B-LORE

Bayesian multiple logistic regression for case-control GWAS

Saikat Banerjee[1], Lingyao Zeng[2], Heribert Schunkert[2] and Johannes Söding[1]

[1] Max Planck Institute for Biophysical Chemistry, 37077 Göttingen, Germany
[2] German Heart Centre, 80636 Munich, Germany

saikat.banerjee@mpibpc.mpg.de, soeding@mpibpc.mpg.de

1 Motivation

In genome-wide association studies (GWAS), genetic variants are tested for disease association mostly using **simple regression**, one variant at a time. This is straightforward, fast and easy to interpret but ignores the complexity of the data. Improvement in GWAS methods have explored several directions:

Meta-Analysis improve power by combining summary statistics from many studies. It is only used with simple regression.

Multiple regression aggregate evidence from multiple nearby variants. It can distinguish disease-coupled variants from those which are merely correlated with a coupled variant. It requires full genotype data. Multiple logistic regression use inefficient sampling schemes.

Functional Genomics data from other sources improve finemapping *i.e.* pinpoint causal SNPs. Finemapping for a meta study use approximations for LD structure of each population.

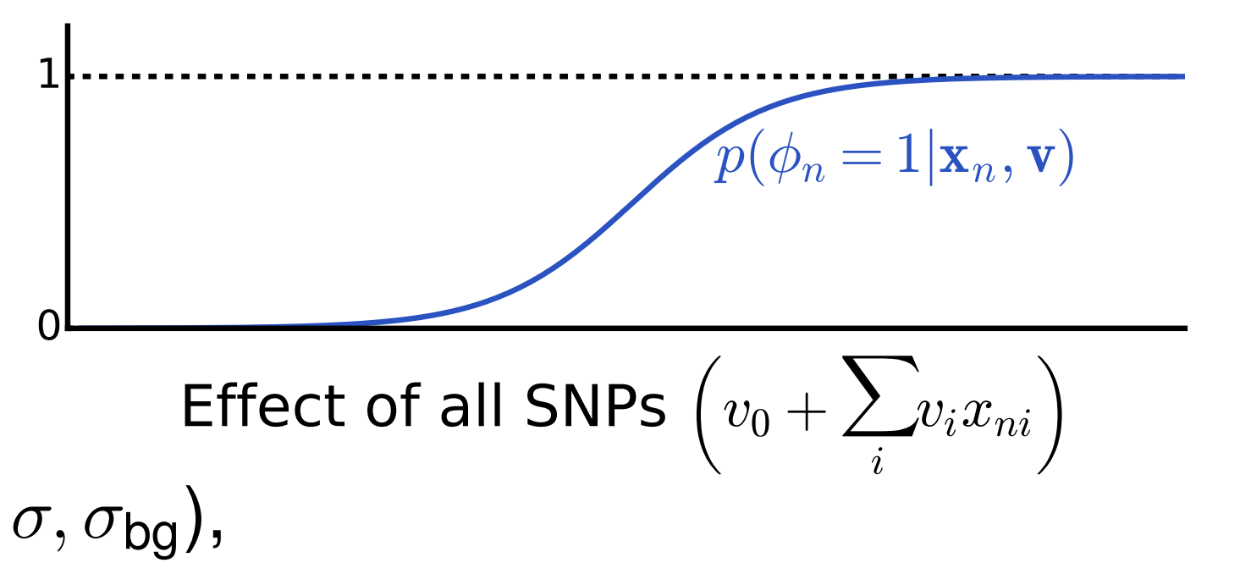
Can we use multiple logistic regression in a meta-analysis of case control GWAS and prioritize variants with functional genomics data?

We attempt to solve this in B-LORE, which uses a novel **quasi-Laplace approximation** to analytically integrate over variant effect sizes.

2 Model and Priors

Probability of n^{th} individual with genotype $\mathbf{x}_n$ to be diseased:

$$\boxed{p\left(\phi_n = 1 \mid \mathbf{x}_n, \mathbf{v}\right) = \frac{\exp\left(\mathbf{v}^\intercal \mathbf{x}_n\right)}{1 + \exp\left(\mathbf{v}^\intercal \mathbf{x}_n\right)}}$$



Prior on effect sizes given hyperparameters $\boldsymbol{\theta}$ $(\pi, \mu, \sigma, \sigma_{\text{bg}})$,

$$p\left(v_i \mid \boldsymbol{\theta}\right)$$

$$= \overbrace{\pi_i \mathcal{N}\left(v_i \mid \mu, \sigma^2\right)}^{\text{Causal}} + \overbrace{\left(1 - \pi_i\right) \mathcal{N}\left(v_i \mid 0, \sigma^2_{\text{bg}}\right)}^{\text{Non-causal}}$$

$$= \sum_{z_i = 0,1} \pi_i^{z_i} \left(1 - \pi_i\right)^{(1 - z_i)} \times \mathcal{N}\left(v_i \mid \boldsymbol{\mu}_{\mathbf{z},i}, \operatorname{diag}(\boldsymbol{\sigma}^2_{\mathbf{z},i})\right)$$

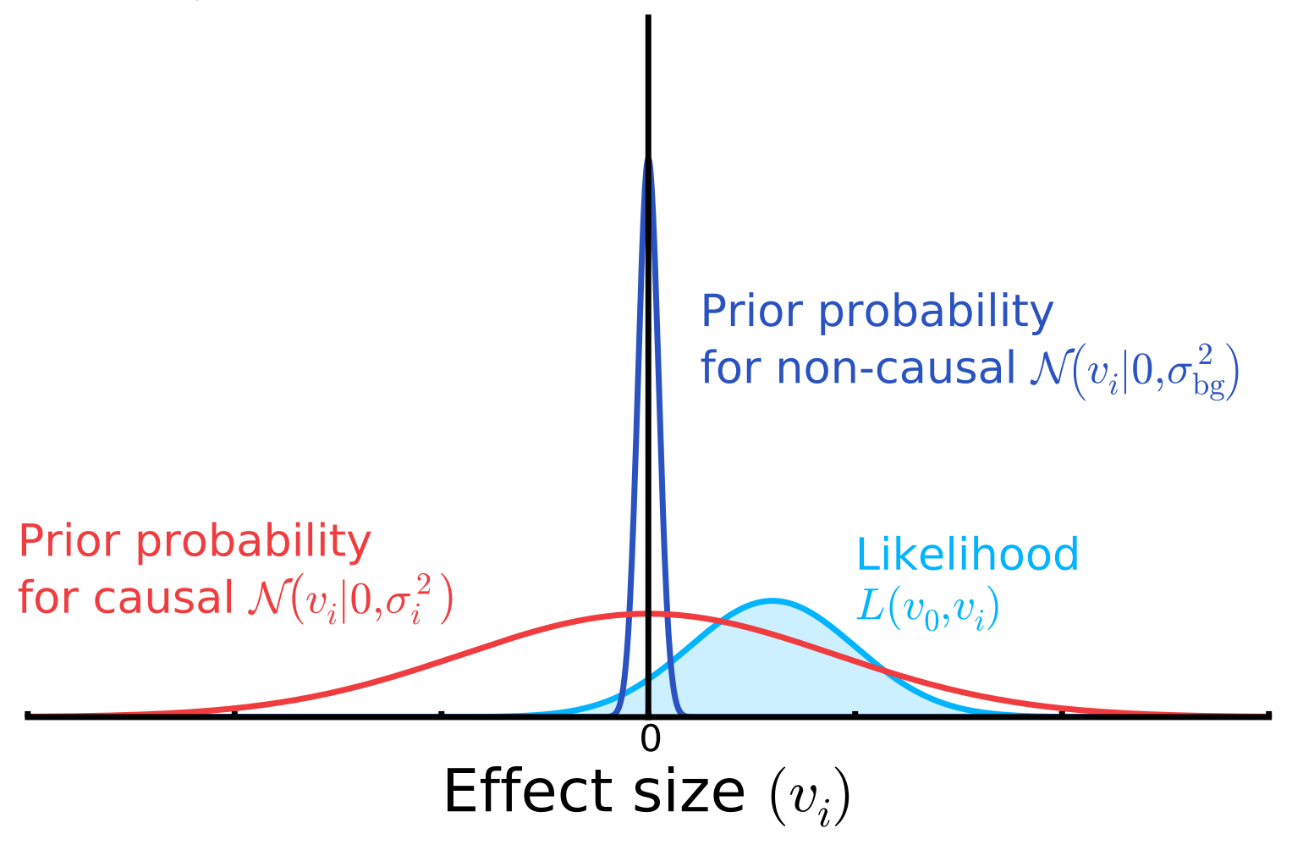
$$\mu_{\mathbf{z},i} = z_i \mu$$

$$\sigma^2_{\mathbf{z},i} = \sigma^2_{\text{bg}} + z_i \left(\sigma^2 - \sigma^2_{\text{bg}}\right)$$

$$\pi_i = \frac{1}{1 + \exp\left(-\boldsymbol{\xi}_i^\intercal \boldsymbol{\beta}_\pi\right)}$$

$z_i \in \{0, 1\} \Rightarrow$ Indicator variable of causality

$\boldsymbol{\xi}_i \Rightarrow$ vector of local genomic features



- $z_i = 1$ SNP i is causal
- $z_i = 0$ SNP i is non-causal

3 Optimization

Evidence approximation: maximizing the marginal likelihood

$$mL(\boldsymbol{\theta}) = p\left(\boldsymbol{\phi} \mid \mathbf{x}, \boldsymbol{\theta}\right) = \sum_{\mathbf{z}} p\left(\mathbf{z} \mid \boldsymbol{\theta}\right) \int p\left(\boldsymbol{\phi} \mid \mathbf{x}, \mathbf{v}\right) \mathcal{N}\left(\mathbf{v} \mid \boldsymbol{\mu}_{\mathbf{z}}, \operatorname{diag}\left(\boldsymbol{\sigma}^2_{\mathbf{z}}\right)\right) d\mathbf{v} \rightarrow \max$$

Quasi-Laplace approximation:

$$p\left(\boldsymbol{\phi} \mid \mathbf{x}, \mathbf{v}\right) \mathcal{N}\left(\mathbf{v} \mid \boldsymbol{\mu}_{\mathbf{z}}, \operatorname{diag}\left(\boldsymbol{\sigma}^2_{\mathbf{z}}\right)\right) = \underbrace{p\left(\boldsymbol{\phi} \mid \mathbf{x}, \mathbf{v}\right) \mathcal{N}\left(\mathbf{v} \mid \tilde{\boldsymbol{\mu}}, \operatorname{diag}\left(\tilde{\boldsymbol{\sigma}}^2\right)\right)}_{\propto \mathcal{N}\left(\mathbf{v} \mid \tilde{\mathbf{v}}, \tilde{\boldsymbol{\Lambda}}^{-1}\right)} \times \frac{\mathcal{N}\left(\mathbf{v} \mid \boldsymbol{\mu}_{\mathbf{z}}, \operatorname{diag}\left(\boldsymbol{\sigma}^2_{\mathbf{z}}\right)\right)}{\mathcal{N}\left(\mathbf{v} \mid \tilde{\boldsymbol{\mu}}, \operatorname{diag}\left(\tilde{\boldsymbol{\sigma}}^2\right)\right)}$$

The optimization can be done over multiple studies,

$$mL(\boldsymbol{\theta}) = p\left(\boldsymbol{\phi} \mid \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_S, \boldsymbol{\theta}\right) = \int p\left(\boldsymbol{\phi} \mid \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_S, \mathbf{v}\right) p\left(\mathbf{v} \mid \boldsymbol{\theta}\right) d\mathbf{v} \rightarrow \max$$

assuming that the quasi-Laplace approximation holds for each individual study

$$\prod_{s=1}^{S} \left[p\left(\boldsymbol{\phi} \mid \mathbf{x}_s, \mathbf{v}\right) \mathcal{N}\left(\mathbf{v} \mid \tilde{\boldsymbol{\mu}}_{\mathbf{z},s}, \operatorname{diag}(\tilde{\boldsymbol{\sigma}}^2_{\mathbf{z},s})\right)\right] \propto \prod_{s=1}^{S} \mathcal{N}\left(\mathbf{v} \mid \tilde{\mathbf{v}}_s, \tilde{\boldsymbol{\Lambda}}_s^{-1}\right) = \mathcal{N}\left(\mathbf{v} \mid \tilde{\mathbf{v}}, \tilde{\boldsymbol{\Lambda}}^{-1}\right)$$

where $\tilde{\boldsymbol{\Lambda}} = \sum_{s=1}^{S} \tilde{\boldsymbol{\Lambda}}_s$ and $\tilde{\mathbf{v}} = \tilde{\boldsymbol{\Lambda}}^{-1} \sum_{s=1}^{S} \tilde{\boldsymbol{\Lambda}}_s \tilde{\mathbf{v}}_s$.

B-LORE schema

1. Two optimizations at each cohort to estimate $(\tilde{\boldsymbol{\mu}}_s, \tilde{\boldsymbol{\sigma}}_s)$ and $(\tilde{\mathbf{v}}_s, \tilde{\boldsymbol{\Lambda}}_s)$.
2. Optimize summary statistics to estimate the hyperparameters.

4 Inference

Prediction of causality of each locus.

The probability for a locus to be causally associated with the disease is

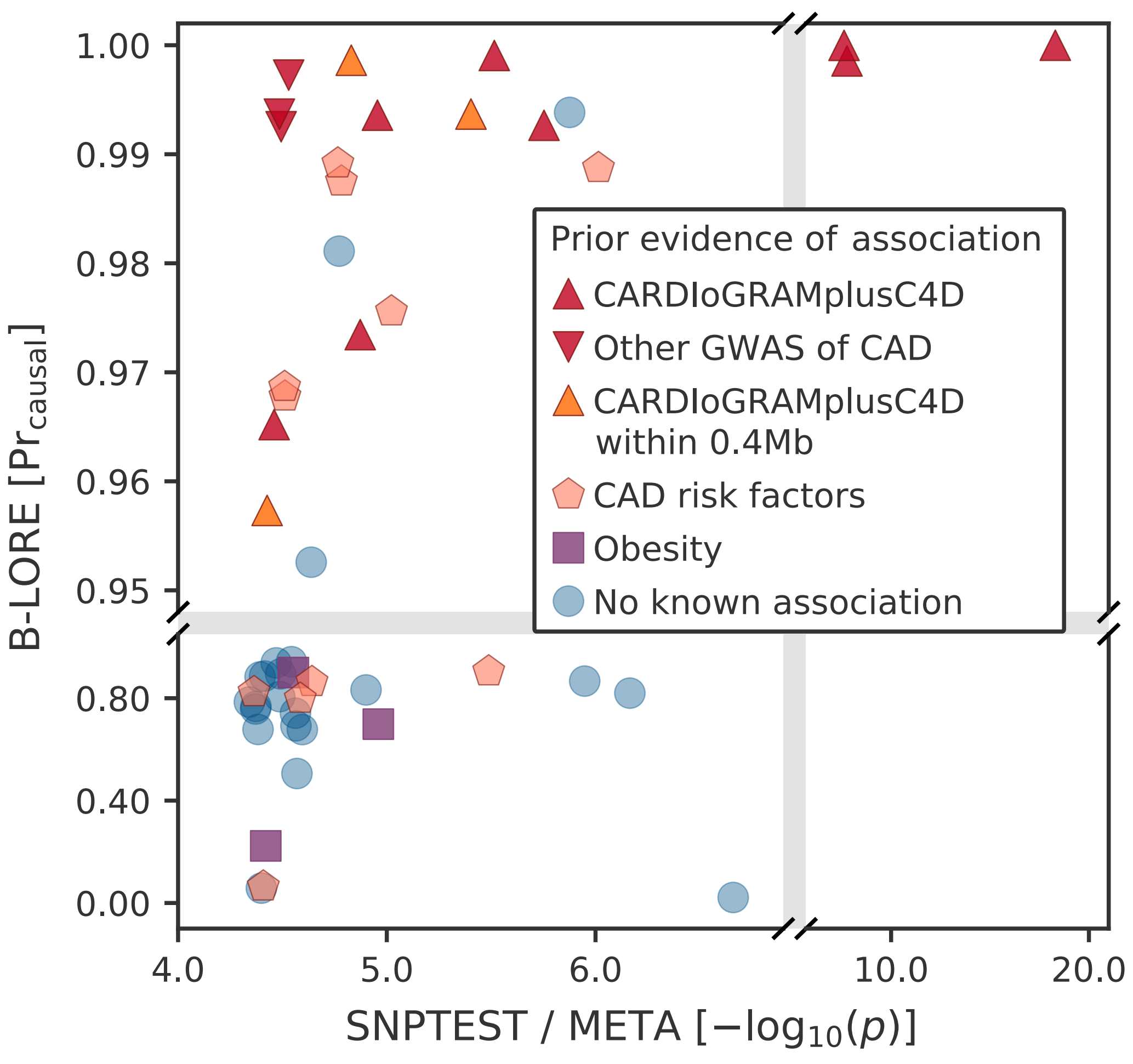
$$\boxed{\begin{aligned} \Pr_{\text{causal}} &= p\left(\text{locus is causal} \mid \boldsymbol{\phi}, \mathbf{X}, \hat{\boldsymbol{\theta}}\right) \\ &= 1 - p\left(\mathbf{z} = 0 \mid \boldsymbol{\phi}, \mathbf{X}, \hat{\boldsymbol{\theta}}\right) \end{aligned}}$$

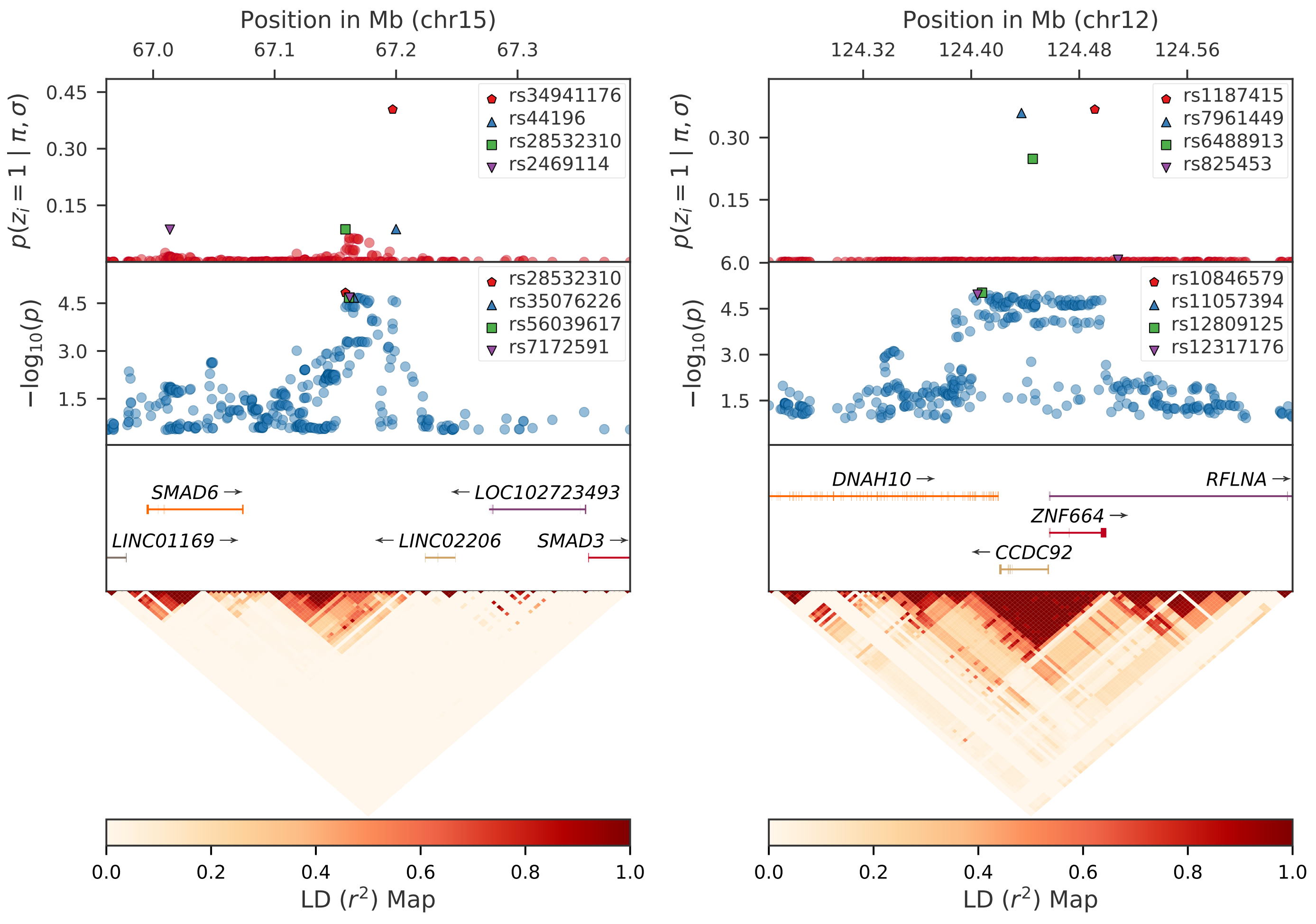
Statistical finemapping of causal variants.

The posterior probability for SNP i to be causal is

$$\boxed{p\left(z_i = 1 \mid \boldsymbol{\phi}, \mathbf{X}, \hat{\boldsymbol{\theta}}\right)}$$

5 Application on Coronary Artery Disease

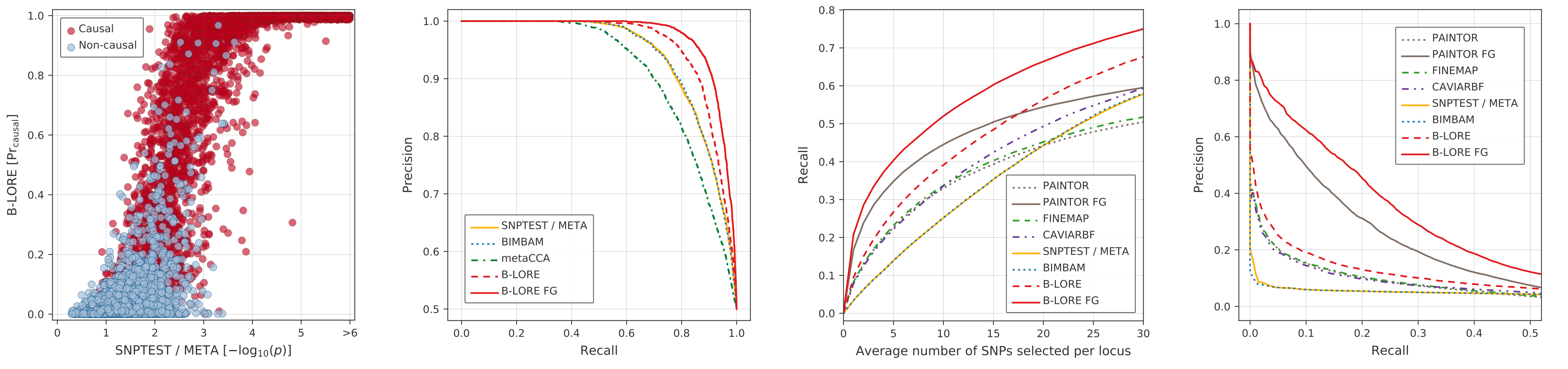




Meta-analysis of 5 cohorts (Germal Myocardial Infarction Family Study, GerMIFS I-V) with a total of 6234 cases and 6848 controls from white European ancestry. We pre-selected the top 50 loci with SNPTEST / META, and applied B-LORE, using 112 functional genomics features for each SNP from DNase-seq data of the ENCODE project.

6 Meta-analyses with real genotype and simulated phenotype

We selected 200 loci each with 200 SNPs from GerMIFS. We used 112 functional genomics features. We randomly selected 3 features as significant and simulated binary phenotype for $\sim$ 13000 patients. Each simulation had 100 causal loci and $\sim$ 450 causal SNPs with a total heritability of 0.25.



7 References

1. Banerjee *et al.* bioRxiv 2017, doi:10.1101/198911
2. Servin *et al.* PLOS Genet 2007, doi:10.1371/journal.pgen.0030114
3. Guan *et al.* Ann Appl Stat 2011, doi:10.1214/11-AOAS455
4. Schunkert *et al.* , Nat Genet 2011, doi:10.1038/ng.784
5. CARDIoGRAMplusC4D Nat Genet 2015, doi:10.1038/ng.3396

8 Acknowledgement

We thank Prof. Dr. Jeanette Erdmann and all members of the Söding lab for helpful suggestions and discussions. This work was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the e:Med research and funding concept (grant 01ZX1313A-2014).

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

e:Med
SYSTEMS MEDICINE