香港城市大學
City University of Hong Kong

Advanced Approaches to Genetic Association Analysis for High-Dimensional Annotation Integration and Single-Cell eQTL Mapping

Speaker: LI Yuekai

Supervisor: CAI Mingxuan

CONTENTS

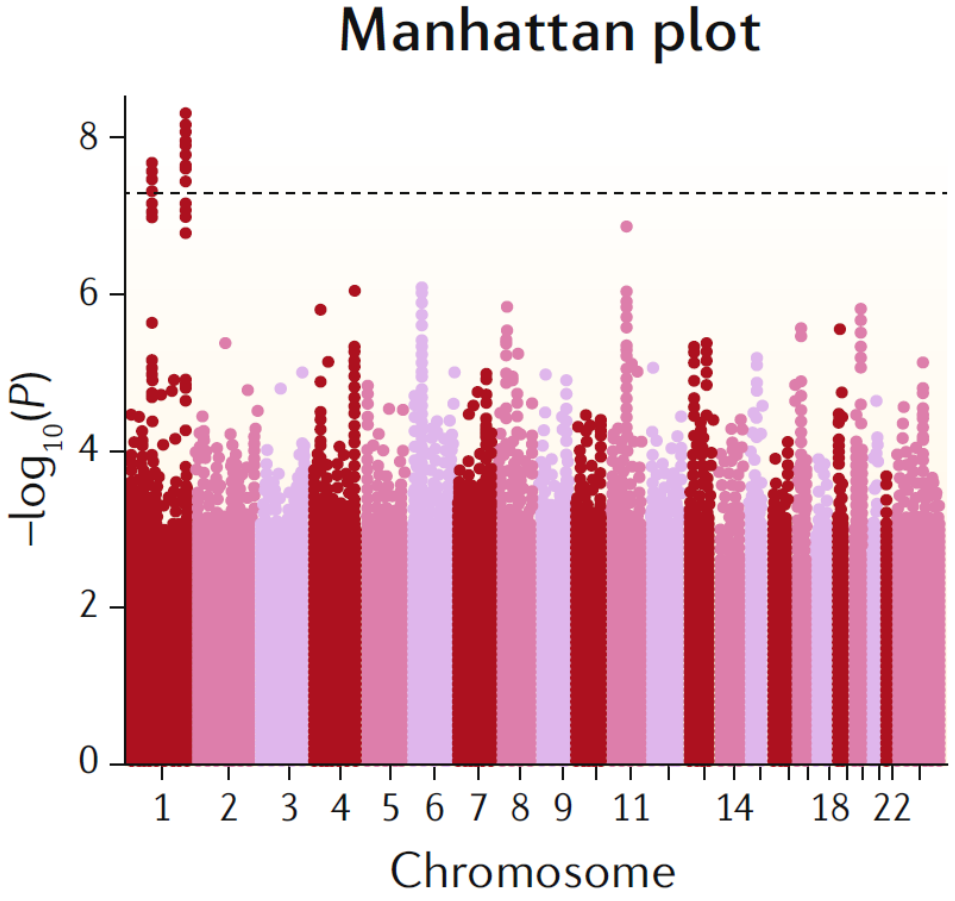
香港城市大學
City University of Hong Kong

香港城市大學
City University of Hong Kong

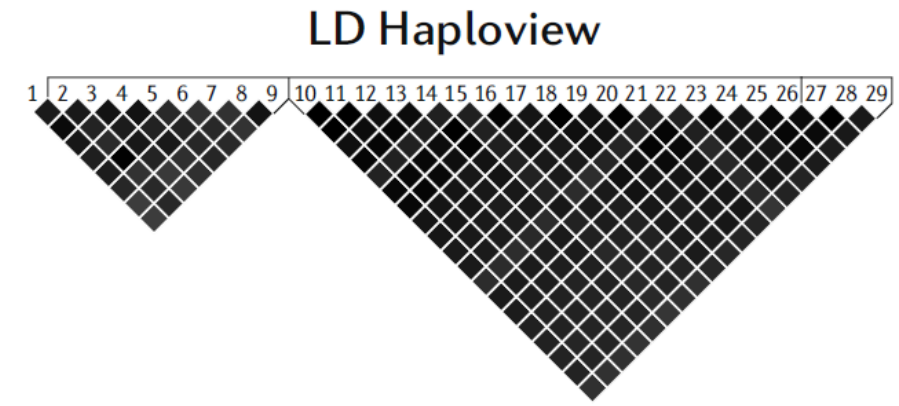
01

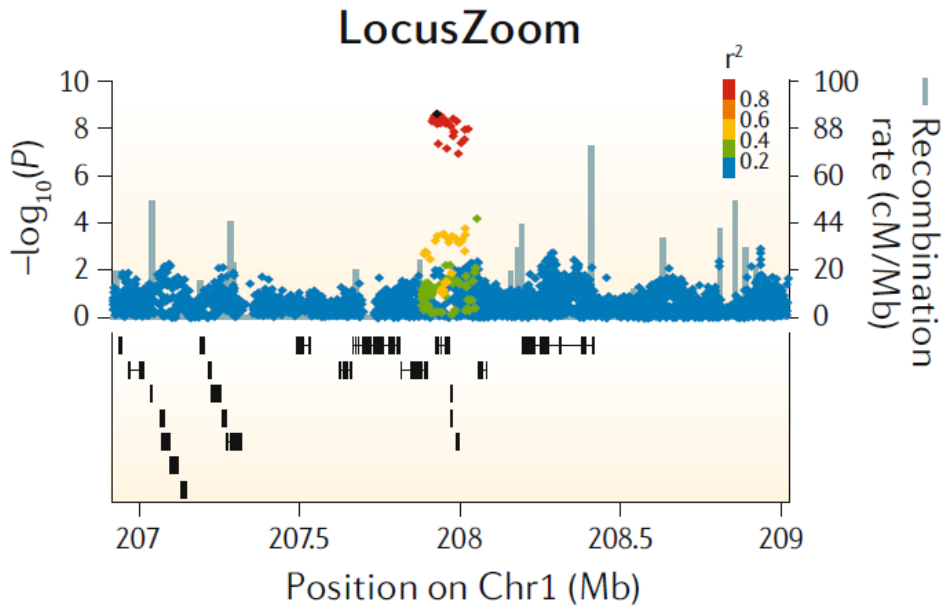
Introduction

GWAS and Linkage Disequilibrium (LD)

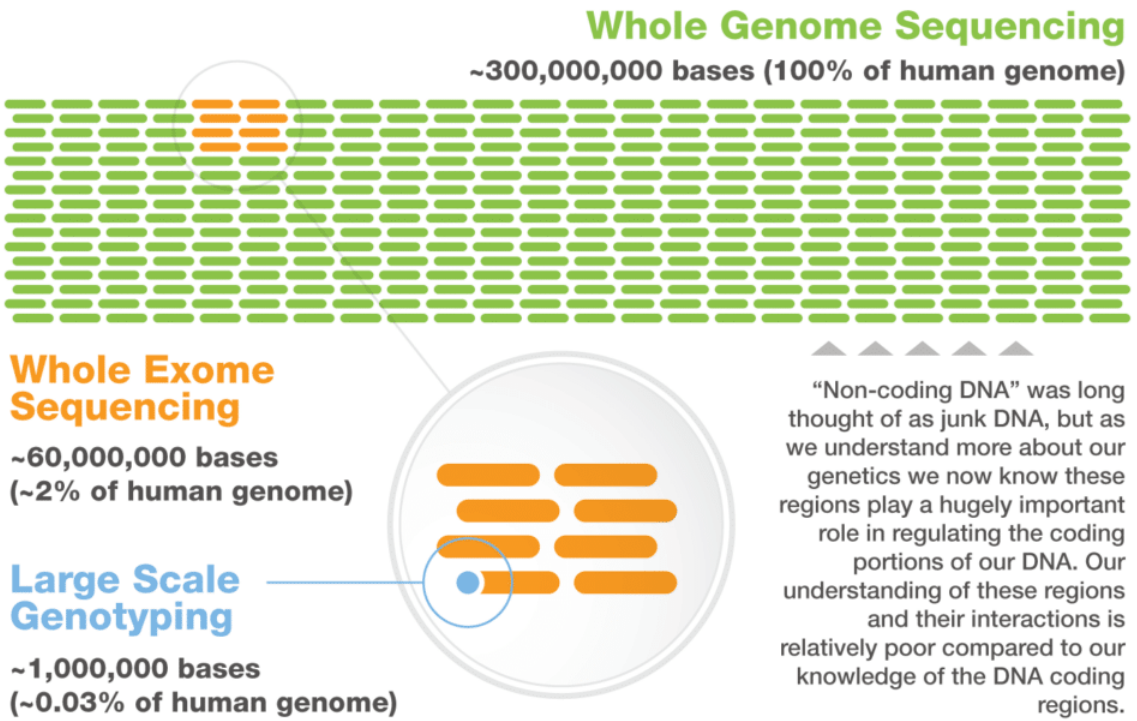


https://doi.org/10.1038/s41576-018-0016-z



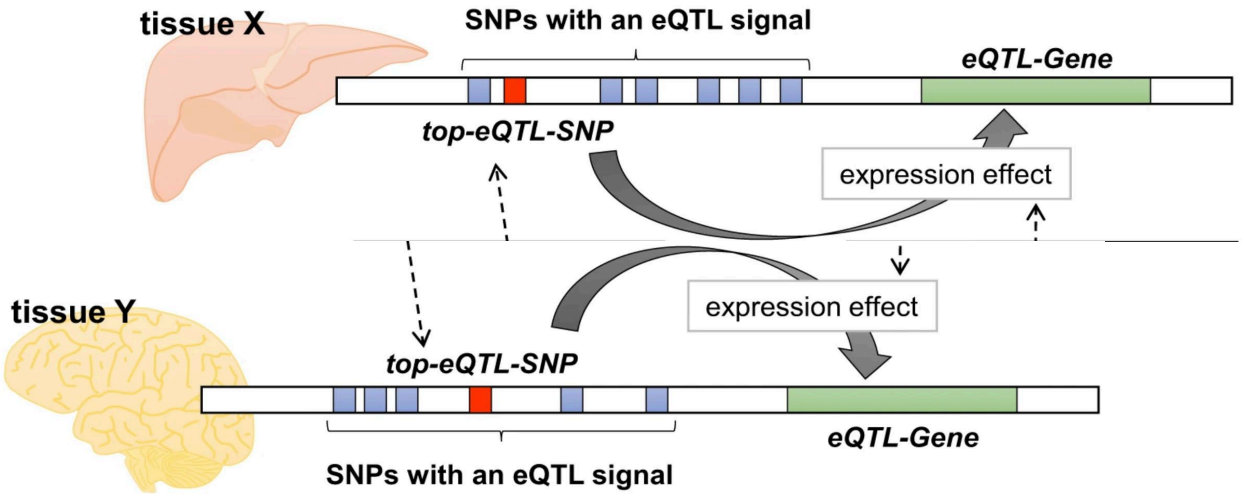


SNPs in non-coding regions and complex regulatory mechanisms



- Most SNPs are located in the non-coding regions of DNA.

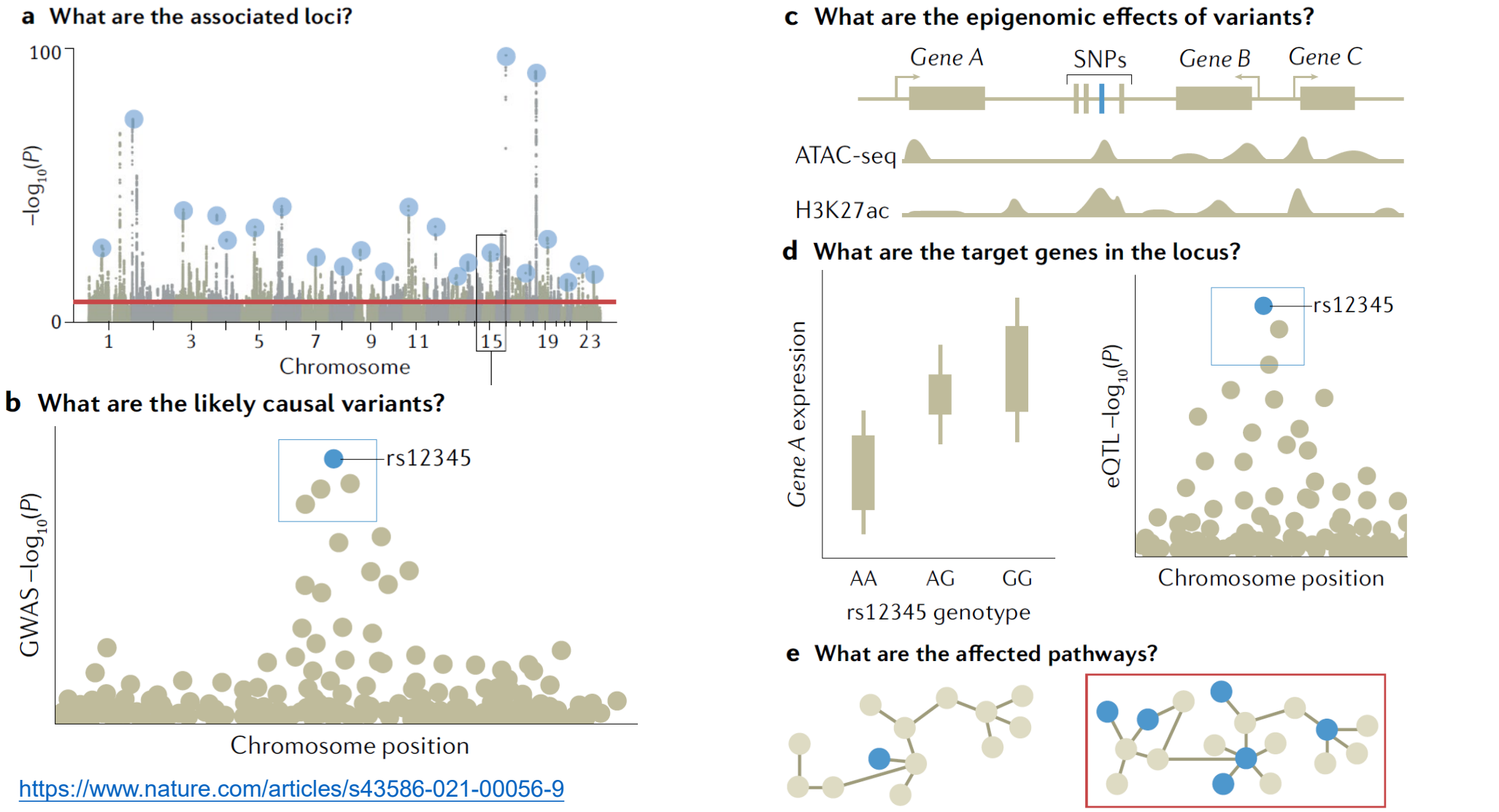
https://www.mygenefood.com/blog/finding-best-dna-test-genotype-sequence/



- SNPs in different tissues may have different regulatory effects on gene expression.

https://www.nature.com/articles/s41431-019-0468-4

Fine-mapping and downstream analysis



https://www.nature.com/articles/s43586-021-00056-9

The Sum of Single Effects (SuSiE) model

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

$$\boldsymbol{e} \sim N_n(0, \sigma^2 I_n)$$

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

$$\boldsymbol{b}_l = \boldsymbol{\gamma}_l b_l$$

$$\boldsymbol{\gamma}_l \sim \mathrm{Mult}(1, \boldsymbol{\pi})$$

$$\mathrm{b}_l \sim N_1(0, \sigma_{0l}^2)$$

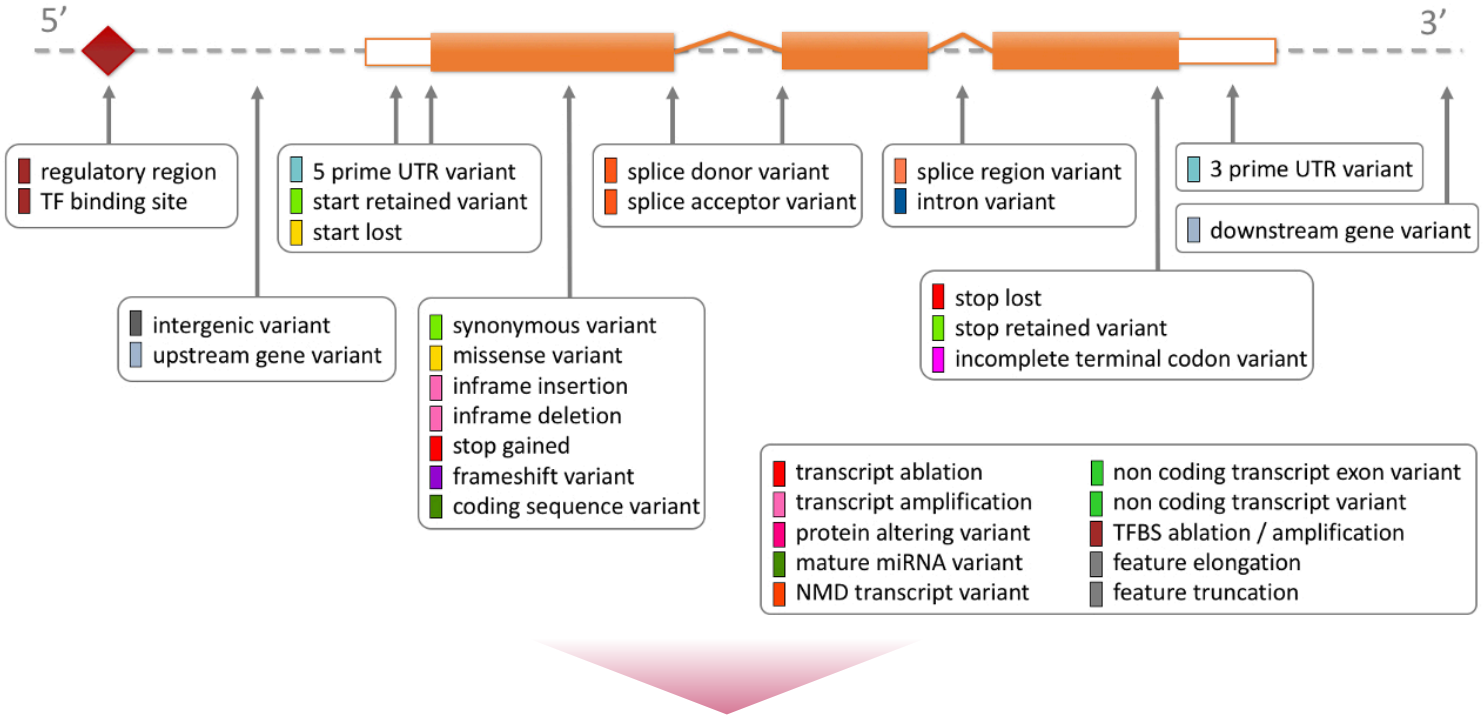
- The variational “sum of single effects” assumption treats causal signals as independent, giving simple updates and credible sets.
- By iteratively adding single-effect components, the model can capture multiple causal variants even in complex LD patterns.
- Variational EM makes it faster than MCMC, and the framework is straightforward to extend with additional priors or phenotypes.

香港城市大學
City University of Hong Kong

02

Funmap: integrating high-dimensional functional annotations to improve complex traits fine-mapping

Introduction of functional annotations



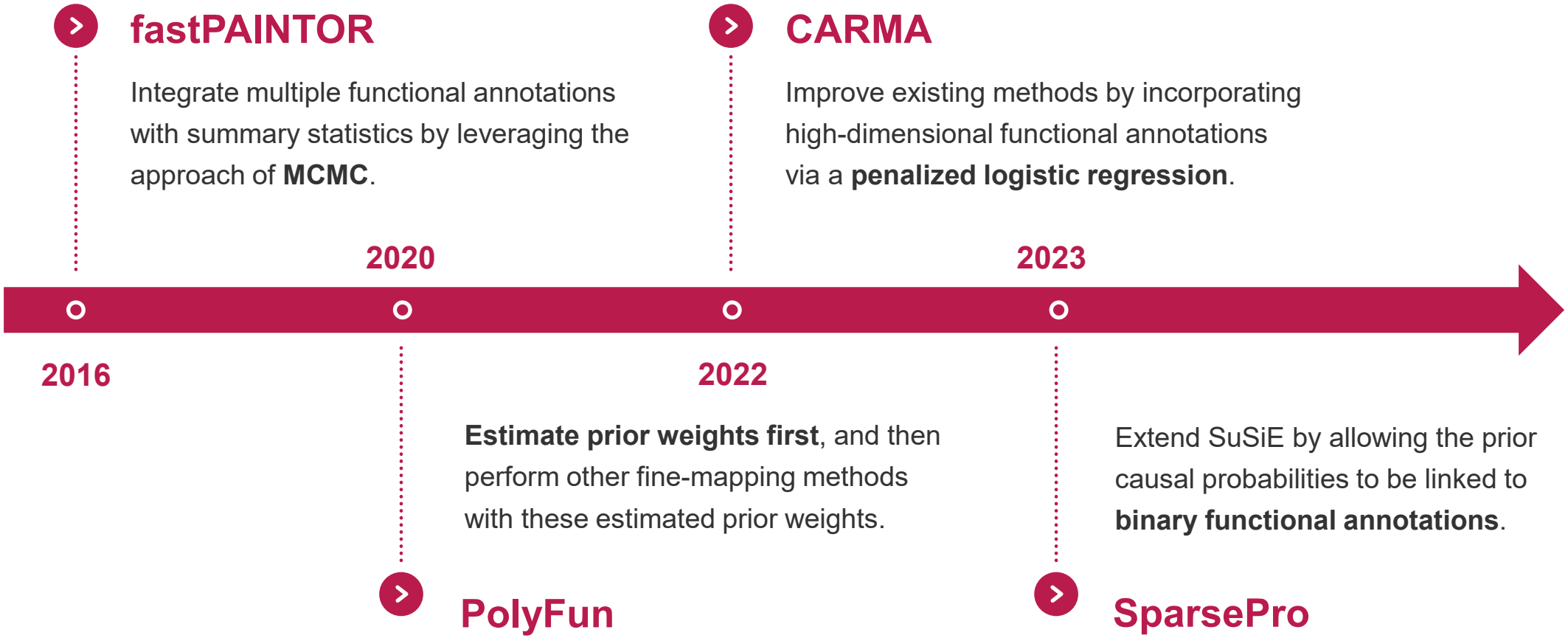
Example of Baseline-LF v2.2. UKB annotations

	m_1	m_2	⋯	m_{187}
rs1124048	0	0	…	11.95
rs10494829	0	-0.205	…	11.95
rs4915210	1	0.557	…	11.85
rs56368827	1	0	…	12.0
⋮	⋮	⋮	…	⋮
rs3738255	0	-2.079	…	11.95
rs296568	0	0	…	12.0
rs296567	1	0	…	12.4
rs296566	0	0.966	…	12.0

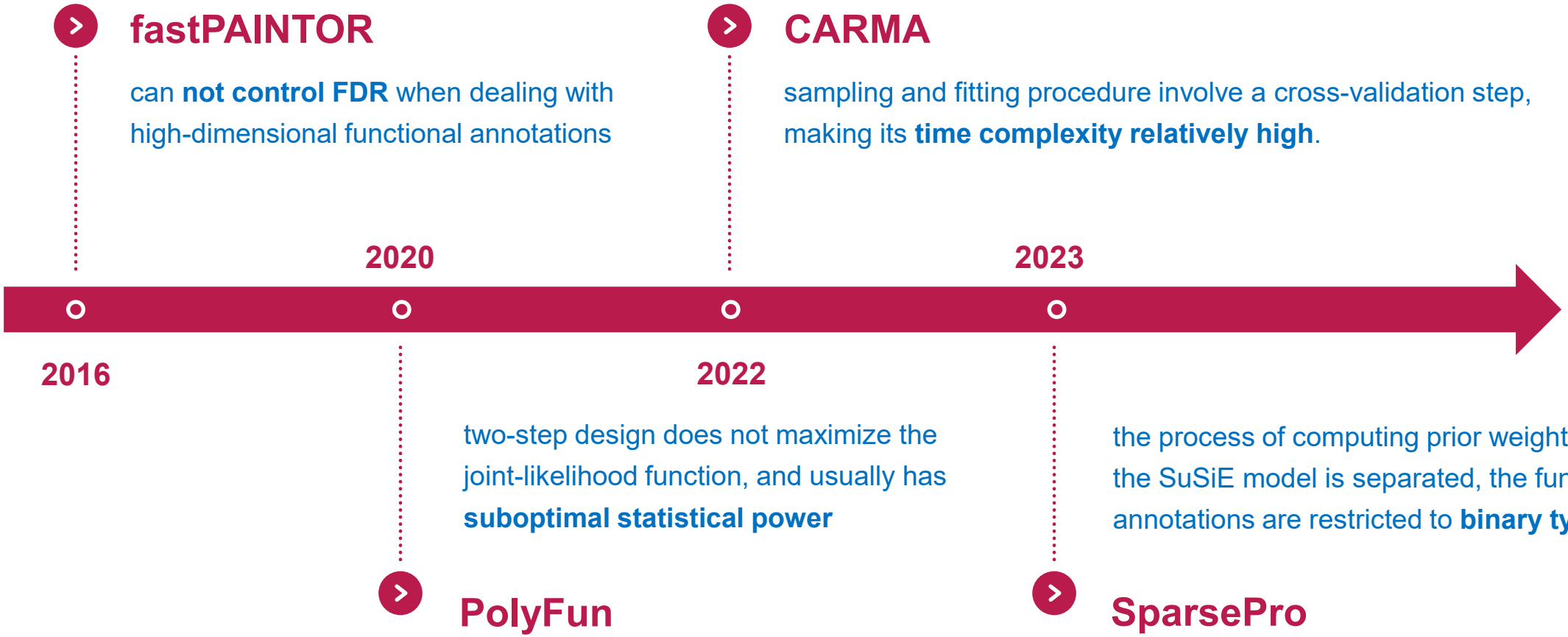
Functional annotations offer promising auxiliary information for fine-mapping

https://asia.ensembl.org/info/genome/variation/prediction/predicted_data.html

Existing fine-mapping methods with functional annotations



Existing fine-mapping methods with functional annotations



The Funmap model

We relate the phenotype $\mathbf{y}$ to genotypes $\mathbf{X}$ with the linear model, and we consider the following sum-of-single-effects structure:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e},\ \mathbf{e} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

$$\mathbf{b} = \sum_{l=1}^{L} \boldsymbol{\gamma}_l \cdot b_l,\ b_l \sim \mathcal{N}_1(0, \sigma_{bl}^2),\ \boldsymbol{\gamma}_l \sim \mathrm{Mult}(1, \boldsymbol{\pi}_l)$$

- $\mathbf{b} \in \mathbf{R}^p$ is the sparse vector of SNP effect sizes; b_l is the effect size of the l-th causal signal;
- $\boldsymbol{\gamma}_l$ is a binary vector with $\gamma_{lj} = 1$ indicating the l-th causal signal is attributed to the j-th SNP;
- $\boldsymbol{\pi}_l$ is the vector of prior causal probabilities with $\sum_{j=1}^{p} \pi_{lj} = 1$.

We link the prior probability $\boldsymbol{\pi}_l$ to SNP annotations $\mathbf{A}$ with the following softmax model with random effects:

$$\boldsymbol{\pi}_l = \mathrm{softmax}(\mathbf{A}\mathbf{w}_l),\ \mathbf{w}_l \sim \mathcal{N}_m(\mathbf{0}, \sigma_{wl}^2\mathbf{I}_m)$$

where $\mathbf{w}_l \in \mathbf{R}^m$ represents the random effects of annotations on the causal probability of the l-th component.

The Funmap model

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e},\ \mathbf{e} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

$$\mathbf{b} = \sum_{l=1}^{L} \mathbf{b}_l,\ \mathbf{b}_l = \boldsymbol{\gamma}_l \cdot b_l,\ b_l \sim \mathcal{N}_1(0, \sigma_{bl}^2)$$

$$\boldsymbol{\gamma}_l = \mathrm{Mult}(1, \boldsymbol{\pi}_l),\ \boldsymbol{\pi}_l = \mathrm{softmax}(\mathbf{A}\mathbf{w}_l),\ \mathbf{w}_l \sim \mathcal{N}_m(\mathbf{0}, \sigma_{wl}^2\mathbf{I}_m)$$

We propose a **component-specific random effects** assumption on the annotation weights w_l

Funmap allows the **annotation weights to vary across causal signals**, which better characterizes the real genetic architecture.

Funmap can adaptively estimate the annotation weights w_l from the data while **shrinking those of redundant annotations to avoid over-fitting**.

Fitting SuSiE: Iterative Bayesian Stepwise Selection (IBSS) algorithm

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 $\bar{\mathbf{b}}_l = 0$, for $l = 1, \ldots, L$ ▷ other initializations are possible (see algorithm 3 in the on-line appendix B)
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 $\boldsymbol{\alpha}_1, \boldsymbol{\mu}_{11}, \boldsymbol{\sigma}_{11}, \ldots, \boldsymbol{\alpha}_L, \boldsymbol{\mu}_{1L}, \boldsymbol{\sigma}_{1L}$

Fitting SuSiE: Iterative Bayesian Stepwise Selection (IBSS) algorithm

Factorization assumption:

$$q(\mathbf{b}_1,\ldots,\mathbf{b}_L)=\prod_{l=1}^{L} q_l(\mathbf{b}_l)$$

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 $\text{SER}(\mathbf{X},\mathbf{y};\sigma^2,\sigma_0^2)\rightarrow(\boldsymbol{\alpha},\boldsymbol{\mu}_1,\boldsymbol{\sigma}_1)$ that computes the posterior distribution for $\mathbf{b}_l$ under the SER model; see expression (2.11)

1, initialize posterior means
   $\bar{\mathbf{b}}_l=0$, for $l=1,\ldots,L$ ▷ other initializations are possible (see algorithm 3 in the on-line appendix B)
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 $\boldsymbol{\alpha}_1,\boldsymbol{\mu}_{11},\boldsymbol{\sigma}_{11},\ldots,\boldsymbol{\alpha}_L,\boldsymbol{\mu}_{1L},\boldsymbol{\sigma}_{1L}$

$$\arg\max_{q_l} F(q_1,\ldots,q_L;\sigma^2,\boldsymbol{\sigma}_0^2)=\text{SER}(\mathbf{X},\mathbf{r}_l;\sigma^2,\sigma_{0l}^2)$$

IBSS is a coordinate ascent algorithm for maximizing the ELBO F over q satisfying the approximation.

The Evidence Lower Bound (ELBO) of Funmap

log marginal likelihood:

$$\log \Pr(\mathbf{y} \mid \mathbf{X}, \mathbf{A}; \theta) = \log \sum_{\tilde{\gamma}} \int_{\tilde{\mathrm{b}}} \int_{\tilde{\mathbf{w}}} \Pr(\mathbf{y}, \tilde{\mathrm{b}}, \tilde{\gamma}, \tilde{\mathbf{w}} \mid \mathbf{X}, \mathbf{A}; \theta) \, \mathrm{d}\tilde{\mathbf{w}} \, \mathrm{d}\tilde{\mathrm{b}}$$

(*Jensen's inequality*)

$$\geq \sum_{\tilde{\gamma}} \int_{\tilde{\mathrm{b}}} \int_{\tilde{\mathbf{w}}} q(\tilde{\mathrm{b}}, \tilde{\gamma}, \tilde{\mathbf{w}}) \log \frac{\Pr(\mathbf{y}, \tilde{\mathrm{b}}, \tilde{\gamma}, \tilde{\mathbf{w}} \mid \mathbf{X}, \mathbf{A}; \theta)}{q(\tilde{\mathrm{b}}, \tilde{\gamma}, \tilde{\mathbf{w}})} \, \mathrm{d}\tilde{\mathbf{w}} \, \mathrm{d}\tilde{\mathrm{b}}$$

$$= \mathbb{E}_q \left[\log \Pr(\mathbf{y}, \tilde{\mathrm{b}}, \tilde{\gamma}, \tilde{\mathbf{w}} \mid \mathbf{X}, \mathbf{A}; \theta) \right] - \mathbb{E}_q \left[\log q(\tilde{\mathrm{b}}, \tilde{\gamma}, \tilde{\mathbf{w}}) \right]$$

log joint likelihood can be decomposed as follows:

$$\sum_{l=1}^{L} \{ \log \Pr(\mathbf{y} \mid \mathbf{X}, \mathrm{b}_l, \gamma_l, \mathbf{w}_l; \sigma^2) + \log \Pr(\mathrm{b}_l, \gamma_l \mid \mathbf{A}, \mathbf{w}_l; \sigma_{bl}^2) + \log \Pr(\mathbf{w}_l; \sigma_{wl}^2) \}$$

factorizable mean-field assumption:

$$q(\tilde{\mathrm{b}}, \tilde{\gamma}, \tilde{\mathbf{w}}) = \prod_{l=1}^{L} q_l(\tilde{\mathrm{b}}, \tilde{\gamma}) \cdot q_l(\tilde{\mathbf{w}}_l)$$

It is difficult to exactly evaluate the integration due to the softmax function.

The Evidence Lower Bound (ELBO) of Funmap

In 2007, *Bouchard* proposed an upper bound on the logarithm of the sum of exponential functions:

$$\log \sum_{k=1}^{K} e^{x_k} \leq \alpha + \sum_{k=1}^{K} \left[\frac{x_k - \alpha - \xi_k}{2} + \lambda(\xi_k) \cdot \left((x_k - \alpha)^2 - \xi_k^2 \right) + \log\left(1 + e^{\xi_k}\right) \right]$$

where $\lambda(\xi_{lj}) = \frac{1}{2\xi_{lj}} \left(\frac{1}{1+\exp(-\xi_{lj})} - \frac{1}{2} \right), \xi_{lj} \geq 0, \rho_l \in \mathbf{R}$ are the parameters of the inequality

Applying this inequality to the denominator of the softmax function leads to a tractable bound with quadratic form of w_l:

$$\begin{aligned}
\log \Pr(\mathrm{b}_l, \boldsymbol{\gamma}_l \mid \mathbf{A}, \mathbf{w}_l; \sigma_{bl}^2) &= -\frac{1}{2}\log\left(2\pi\sigma_{b_l}^2\right) - \frac{1}{2\sigma_{b_l}^2} b_l^2 + \sum_{j=1}^{p} \gamma_{lj} \cdot \log \frac{\exp\left(\mathbf{A}_j^T \tilde{\mathbf{w}}_l\right)}{\sum_{j'=1}^{p} \exp\left(\mathbf{A}_{j'}^T \tilde{\mathbf{w}}_l\right)} \\
&\geq -\frac{1}{2}\log\left(2\pi\sigma_{b_l}^2\right) - \frac{1}{2\sigma_{b_l}^2} b_l^2 + \sum_{j=1}^{p} \gamma_{lj} \cdot A_j^T w_l \\
&\quad - \sum_{j=1}^{p} \gamma_{lj} \left(\rho_l + \sum_{j'=1}^{p} \frac{A_{j'}^T w_l - \rho_l - \xi_{lj'}}{2} + \sum_{j'=1}^{p} \lambda(\xi_{lj'}) \left((A_{j'}^T w_l - \rho_l)^2 - \xi_{lj'}^2 \right) + \sum_{j'=1}^{p} \log\left(1 + e^{\xi_{lj'}}\right) \right)
\end{aligned}$$

The Evidence Lower Bound (ELBO) of Funmap

log marginal likelihood:

$$\log \Pr(\mathbf{y} \mid \mathbf{X}, \mathbf{A}; \theta) = \log \sum_{\tilde{\gamma}} \int_{\tilde{\mathrm{b}}} \int_{\tilde{\mathbf{w}}} \Pr(\mathbf{y}, \tilde{\mathrm{b}}, \tilde{\gamma}, \tilde{\mathbf{w}} \mid \mathbf{X}, \mathbf{A}; \theta) \, \mathrm{d}\tilde{\mathbf{w}} \, \mathrm{d}\tilde{\mathrm{b}}$$

(*Jensen's inequality*)

$$\geq \sum_{\tilde{\gamma}} \int_{\tilde{\mathrm{b}}} \int_{\tilde{\mathbf{w}}} q(\tilde{\mathrm{b}}, \tilde{\gamma}, \tilde{\mathbf{w}}) \log \frac{\Pr(\mathbf{y}, \tilde{\mathrm{b}}, \tilde{\gamma}, \tilde{\mathbf{w}} \mid \mathbf{X}, \mathbf{A}; \theta)}{q(\tilde{\mathrm{b}}, \tilde{\gamma}, \tilde{\mathbf{w}})} \, \mathrm{d}\tilde{\mathbf{w}} \, \mathrm{d}\tilde{\mathrm{b}}$$

$$= \mathbb{E}_q \left[\log \Pr(\mathbf{y}, \tilde{\mathrm{b}}, \tilde{\gamma}, \tilde{\mathbf{w}} \mid \mathbf{X}, \mathbf{A}; \theta) \right] - \mathbb{E}_q \left[\log q(\tilde{\mathrm{b}}, \tilde{\gamma}, \tilde{\mathbf{w}}) \right]$$

log joint likelihood can be decomposed as follows:

$$\sum_{l=1}^{L} \{ \log \Pr(\mathbf{y} \mid \mathbf{X}, \mathrm{b}_l, \gamma_l, \mathbf{w}_l; \sigma^2) + \log \Pr(\mathrm{b}_l, \gamma_l \mid \mathbf{A}, \mathbf{w}_l; \sigma_{bl}^2) + \log \Pr(\mathbf{w}_l; \sigma_{wl}^2) \}$$

factorizable mean-field assumption:

$$q(\tilde{\mathrm{b}}, \tilde{\gamma}, \tilde{\mathbf{w}}) = \prod_{l=1}^{L} q_l(\tilde{\mathrm{b}}, \tilde{\gamma}) \cdot q_l(\tilde{\mathbf{w}}_l)$$

Define the lower bound of log complete-data likelihood as: $\log f(\mathbf{y}, \tilde{\mathrm{b}}, \tilde{\gamma}, \tilde{\mathbf{w}} \mid \mathbf{X}, \mathbf{A}; \boldsymbol{\theta})$

Define the evidence lower bound as: $F(q; \Theta) = \mathbb{E}_q \left[\log f(\mathbf{y}, \tilde{\mathrm{b}}, \tilde{\gamma}, \tilde{\mathbf{w}} \mid \mathbf{X}, \mathbf{A}; \boldsymbol{\theta}) - \log q(\tilde{\mathrm{b}}, \tilde{\gamma}, \tilde{\mathbf{w}}) \right]$

The posterior distribution under the factorizable mean-field assumption

$$\begin{aligned}\log q_l(b_l, \boldsymbol{\gamma}_l) &= \mathbb{E}_{\mathbf{w}_l}\left[\log f(\mathbf{y}, \tilde{\mathrm{b}}, \tilde{\boldsymbol{\gamma}}, \tilde{\mathbf{w}} \mid \mathbf{X}, \mathbf{A}; \boldsymbol{\theta})\right] \\ &= -\frac{1}{2\sigma^2}\left(-2\, b_l\, \boldsymbol{\gamma}_l^T \mathbf{X}^T \mathbf{y} + b_l^2\, \boldsymbol{\gamma}_l^T \mathbf{X}^T \mathbf{X}\, \boldsymbol{\gamma}_l\right) - \frac{1}{2\sigma_{b_l}^2} b_l^2 + \text{const} \\ &= \left(-\frac{1}{2\sigma^2}\, \boldsymbol{\gamma}_l^T \mathbf{X}^T \mathbf{X}\, \boldsymbol{\gamma}_l - \frac{1}{2\sigma_{b_l}^2}\right) b_l^2 + \frac{1}{\sigma^2}\, \boldsymbol{\gamma}_l^T \mathbf{X}^T \mathbf{y}\, b_l + \text{const}\end{aligned}$$

$$q_l(\mathrm{b}_l, \boldsymbol{\gamma}_l) \sim \prod_{j=1}^{p} (\alpha_{lj} \cdot \mathcal{N}(\mu_{blj}, s_{blj}^2))^{\gamma_{lj}}$$

$$\begin{aligned}\mathrm{PIP}_{(j)} &= 1 - \prod_{l=1}^{L}(1 - q(\gamma_{lj} = 1)) \\ &= 1 - \prod_{l=1}^{L}(1 - \alpha_{lj})\end{aligned}$$

$$\begin{aligned}\log q_l(\mathbf{w}_l) &= \mathbb{E}_{b_l, \boldsymbol{\gamma}_l}\left[\log f(\mathbf{y}, \tilde{\mathrm{b}}, \tilde{\boldsymbol{\gamma}}, \tilde{\mathbf{w}} \mid \mathbf{X}, \mathbf{A}; \boldsymbol{\theta})\right] \\ &= \mathbf{w}_l^T \left(-\frac{1}{2\sigma_{w_l}^2}\, \mathbf{I}_M - \sum_{j=1}^{p} \lambda(\xi_{lj})\, \mathbf{A}_j \mathbf{A}_j^T\right) \mathbf{w}_l \\ &\quad + \mathbf{w}_l^T \sum_{j=1}^{p} \left(\alpha_{lj} - \tfrac{1}{2} + 2\,\lambda(\xi_{lj})\, \alpha_{lj}\right) \mathbf{A}_j + \text{const}\end{aligned}$$

$$q_l(\mathbf{w}_l) \sim \mathcal{N}(\boldsymbol{\mu}_{wl}, \boldsymbol{\Sigma}_{wl})$$

Update parameters of Funmap in M-steps

Update the variational parameters ρ_l, ξ_{lj}^2 :

$$\xi_{lj}^2 = \mathbf{A}_j^T\left(\boldsymbol{\Sigma}_{wl} + \boldsymbol{\mu}_{wl}\boldsymbol{\mu}_{wl}^T\right)\mathbf{A}_j + \rho_l^2 - 2\,\rho_l\,\mathbf{A}_j^T\boldsymbol{\mu}_{wl}$$

$$\rho_l = \frac{\frac{1}{2}\left(\frac{p}{2} - 1\right) + \sum_{j=1}^{p}\lambda\left(\xi_{lj}\right)\mathbf{A}_j^T\boldsymbol{\mu}_{wl}}{\sum_{j=1}^{p}\lambda\left(\xi_{lj}\right)}$$

Update the prior variance parameters $\sigma_{bl}^2, \sigma_{wl}^2$:

$$\sigma_{b_l}^2 = \sum_{j=1}^{p}\alpha_{lj}\left(\mu_{blj}^2 + s_{blj}^2\right)$$

$$\sigma_{w_l}^2 = \frac{\mathrm{Tr}\left(\boldsymbol{\Sigma}_{wl} + \boldsymbol{\mu}_{wl}\boldsymbol{\mu}_{wl}^T\right)}{m}$$

Update the residual variance parameter σ^2 :

$$\hat{\sigma}^2 = \frac{1}{n}\mathbb{E}_q\left[\left\|y - \sum_{l=1}^{L}\mathbf{X}\mathbf{b}_l\right\|^2\right]$$

Funmap with GWAS summary data

Consider the z-scores obtained from the univariate linear regression:

$$z_j = \frac{\hat{\beta}_j}{\hat{s}_j}, \quad \hat{\beta}_j = \left(\mathbf{x}_j^T \mathbf{x}_j\right)^{-1} \mathbf{x}_j^T \mathbf{y}, \quad \hat{s}_j = \sqrt{\frac{\| \mathbf{y} - \mathbf{x}_j \hat{\beta}_j \|_2^2}{n \left(\mathbf{x}_j^T \mathbf{x}_j\right)}}$$

Note that the likelihood and its lower bound depend on the GWAS data $\{\mathbf{X}, \mathbf{y}\}$

only through the sufficient statistics: $\mathbf{X}^T\mathbf{X}, \mathbf{X}^T\mathbf{y}$ and $\mathbf{y}^T\mathbf{y}$.

We can replace the sufficient statistics with z-scores and LD matrix with the following relationships:

$$\mathbf{X}^T\mathbf{X} = n\,\mathbf{R}, \quad \mathbf{x}_j^T\mathbf{y} = \frac{n}{\sqrt{n + z_j^2}}\, z_j, \quad \mathbf{y}^T\mathbf{y} = n.$$

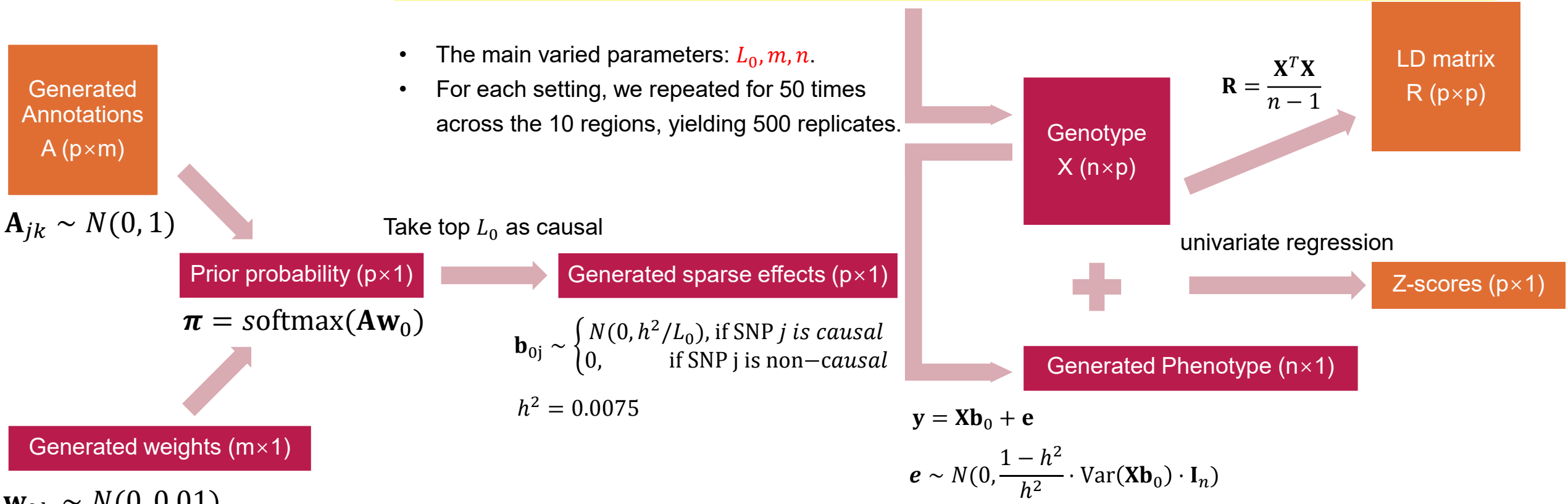
where $\mathbf{R}$ can be computed with genotypes from a subset of GWAS samples or

from a reference panel of similar ancestry background.

Simulation Settings

Genotype is from 10 risk regions identified in GWAS as associated with breast cancer (Fachal et al. 2020).
The total sample size is $n = 50{,}000$ and each region comprises $p = 700 \sim 4200$ variants.

- The main varied parameters: L_0, m, n.
- For each setting, we repeated for 50 times across the 10 regions, yielding 500 replicates.



- Squares represent matrices and rectangle represent vectors.
- The data in  is the simulation inputs of Funmap.

Comparisons of FDR calibration

01

To control the global FDR, we first compute the local FDR of each SNP as:

$$\mathrm{fdr}_j = 1 - \mathrm{PIP}_j$$

02

Then we sort SNPs by local FDR in ascending order and regard the j-th re-ordered SNP as a putative causal SNP if:

$$\mathrm{FDR}_{(j)} = \frac{\sum_{i=1}^{j} fdr_{(i)}}{j} < \xi$$

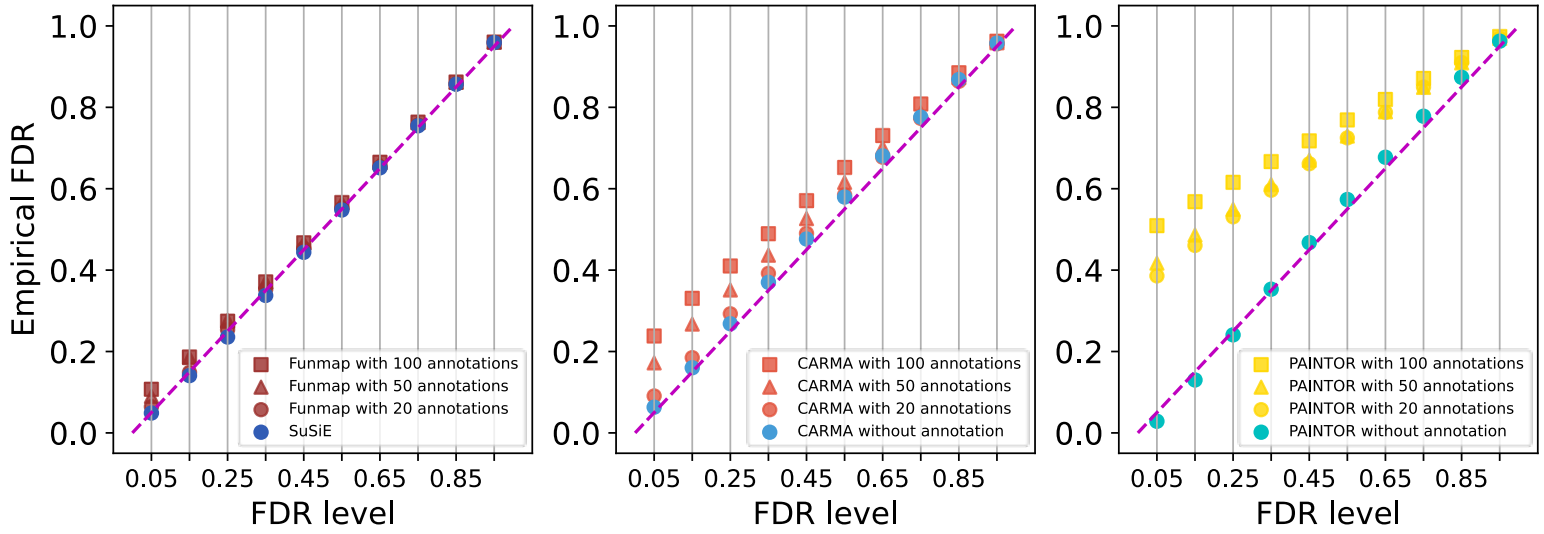
where $\mathrm{fdr}_{(i)}$ is the i-th ordered local FDR, $\mathrm{FDR}_{(j)}$ is the corresponding global FDR, and ξ is the selected FDR level to control the global FDR.

03

We aggregated 500 simulation replicates to improve the precision of empirical FDR and identified putative causal SNPs with a given FDR level and computed the empirical FDR as:

$$\text{empirical FDR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP}}$$

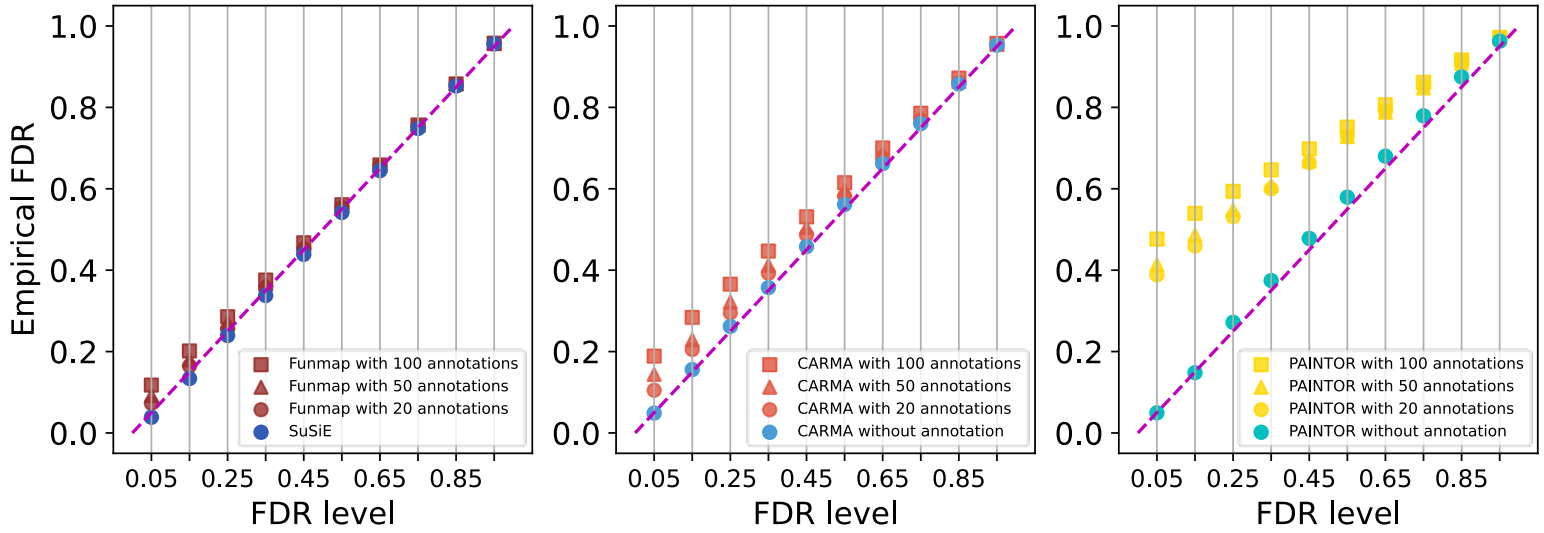
Comparisons of FDR calibration



Sample size: n = 50,000

Number of all SNPs: p ∈ (700, 4200)

Number of causal SNPs: L0 = 2

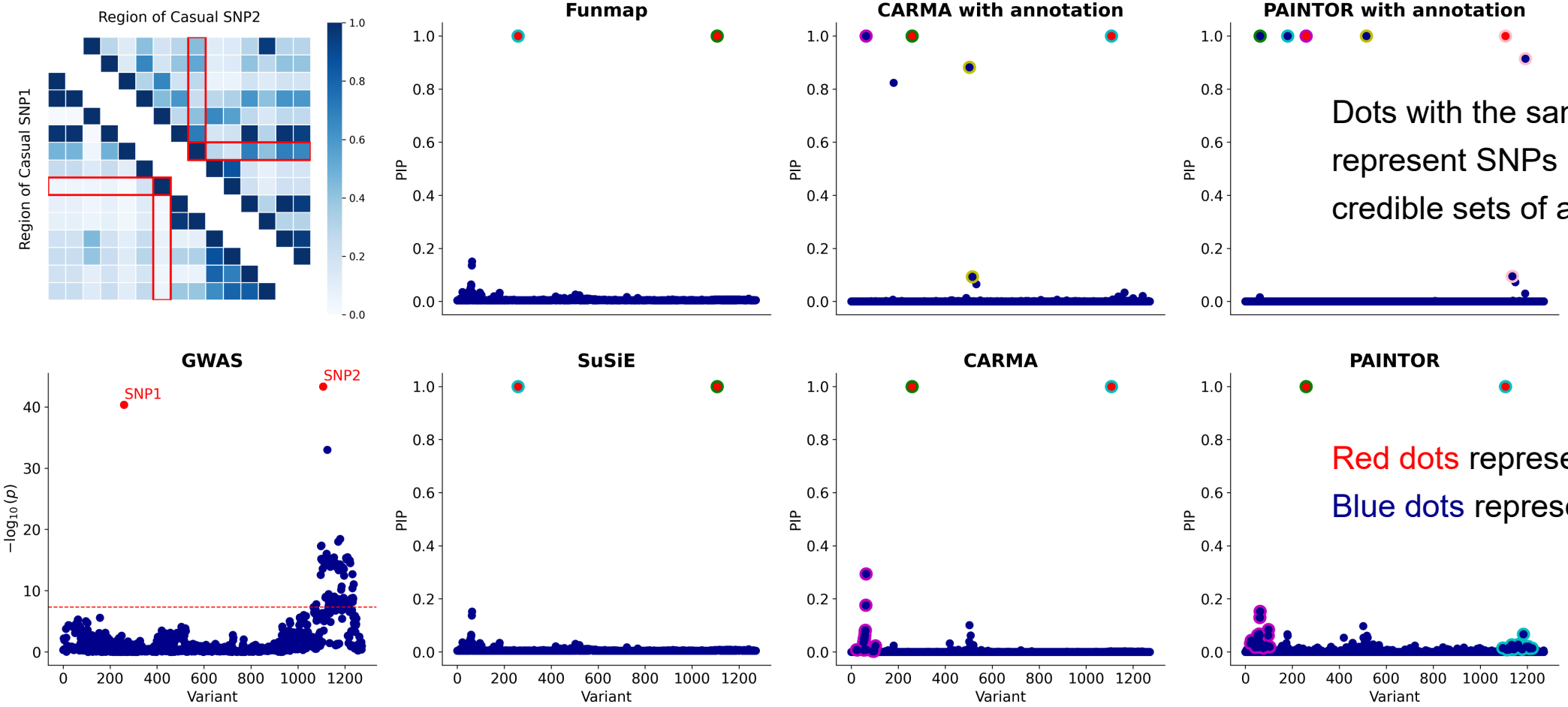


Sample size: n = 50,000

Number of all SNPs: p ∈ (700, 4200)

Number of causal SNPs: L0 = 3

An illustrative example generated by simulation

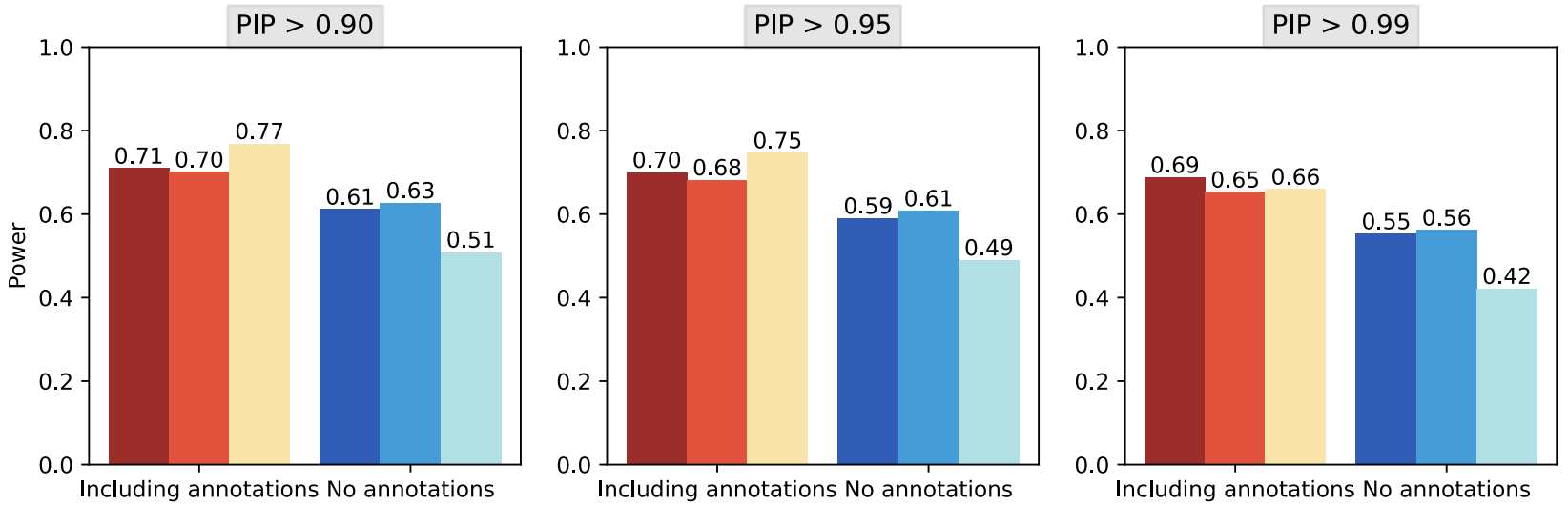


Dots with the same outline color represent SNPs in the level-95% credible sets of a causal signal.

Red dots represent causal SNPs.

Blue dots represent non-causal SNPs.

Comparisons of power at the selected thresholds

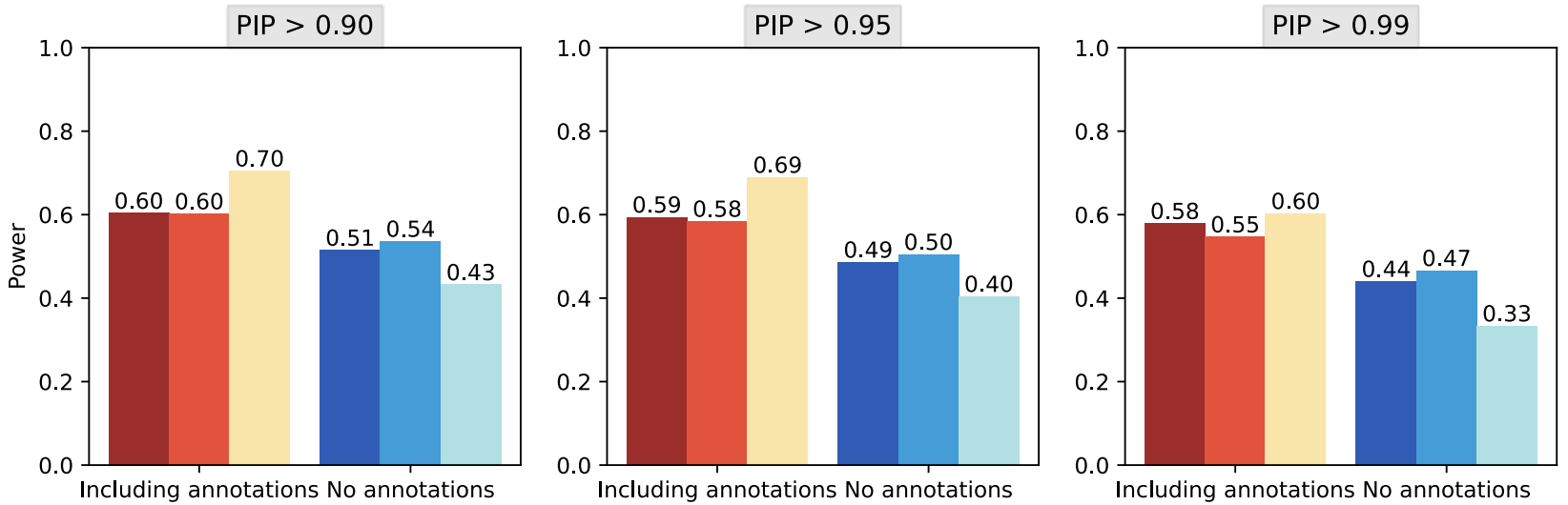


Sample size: n = 50,000

Number of annotations: m = 100

Number of all SNPs: p ∈ (700, 4200)

Number of causal SNPs: L0 = 2



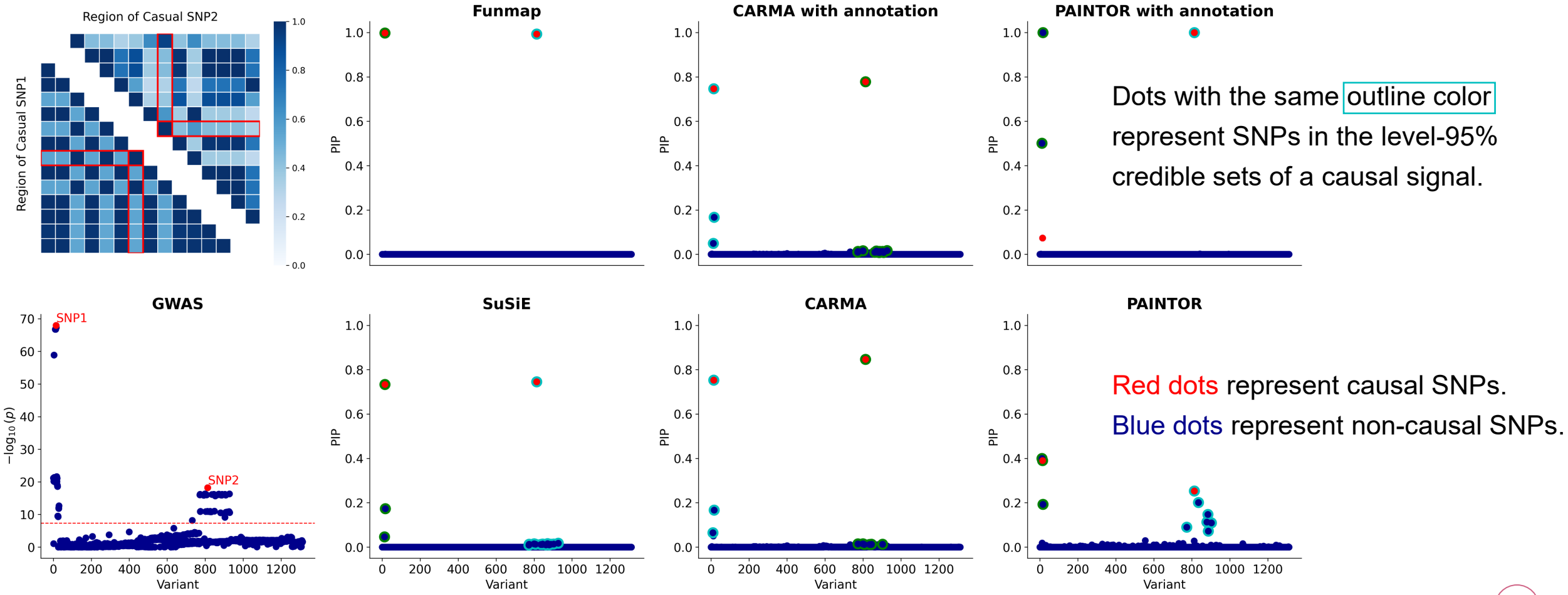
Sample size: n = 50,000

Number of annotations: m = 100

Number of all SNPs: p ∈ (700, 4200)

Number of causal SNPs: L0 = 3

Another illustrative example generated by simulation



Dots with the same outline color represent SNPs in the level-95% credible sets of a causal signal.

Red dots represent causal SNPs.

Blue dots represent non-causal SNPs.

Comparison of models' computational efficiency

Table 1: Average time consumption of different methods under a fixed number of annotations $m = 100$ (in seconds)

Methods	$m = 100$				
	$p = 854$	$p = 1313$	$p = 1833$	$p = 2561$	$p = 4107$
Funmap	37.24	71.90	98.30	224.42	346.89
CARMA+anno	343.65	797.59	978.62	983.73	2099.52
PAINTOR+anno	84.66	95.71	195.52	240.73	304.19
SuSiE	0.39	1.27	1.53	3.33	5.87
CARMA	383.06	703.22	1029.87	991.66	1298.16
PAINTOR	67.37	88.60	124.94	169.45	263.08

Table 2: Average time consumption of different methods with fixed number of variables $p = 1833$ (in seconds)

Methods	$p = 1833$				
	$m = 10$	$m = 50$	$m = 100$	$m = 150$	$m = 200$
Funmap	98.30	85.01	98.30	217.69	317.33
CARMA+anno	978.62	1191.52	978.62	953.04	644.12
PAINTOR+anno	195.52	149.77	195.52	99.90	121.45

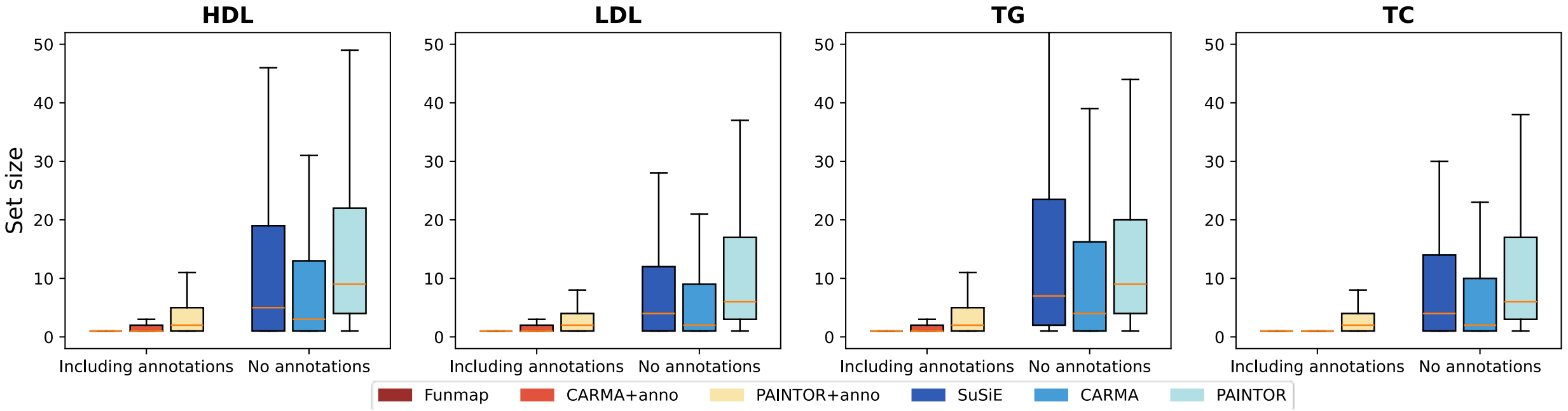
Introduction of real data analysis

- **GWAS summary data**: four lipid related traits, HDL, LDL, TC, TG. (from about 315,133 UKBB individuals of European ancestry, 864 genomic regions (190–347 per trait) with 434–8646 SNPs)

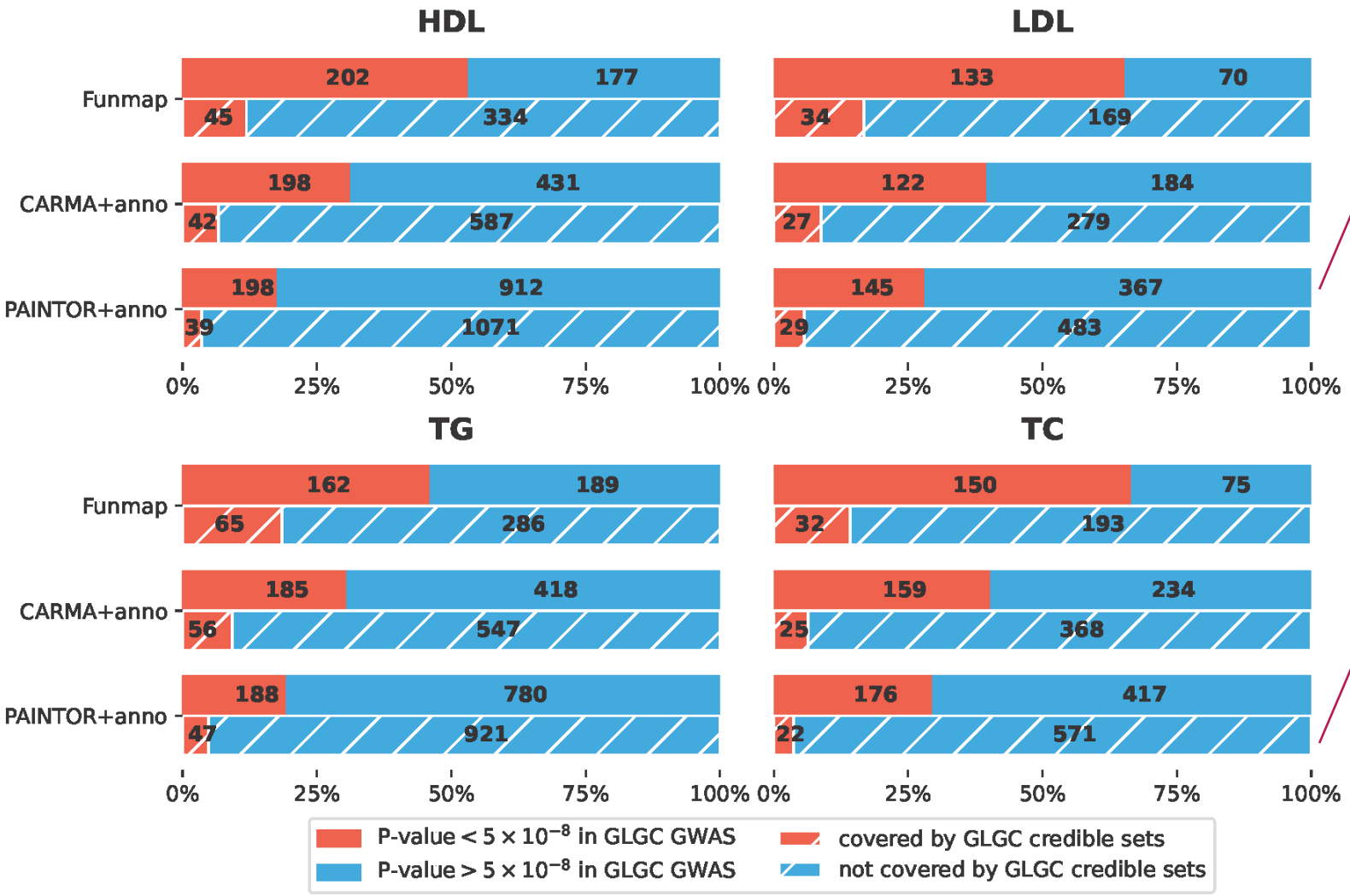
- **LD reference**: LD matrices of UK Biobank participants of a British ancestry, based on imputed genotypes prepared by Prof. Alkes Price‘s group. (from 337,199 UKBB individuals of British ancestry)

- **Functional annotations**: Baseline-LF v2.2 UKB annotations (187 annotations)

Box plots of credible set size across four lipid traits



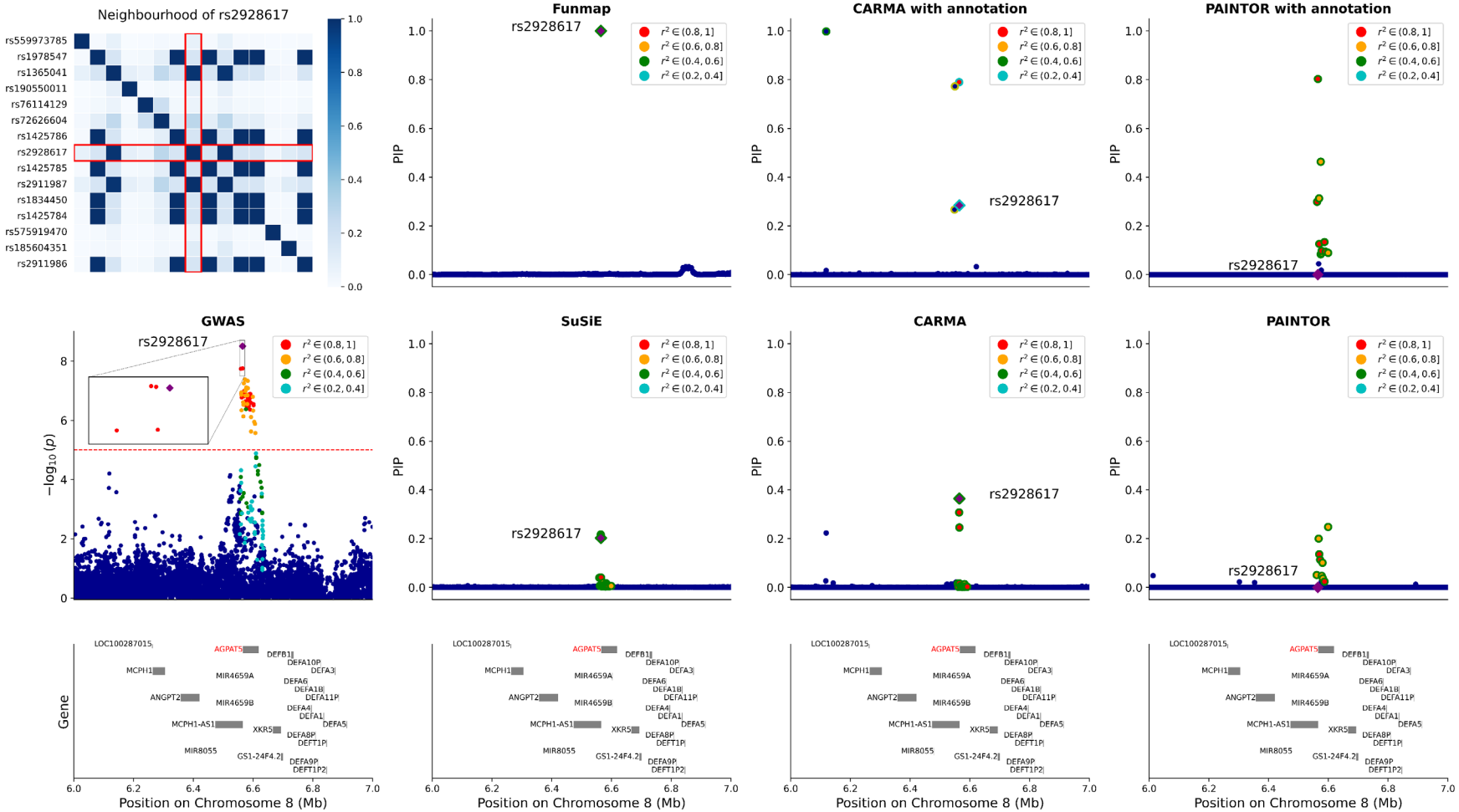
Replication analysis



Bar charts on the top shows the fraction and number of newly identified SNPs with P-value < 5 × 10−8 in the replication cohorts of GLGC GWAS.

Bar charts on the bottom shows the fraction and number of newly identified SNPs that are included in the 95%-level credible sets generated from GLGC GWAS with SuSiE.

Comparison of fine-mapping results from a region of cholesterol GWAS.



Colors of the points represent the correlation between neighboring SNPs and rs2928617.

Dots with the same outline color represent SNPs in the level-95% credible sets of a causal signal.

rs2928617 is represented by purple square ◆

locates on the 2Kb upstream of

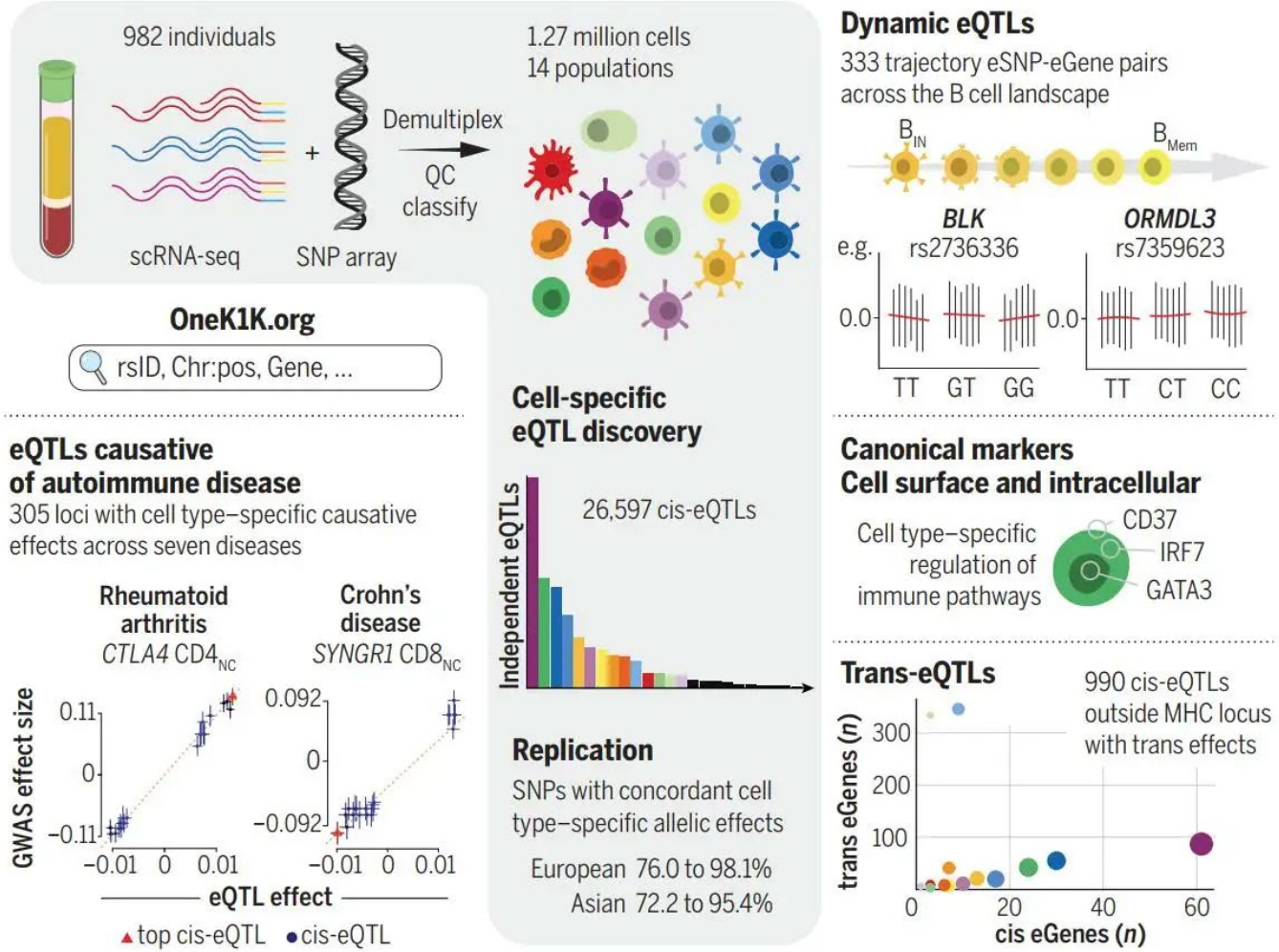
AGPAT5 encodes an integral membrane protein of the 1-acylglycerol-3-phosphate O-acyltransferase family.

香港城市大學
City University of Hong Kong

03

scSuSiE: eQTL mapping with single-cell data

OneK1K cohort



https://www.science.org/doi/10.1126/science.abf3041

The scSuSiE model

We extend the SuSiE framework to handle single-cell gene expression data for eQTL fine-mapping:

$$
\begin{aligned}
y_{ik} \mid \mu_{ik} &\sim \text{Poisson}(\mu_{ik}) \\
\mu_{ik} &= \exp(\eta_{ik}) \\
\eta_{ik} &= \mathbf{x}_{ik}^{c\,T}\boldsymbol{\alpha}^{c} + \mathbf{x}_{i}^{dT}\boldsymbol{\alpha}^{d} + \sum_{l=1}^{L} \tilde{\mathbf{g}}_{i}^{T}\mathbf{b}_{l} + c_{i}
\end{aligned}
$$

where the i-th sample from n individuals contributes m_i cells, resulting in a total of $M = \sum_{i=1}^{n} m_i$.

- $y_{ik} \in N_0$ denotes the UMI count for a target gene in cell k of individual i.
- $\mathrm{x}_{\mathrm{ik}}^{c} \in R^{p_c+1}$ represents cell-level covariates (gene expression PCs, mitochondrial content, etc.).
- $\mathrm{x}_{\mathrm{i}}^{d} \in R^{p_d}$ contains individual-level covariates (geno-type PCs, age, sex, etc.).
- $\tilde{\mathrm{g}}_{\mathrm{i}}^{T} \in R^{p}$ is the standardized genotype vector for individual i at p variants in *cis*.
- $\mathbf{b}_l \in R^{p}$ represents the random effect vector for the l-th single-effect component.
- c_i represents individual-specific random effects with $c_i \sim N(0, \tau)$.

The scSuSiE model

We extend the SuSiE framework to handle single-cell gene expression data for eQTL fine-mapping:

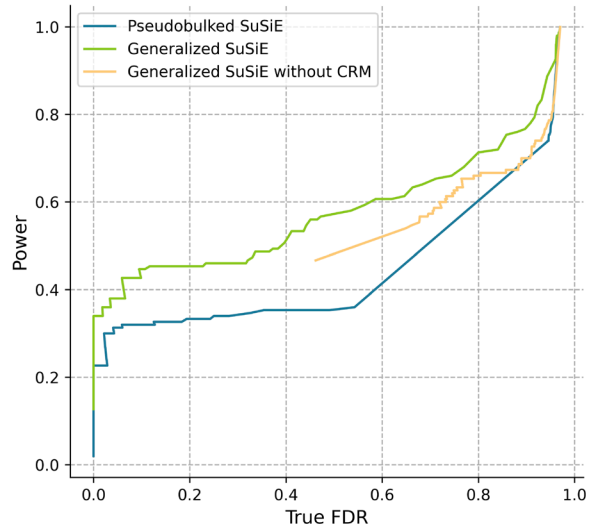
$$
\begin{aligned}
y_{ik} \mid \mu_{ik} &\sim \text{Poisson}(\mu_{ik}) \\
\mu_{ik} &= \exp(\eta_{ik}) \\
\eta_{ik} &= \mathbf{x}_{ik}^{c\,T}\boldsymbol{\alpha}^{c} + \mathbf{x}_{i}^{dT}\boldsymbol{\alpha}^{d} + \sum_{l=1}^{L} \tilde{\mathbf{g}}_{i}^{T}\mathbf{b}_{l} + c_{i}
\end{aligned}
$$

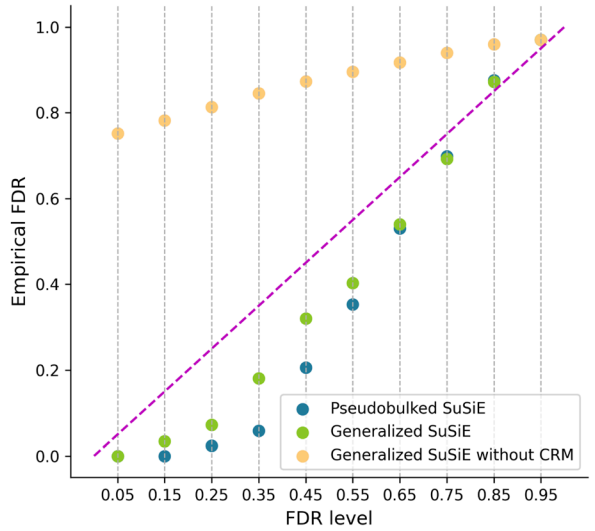
Following the SuSiE framework, we impose the sparse multiple-effects prior:

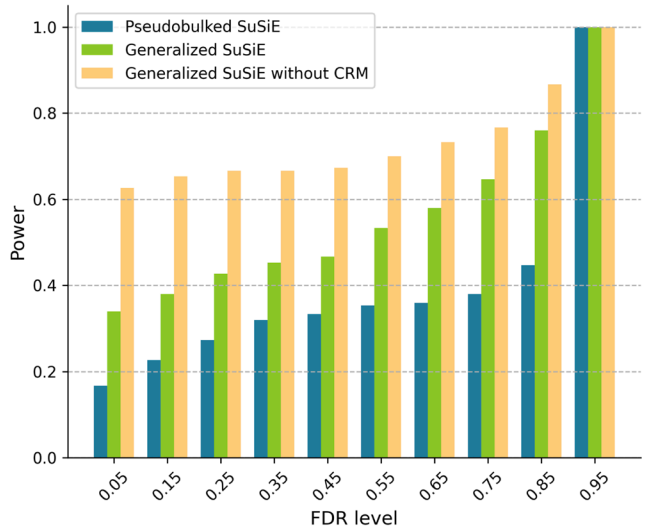
$$
\mathbf{b}_l = b_l \cdot \boldsymbol{\gamma}_l, \quad b_l \sim \mathcal{N}(0, \sigma_l^2), \quad \boldsymbol{\gamma}_l \sim \text{Multinomial}(1, \boldsymbol{\pi}), \quad l = 1, \ldots, L
$$

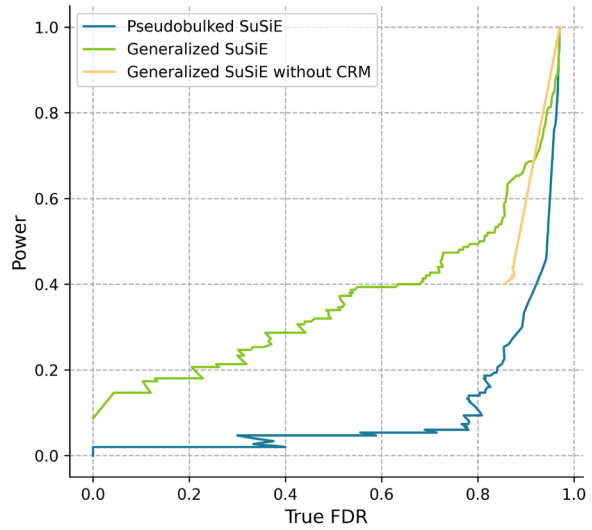
where $\boldsymbol{\gamma}_l \in \{0,1\}^p$ is a binary vector and we set $\boldsymbol{\pi} = \left[\frac{1}{p}, .., \frac{1}{p}\right]^T$ for computational convenience.

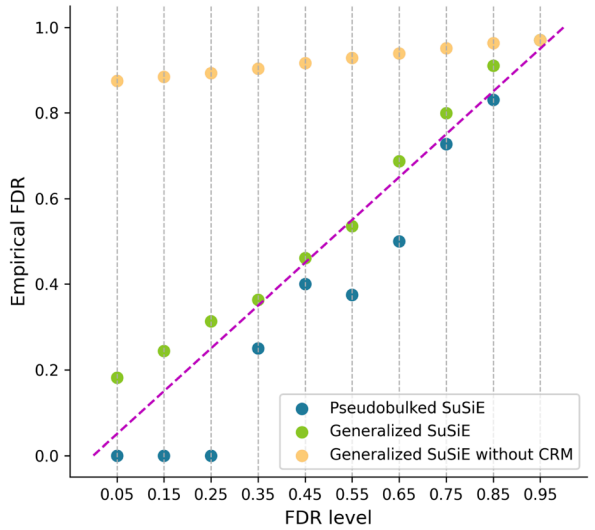
Preliminary simulation results

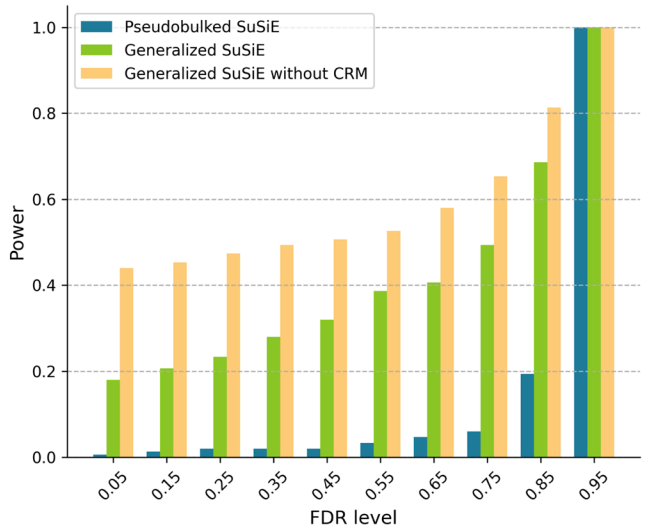












香港城市大學
City University of Hong Kong

04

Discussion

Discussion

Methodological Innovation

Funmap introduces a component-specific random effects model with adaptive shrinkage that effectively controls false discovery rates in high-dimensional functional annotation integration.

scSuSiE addresses single-cell eQTL analysis limitations through proper modelling of cell-to-cell correlations via the cell-cell relatedness matrix.

Superior Performance

Funmap demonstrated substantially higher replication rates in lipid traits analysis while maintaining proper FDR calibration.

scSuSiE achieved optimal balance on single-cell mapping tasks between enhanced statistical power from individual cell analysis and proper false positive control.

Publications

1. **Yuekai Li**, Jiashun Xiao, Jingsi Ming, Yicheng Zeng*, Mingxuan Cai*,“Funmap: integrating high-dimensional functional annotations to improve fine-mapping”, *Bioinformatics*, 2025.

香港城市大學
City University of Hong Kong

THANK YOU

Speaker: LI Yuekai

Supervisor: CAI Mingxuan