*.
- **Do not** assume q_l factorizes over the elements of $\boldsymbol{b}_l$.

Iterative Bayesian forward selection (IBSS) for SuSiE inference

Table 1. Algorithm 1: IBSS

Require data $\mathbf{X}$, $\mathbf{y}$
Require number of effects, L, and hyperparameters σ^2, $\boldsymbol{\sigma}_0^2$
Require a function $\text{SER}(\mathbf{X}, \mathbf{y}; \sigma^2, \sigma_0^2) \to (\boldsymbol{\alpha}, \boldsymbol{\mu}_1, \boldsymbol{\sigma}_1)$ that computes the posterior distribution for $\mathbf{b}_l$ under the SER model; see expression (2.11)

1, initialize posterior means ▷ other initializations are possible (see algorithm 3 in the on-line appendix B)
$\bar{\mathbf{b}}_l = 0$, for $l = 1, \ldots, L$
2, *repeat*
3, *for* l in $1, \ldots, L$ *do*
4, $\bar{\mathbf{r}}_l \leftarrow \mathbf{y} - \mathbf{X}\Sigma_{l' \neq l}\bar{\mathbf{b}}_{l'}$. ▷ expected residuals without lth single effect
5, $(\boldsymbol{\alpha}_l, \boldsymbol{\mu}_{1l}, \boldsymbol{\sigma}_{1l}) \leftarrow \text{SER}(\mathbf{X}, \bar{\mathbf{r}}_l; \sigma^2, \sigma_{0l}^2)$ ▷ fit SER to residuals
6, $\bar{\mathbf{b}}_l \leftarrow \boldsymbol{\alpha}_l \circ \boldsymbol{\mu}_{1l}$ ▷ '$\circ$' denotes elementwise multiplication
7, *until* convergence criterion satisfied
return $\boldsymbol{\alpha}_1, \boldsymbol{\mu}_{11}, \boldsymbol{\sigma}_{11}, \ldots, \boldsymbol{\alpha}_L, \boldsymbol{\mu}_{1L}, \boldsymbol{\sigma}_{1L}$

SER: Single Effect Regression

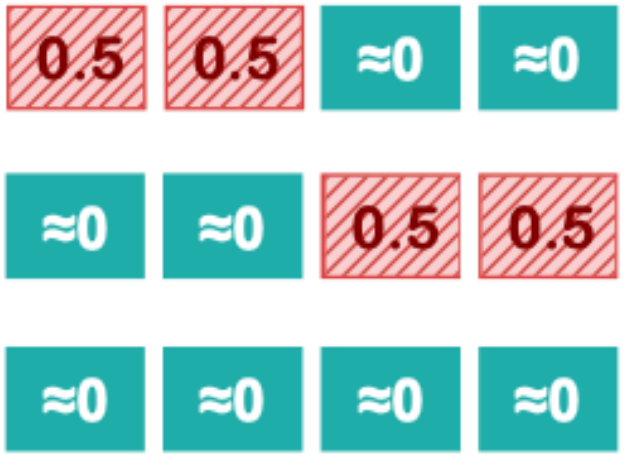
IBSS algorithm illustration

1. At random (zero) initialization, fit single effect model on $\boldsymbol{y}$

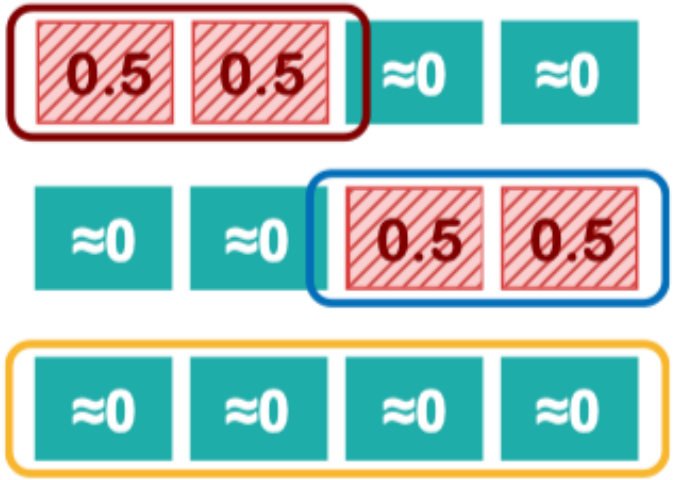
2. Compute residual $\boldsymbol{r}_2$ using fitted model, and do it again

IBSS algorithm illustration

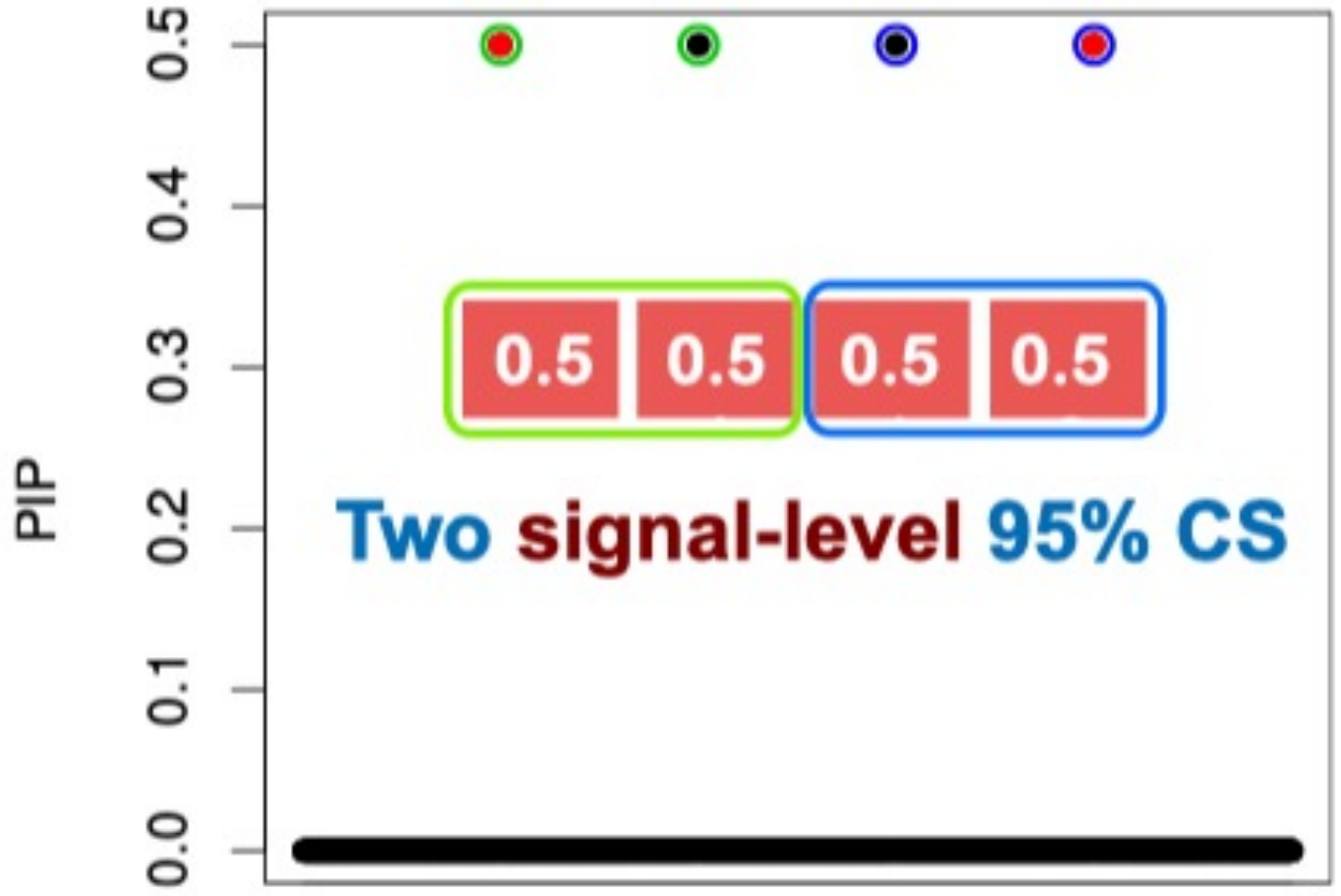
3. Compute residual r_3 using fitted model, and do it again



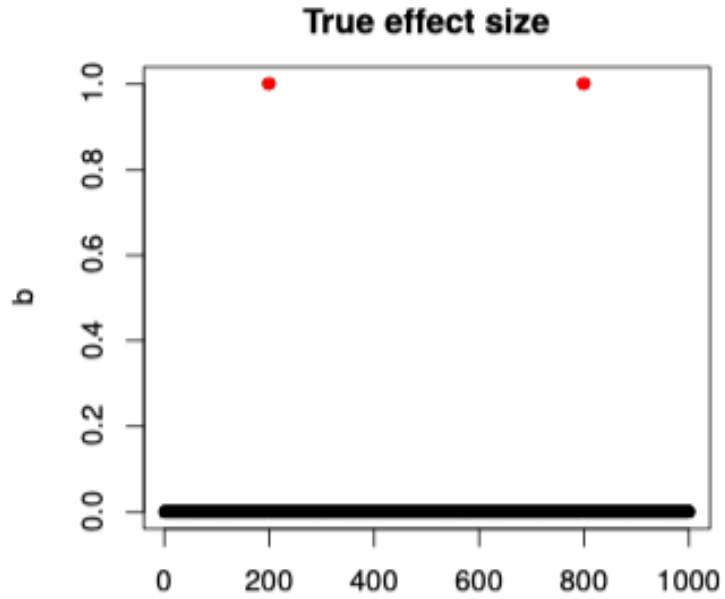
4. Iterate until converge; compute single-effect credible sets

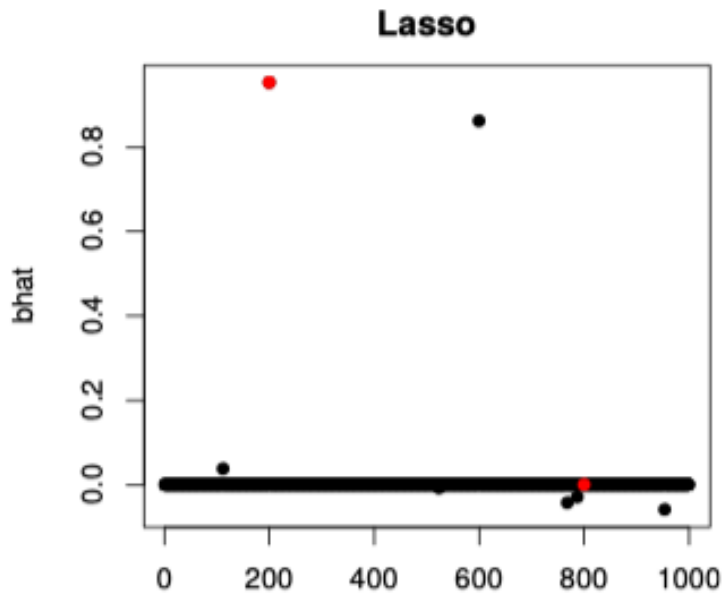


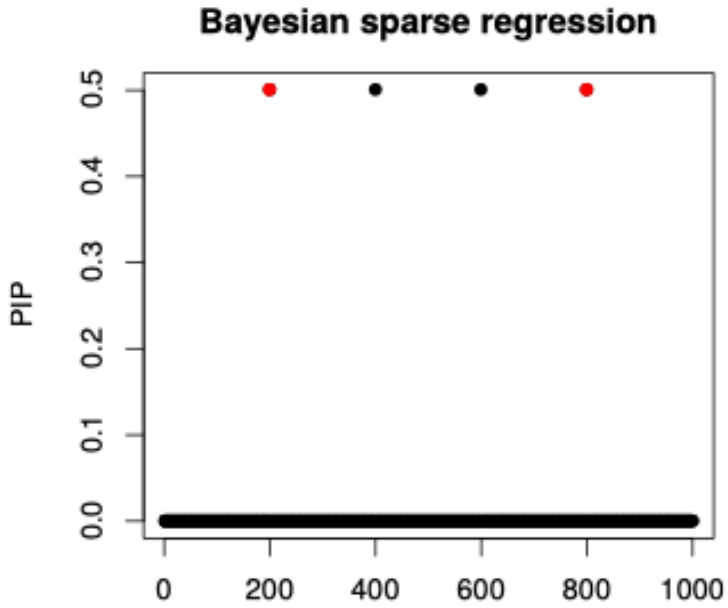
IBSS algorithm illustration

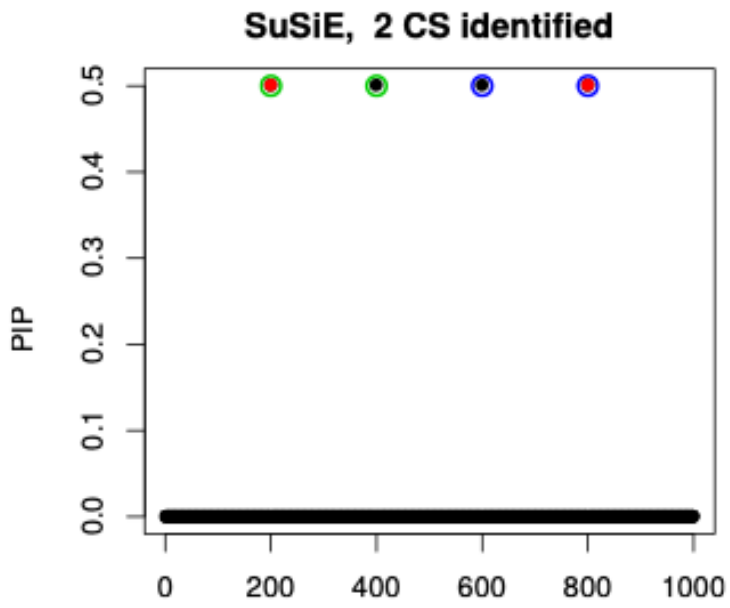


Example Results

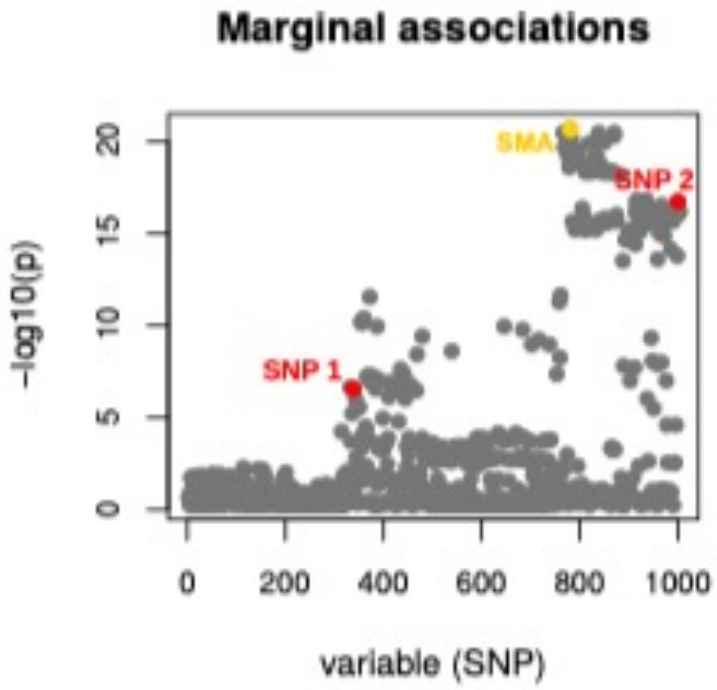


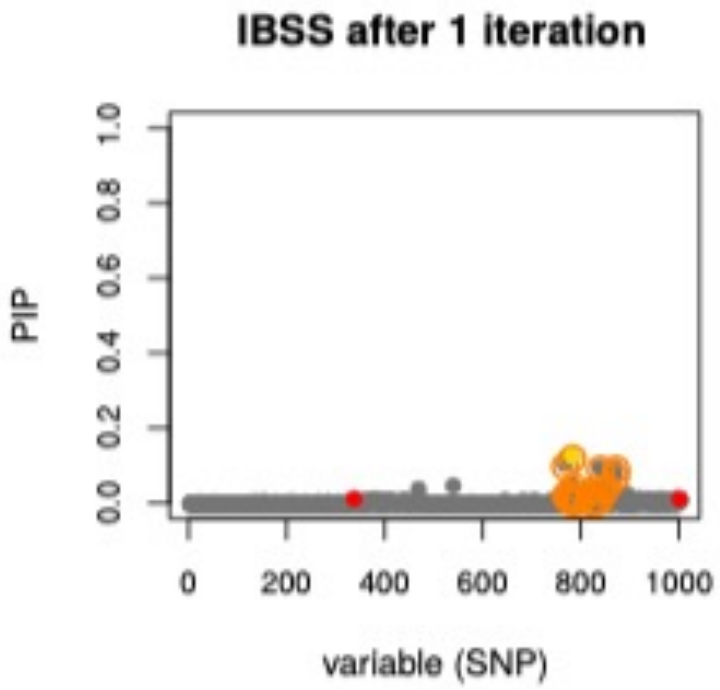


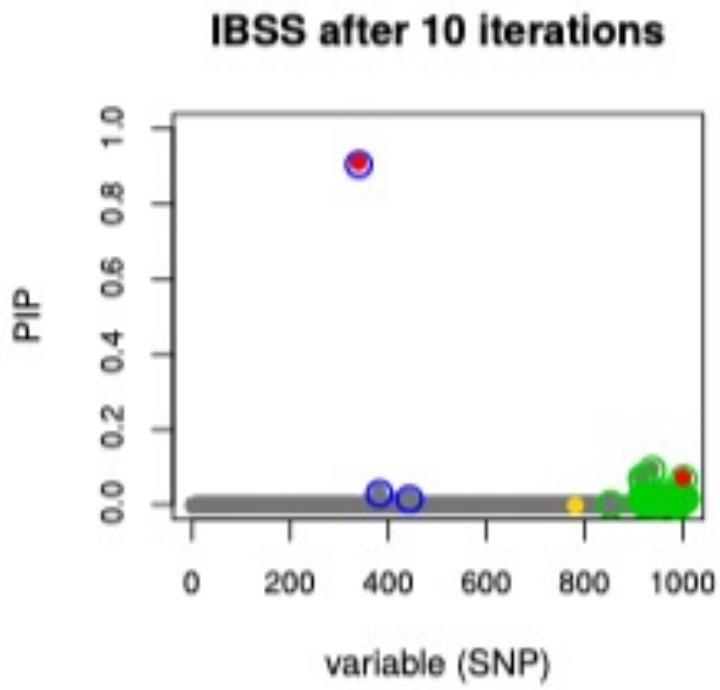




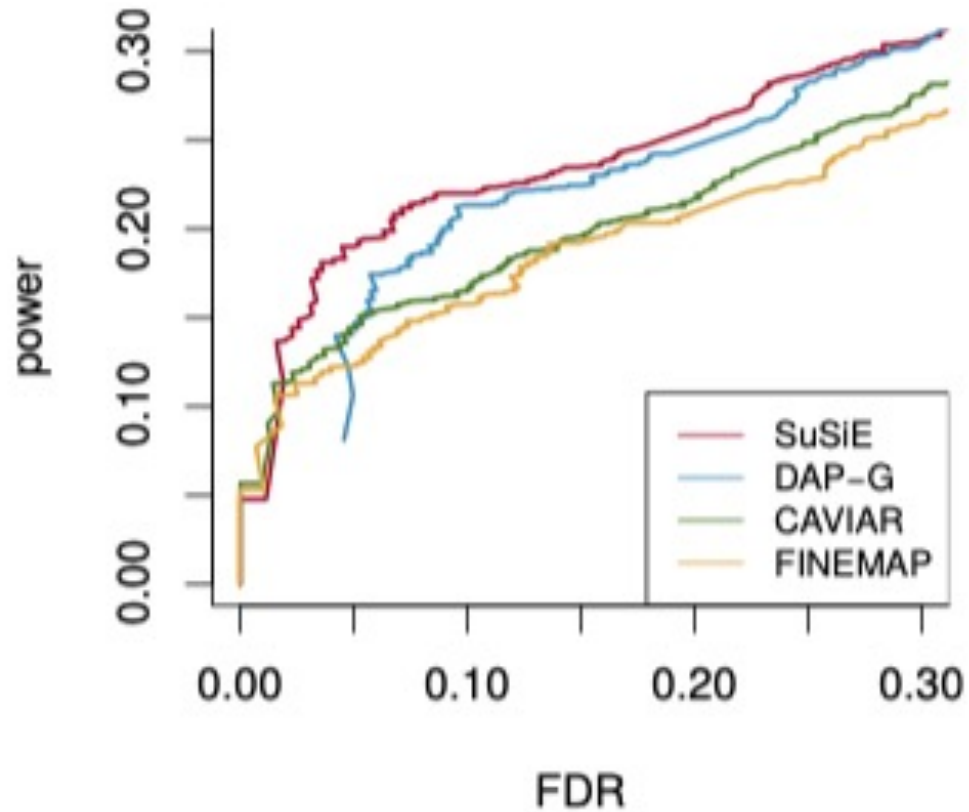
Real-world example illustrated







SuSiE is powerful



† *SuSiE* priors not required as they are **learned from local tests**.

SuSiE is fast

Speed comparison (3 causal variables; unit: sec.), June 2018

Method	Avg.	Min.	Max.
SuSiE †	**0.64**	0.34	2.28
DAP-G	**2.87**	2.23	8.87
FINEMAP	**23.01**	10.99	48.16
CAVIAR	**2907.51**	2637.34	3018.52

† An R implementation of *SuSiE*. Others are implemented in C++.

Fine-mapping using GWAS Summary Statistics

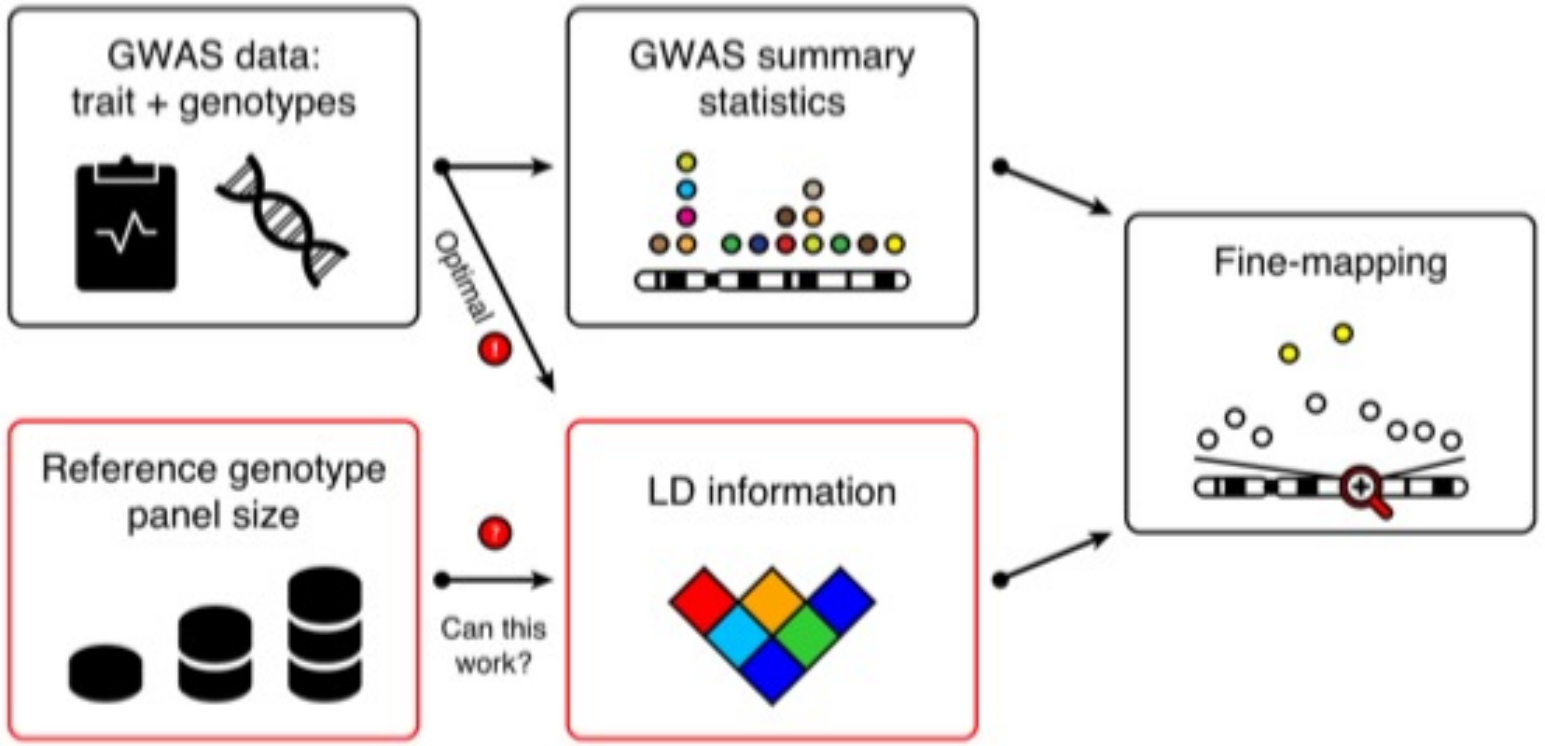


Figure: Benner *et al*. (2017) Am. J. Hum. Genet.

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

$$f(\beta|\text{Summary data}) \propto f(\text{Summary data}|\beta) \times f(\beta)$$

Single Causal variant model fine-mapping using summary statistics

- m SNPs in the region to fine-map
- Prior = each SNP has the same probability to be causal
- Posterior:

$$P(C_j|Z_1, \ldots, Z_m) = \frac{\exp(\frac{Z_j^2}{2})}{\sum_{k=1}^{m} \exp(\frac{Z_k^2}{2})}$$

SuSiE Regression with Summary Statistics (RSS)

"Single effects": $\boldsymbol{z}_l$'s

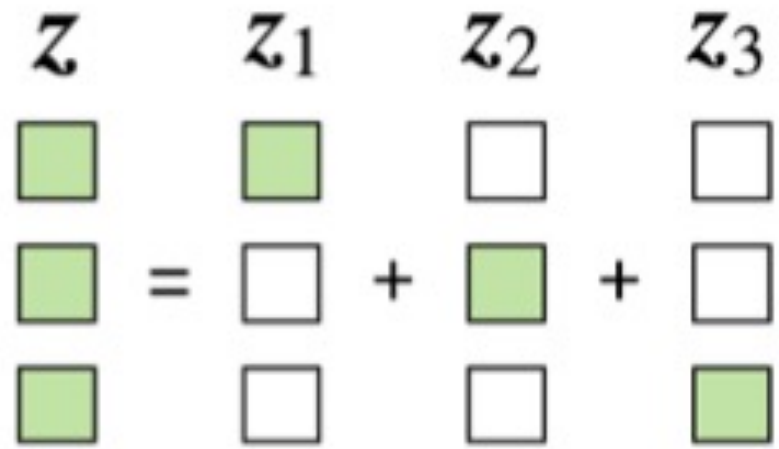
$$\hat{\boldsymbol{z}} \sim N(\hat{\boldsymbol{R}}\boldsymbol{z}, \hat{\boldsymbol{R}})$$

$$\boldsymbol{z} = \sum_{l=1}^{L} \boldsymbol{z}_l$$

$$\boldsymbol{z}_l = \gamma_l z_l$$

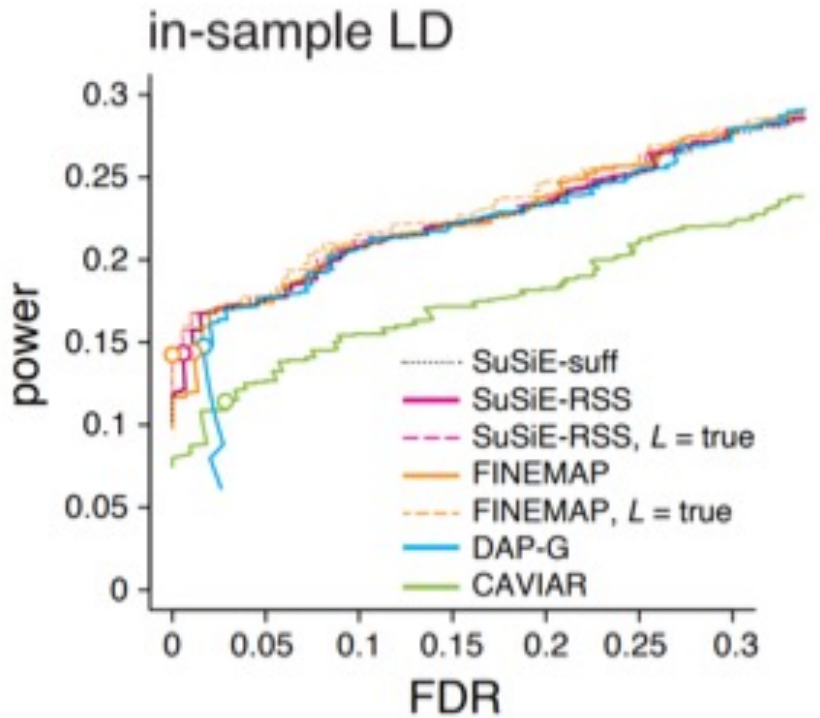
$$z_l \sim N(0, \omega_l^2)$$

$$\gamma_l \sim \text{Mult}(1, \boldsymbol{\pi})$$



Suggested reading: Zou et al (2022) PLoS Genet.

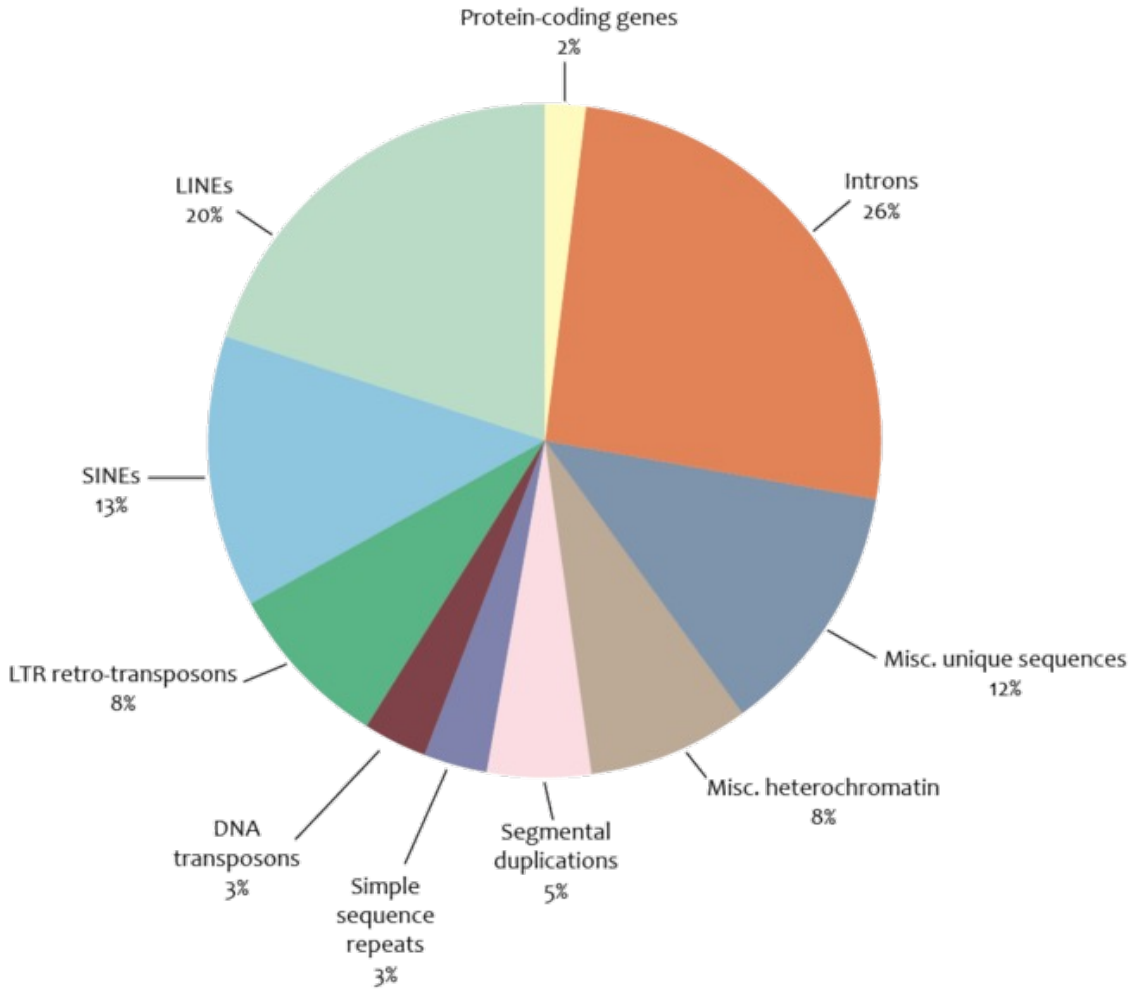
Summary statistics fine-mapping methods comparison



In practice people often use SuSiE RSS, Zou *et al*. (2022) PLoS Genet + FINEMAP, Benner *et al*. (2016) Bioinformatics

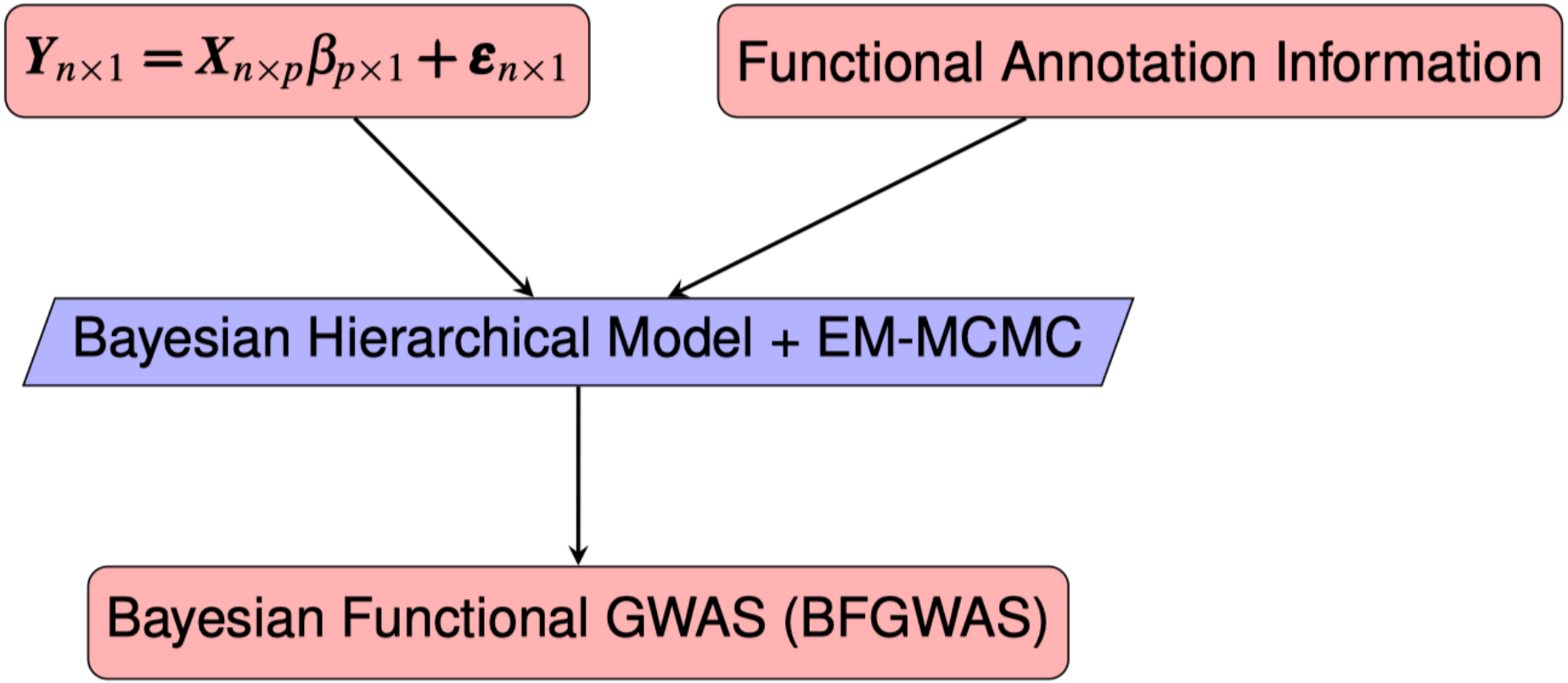
Bayesian Functional GWAS

Account for Functional Annotation



- Prioritize functional SNPs
- Quantify enrichment of each type of functional SNPs with respect to GWAS associations

Bayesian Functional GWAS



Bayesian Hierarchical Model

Joint linear regression model

$$\boldsymbol{Y}_{n\times 1} = \boldsymbol{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}, \quad \varepsilon \sim MN(0, \tau^{-1}I). \tag{1}$$

Prior:

- $\beta_{i_q} \sim \pi_q N(0, \tau^{-1}\sigma_q^2) + (1-\pi_q)\delta_0$, for variants of annotation q
- Introduce a latent indicator vector $\boldsymbol{\gamma}_{p\times 1}$, equivalently

$$\gamma_{i_q} \sim Bernoulli(\pi_q),\ \boldsymbol{\beta}_{-\boldsymbol{\gamma}} \sim \delta_0(\cdot),\ \boldsymbol{\beta}_{\boldsymbol{\gamma}} \sim MVN_{|\boldsymbol{\gamma}|}(0, \tau^{-1}\boldsymbol{V}_{\boldsymbol{\gamma}})$$

J. Yang et. al. AJHG. 2017

Parameters of Interest

- Category-specific (Enrichment parameters):
  - $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_Q)$: Causal probability per annotation
  - $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_Q^2)$: Effect-size variance for associated variants per annotation
- SNP-specific (Association evidence):
  - β_i: Genetic effect-size
  - $E[\gamma_i]$: Bayesian posterior inclusion probability (Bayesian PP), i.e., probability of being an associated SNP
- Region-level (Association evidence):
  - Regional-PP: Regional posterior inclusion probability, i.e., probability of being a risk locus

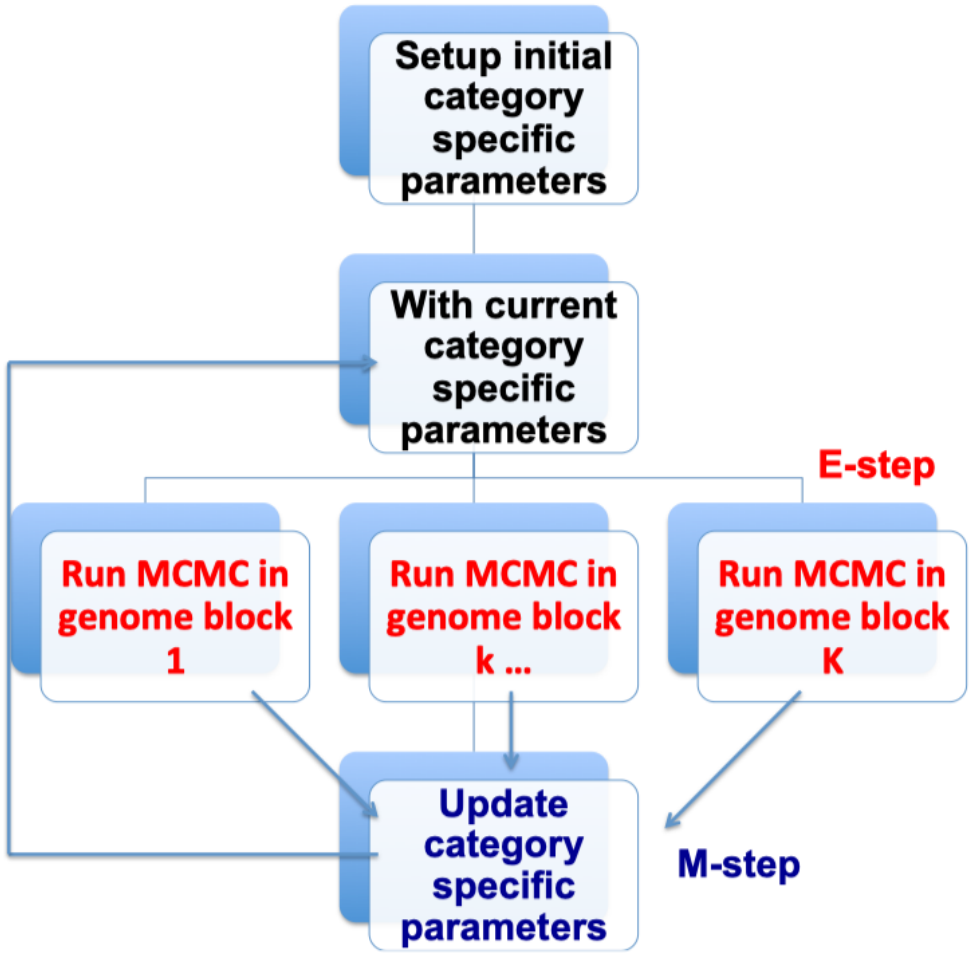
Bayesian Hierarchical Model

- Hierarchical priors
  - $\pi_q \sim Beta(a_q, b_q)$;
  - $\sigma_q^2 \sim InverseGamma(k_1, k_2)$;
  - $\tau \sim Gamma(k_3, k_4)$
- The joint posterior distribution

$$P(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma}^2, \boldsymbol{\pi}, \tau | \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{A}) \propto \tag{2}$$
$$P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \tau) P(\boldsymbol{\beta}|\boldsymbol{A}, \boldsymbol{\pi}, \boldsymbol{\sigma}^2, \tau) P(\boldsymbol{\gamma}|\boldsymbol{\pi}) P(\boldsymbol{\pi}) P(\boldsymbol{\sigma}^2) P(\tau),$$

- Product of Likelihood and Priors
- Challenges of Standard MCMC: memory usage and convergence rate

EM-MCMC Algorithm



Enabled genome-wide analysis

Improved MCMC convergence rate

MCMC Algorithm

Given category-specific parameters (π_q, σ_q^2) and residual variance τ^{-1}:

- Propose a new indicator vector γ
- Calculate conditional posterior likelihood

$$P(\gamma|Y,X) \propto |\Omega|^{-1/2} \exp\left\{\frac{\tau}{2} \boldsymbol{Y}^{\boldsymbol{T}} \boldsymbol{X}_{|\gamma|} V_\gamma \Omega^{-1} \boldsymbol{X}_{|\gamma|}^{\boldsymbol{T}} \boldsymbol{Y}\right\}, \Omega = V_{|\gamma|} \boldsymbol{X}_{|\gamma|}^{\boldsymbol{T}} \boldsymbol{X}_{|\gamma|} + \boldsymbol{I}$$

- Apply Metropolis-Hastings algorithm
- If accepted, update effect-size estimates:

$$\widehat{\beta}_{|\gamma|} = \left[X_{|\gamma|}^T X_{|\gamma|} + V_\gamma^{-1}\right]^{-1} X_{|\gamma|}^T Y$$

- Summary statistics $(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}, \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y})$ can be used here to save computational cost

EM Updates

MAPs (maximum a posteriori estimates):

Let $\widehat{\gamma_{jq}} = E[\gamma_{jq}]$

- Causal probability per annotation

$$\widehat{\pi_q} = \frac{\sum_{j_q=1}^{m_q} \widehat{\gamma_{j_q}} + a_q - 1}{m_q + a_q + b_q - 2}$$

- Effect-size variance per annotation

$$\widehat{\sigma_q^2} = \frac{\tau \sum_{j_q=1}^{m_q} (\widehat{\gamma_{j_q}} \widehat{\beta_{j_q}^2}) + 2k_2}{\sum_{j_q=1}^{m_q} \widehat{\gamma_{j_q}} + 2(k_1 + 1)}$$

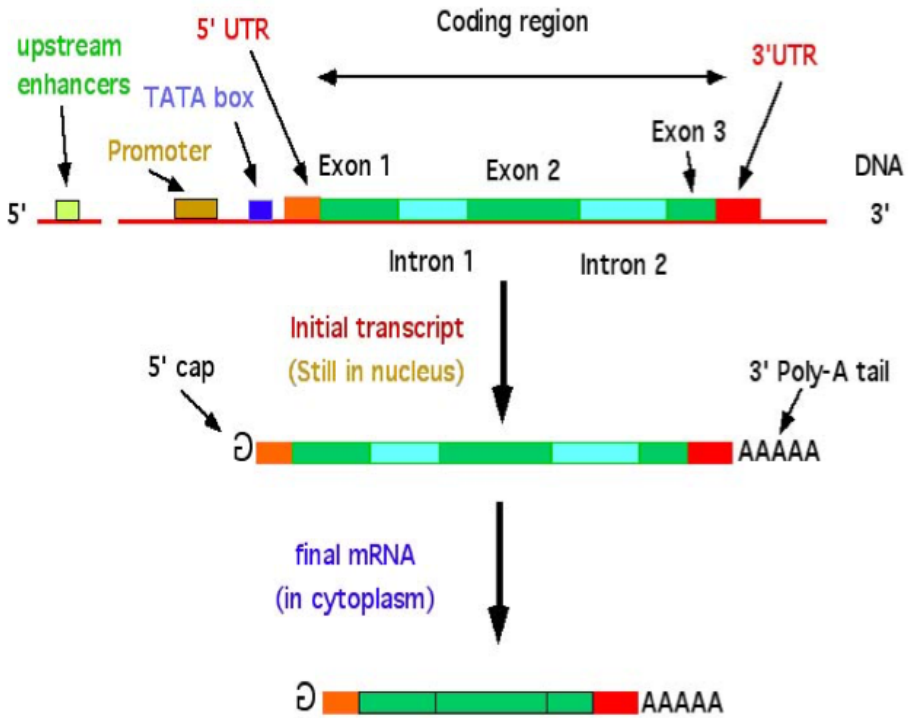
Apply BFGWAS to Study Age-related Macular Degeneration (AMD)

- ∼10M low-frequency and common variants (MAF>0.5%)
- ∼ 16K cases and ∼18K controls (unrelated European)
- Phenotypes adjusted for age, gender, DNA source, and first 2 principal components
- GWAS results with gene-based annotations

Gene-based Annotations

Annotated by SeattleSeq:

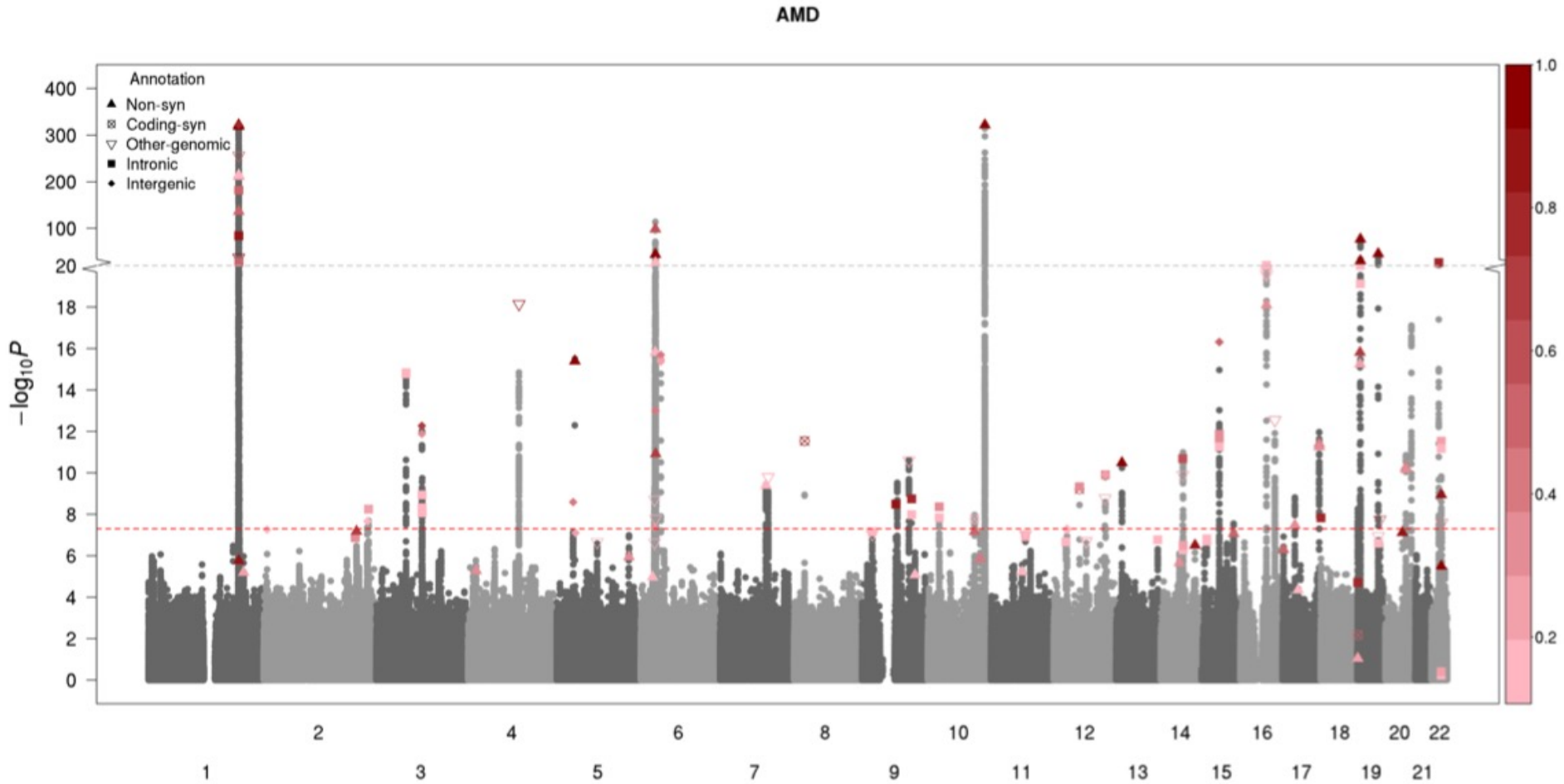
- Non-synonymous (42,005)
- Synonymous (67,165)
- Intronic (3,679,235)
- Intergenic (5,512,423)
- Other genomic (565,916, UTR, non-coding exons, upstream and downstream)



http://nitro.biosci.arizona.edu/

BFGWAS Results with Gene-based Annotations

Colored variants with Bayesian PPs > 0.1068 ($\sim$p-value < 5×10^{-8}).



BFGWAS Results with Gene-based Annotations

By Bayesian PP >0.1068, our method identified 150 variants with association evidence

	Non-syn	Coding-syn	Intronic	Intergenic	Other-genomic
Associations	47	4	54	18	27
Enrichment	72x	4x	0.9x	0.2x	3x

By Regional-PP > 0.95, our method identified 5 potentially novel loci, in addition to 32 known loci (Fritsche LG et al., 2016)

5 Potentially Novel Loci

Annotation	SNP/Gene	Previous Associations
Missense	*rs7562391/PPIL3*	
Missense	*rs61751507/CPN1*	Age-related Hearing Impairment (Fransen E et al., 2015)
Missense	*rs2232613/LBP*	Encodes Lipid Transfer Protein (Masson D et al., 2009)
Downstream	*rs114348558/ZNRD1-AS1*	Lipid Metabolisms (Kettunen J et al., 2012)
Splice	*rs6496562/ABHD2*	Coronary Artery Disease (Nikpay M et al., 2015)

- Known AMD risk loci *CETP, APOE, and LIPC* are also associated with Lipid Metabolisms and Coronary Artery Disease (Kettunen J et al., 2012, Nikpay M et al., 2015)
- Known AMD risk loci *CETP* is part of the Lipid Transfer Protein family (Masson D et al., 2009)

LocusZoom plots around the **Non-synonymous** SNP *rs4151667* (purple triangle).

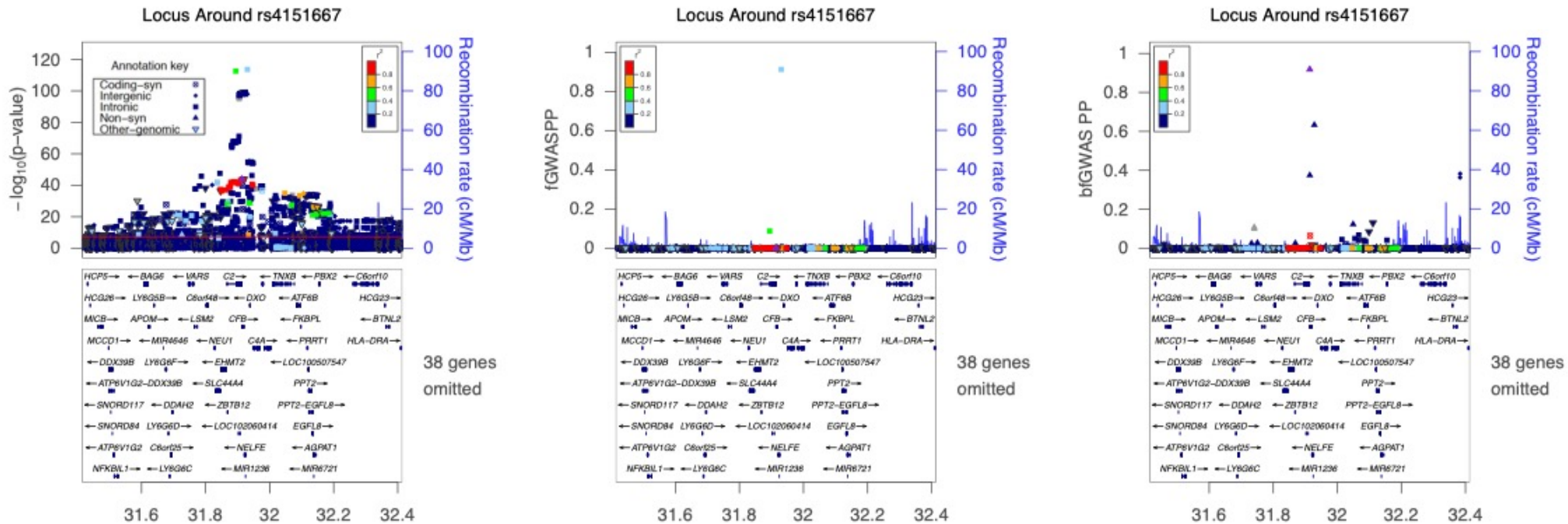


Figure 3: GWAS (left) vs. FGWAS (middle; Pickrell JK, AJHG 2014) vs. BFGWAS (right) for example locus #8.

Model Comparison

- Model1: top 2 SNPs (Intronic) by sequential forward selection
- Model2: top 2 SNPs (Non-synonymous) by BFGWAS

	Model1	Model2	Difference
AIC	95,857.36	**95,752.63**	104.73
BIC	95,891.1	**95,786.36**	104.74
−Log-likelihood	47,924.68	**47,872.31**	52.37

Haplotype Analysis

Haplotype with lead SNP *rs116503776* from standard GWAS and top 2 SNPs *rs4151667, rs115270436* by BFGWAS

rs116503776 *SKIV2L*	*rs4151667* *CFB*	*rs115270436* *SKIV2L*	Freq	OddsRatio	P-value
A	A	G	0.3%	0.364	8.9×10^{-11}
A	T	G	6.6%	0.522	1.5×10^{-86}
A	A	A	3.2%	0.561	5.0×10^{-36}
A	T	A	1.7%	**1.102**	$\mathbf{9.2 \times 10^{-2}}$
G	T	A	**87.8%**	-	Reference

Haplotype analysis by Fritsche LG et al. (2016) also found *rs116503776/SKIV2L* tags two previously identified **Non-synonymous** SNPs *rs4151667/CFB, rs641153/CFB*.

Enrichment Results

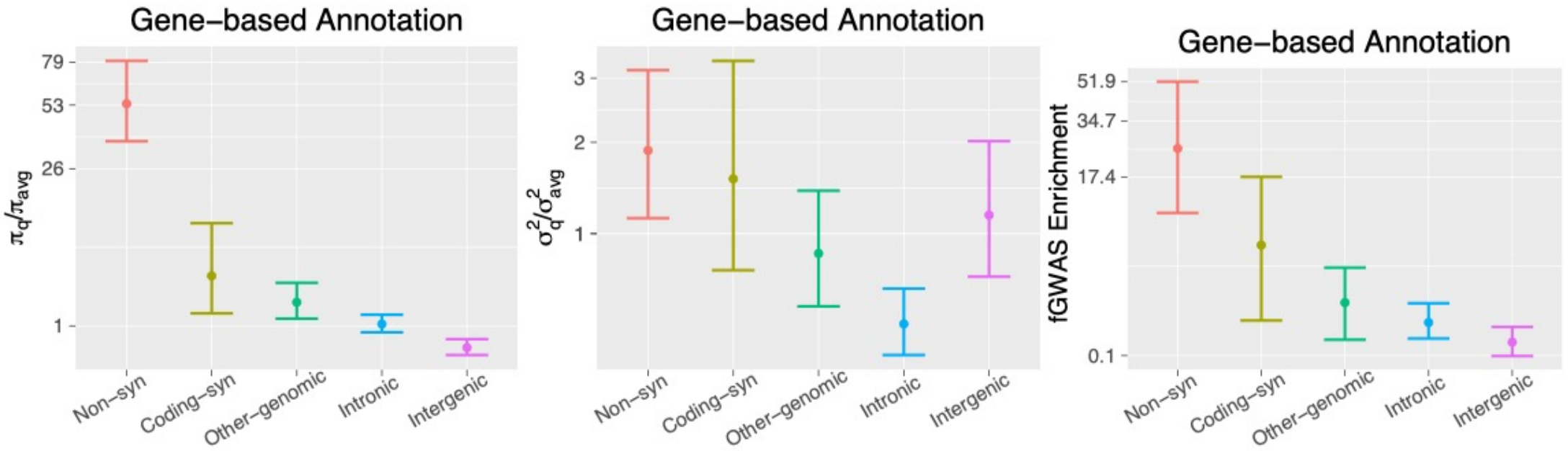


Figure 5: BFGWAS enrichment Results (left, middle) vs. FGWAS (right).

Available Tools

- FINEMAP: Fine-mapping GWAS results: http://www.christianbenner.com

- SuSIE: https://stephenslab.github.io/susieR/
  - Extend for using summary data in a follow-up 2022 paper: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1010299

- BFGWAS_QUANT: https://github.com/yanglab-emory/BFGWAS_QUANT
  - Based on the BVSR model for Bayesian functional GWAS
  - Account for multivariate quantitative annotations