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

https://github.com/soedinglab/b-lore

1 Motivation

Genetic variants in genome-wide association studies (GWAS) are tested for disease association mostly using simple regression, one variant at a time. In post-GWAS analyses, such as finemapping, **MULTIPLE REGRESSION** use multiple SNPs with Bayesian variable selection, in which a sparsity-enforcing prior on effect sizes is used to avoid overtraining. The effect sizes are integrated out for posterior inference.

MULTIPLE LOGISTIC REGRESSION has not yielded clear improvements over the linear model for binary traits in case-control GWAS.

MCMC SAMPLING has proved to be costly and technically challenging to perform the integreation.

LINEAR APPROXIMATION of the logistic function is often used for case-control data.

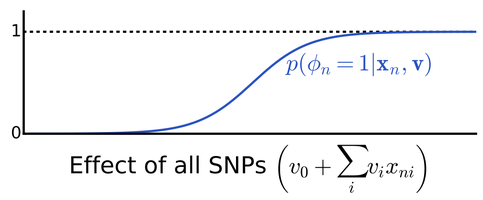
How to perform multiple logistic regression more accurately and faster?

In B-LORE, we introduce the **quasi-Laplace approximation** to analytically integrate over variant effect sizes. B-LORE improves finemapping with increasing number of controls.

2 Model and Priors

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

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



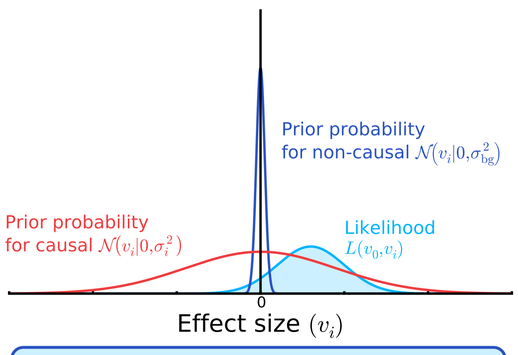
Prior on effect sizes given hyperparameters π and σ,

$p\left(v_i \mid \pi, \sigma\right)$

$$= \overbrace{\pi\,\mathcal{N}\left(v_i \mid 0, \sigma^2\right)}^{\text{Causal}} + \overbrace{(1-\pi)\,\delta_0}^{\text{Non-causal}}$$

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

$$= \sum_{z_i=0,1} p\left(\mathbf{z} \mid \pi\right)\mathcal{N}\left(v_i \mid \mathbf{0}, \mathrm{diag}(\boldsymbol{\sigma}^{\mathbf{2}}_{\mathbf{z,i}})\right)$$



where, $\sigma^2_{\mathbf{z},i} = z_i\sigma^2$

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

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

3 Optimization

Evidence approximation: maximizing the marginal likelihood

$$\mathrm{m}\mathcal{L}(\pi,\sigma) := p\left(\boldsymbol{\phi} \mid \mathbf{x}, \pi, \sigma\right) = \sum_{\mathbf{z}} p\left(\mathbf{z} \mid \pi\right) \int p\left(\boldsymbol{\phi} \mid \mathbf{x}, \mathbf{v}\right)\mathcal{N}\left(\mathbf{v} \mid \mathbf{0}, \mathrm{diag}\left(\boldsymbol{\sigma}^{\mathbf{2}}_{\mathbf{z}}\right)\right)d\mathbf{v} \to \max$$

Quasi-Laplace approximation:

$$p\left(\boldsymbol{\phi} \mid \mathbf{x}, \mathbf{v}\right)\mathcal{N}\left(\mathbf{v} \mid \mathbf{0}, \mathrm{diag}\left(\boldsymbol{\sigma}^{\mathbf{2}}_{\mathbf{z}}\right)\right) = \underbrace{p\left(\boldsymbol{\phi} \mid \mathbf{x}, \mathbf{v}\right)\mathcal{N}\left(\mathbf{v} \mid \mathbf{0},\ \tilde{\sigma}^{\mathbf{2}}\mathbb{I}\right)}_{\propto\ \mathcal{N}\left(\mathbf{v} \mid \tilde{\mathbf{v}}, \tilde{\boldsymbol{\Lambda}}^{-1}\right)} \frac{\mathcal{N}\left(\mathbf{v} \mid \mathbf{0}, \mathrm{diag}\left(\boldsymbol{\sigma}^{\mathbf{2}}_{\mathbf{z}}\right)\right)}{\mathcal{N}\left(\mathbf{v} \mid \mathbf{0},\ \tilde{\sigma}^{\mathbf{2}}\mathbb{I}\right)}$$

Benefits:

- The regularizer pulls the maximum of the regularized likelihood near to the mode of the integral, making it more accurate than Laplace approximation.
- Can be extended to multiple studies.
- Fast gradient-descent optimization.

B-LORE schema

1. Two-step optimization at each cohort to estimate $\tilde{\sigma}$ and $(\tilde{\mathbf{v}}, \tilde{\boldsymbol{\Lambda}})$.
2. Estimation of hyperparameters $(\pi,\ \sigma)$.

4 Inference

Prediction of causality of each locus.

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

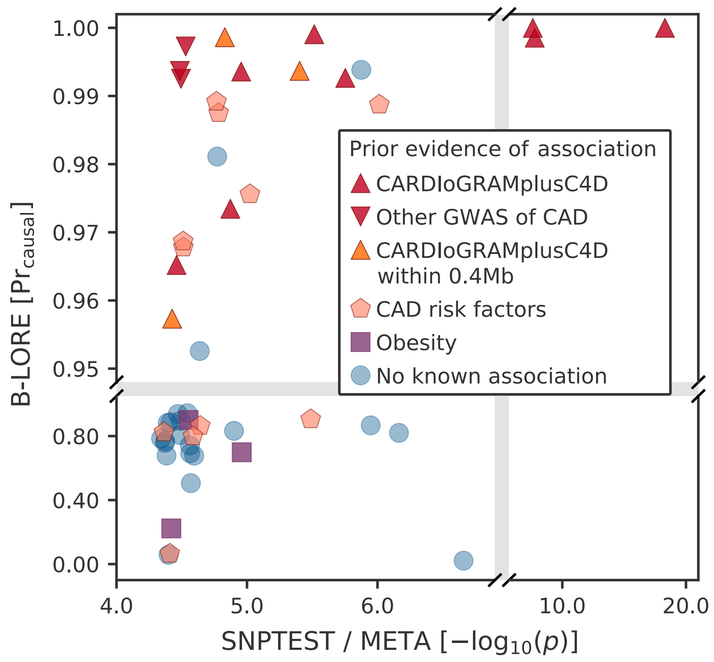
$$\begin{aligned}\mathrm{Pr}_{\text{causal}} &= p\left(\text{locus is causal} \mid \boldsymbol{\phi}, \mathbf{X}, \hat{\pi}, \hat{\sigma}\right)\\ &= 1 - p\left(\mathbf{z} = 0 \mid \boldsymbol{\phi}, \mathbf{X}, \hat{\pi}, \hat{\sigma}\right)\end{aligned}$$

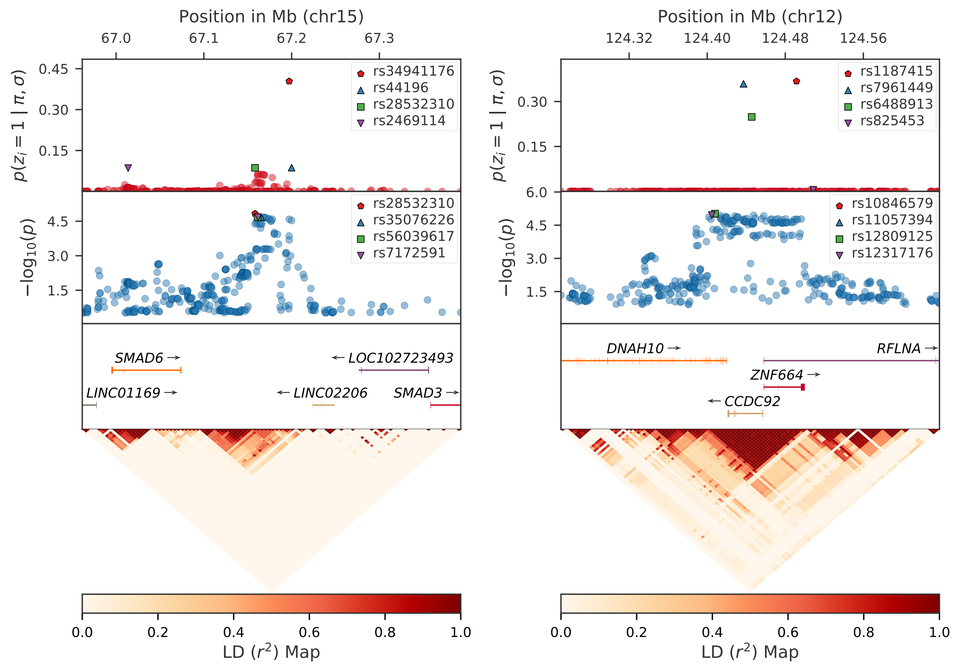
Statistical finemapping of causal variants.

The posterior probability for SNP i to be causal is

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

5 Application on Coronary Artery Disease

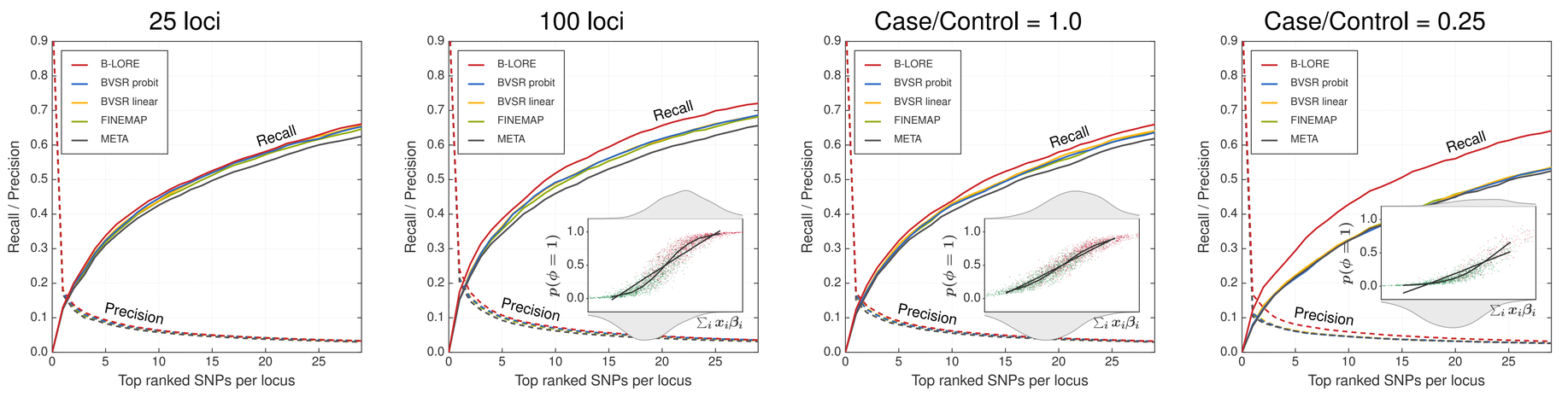




Meta-analysis of 5 cohorts (Germal Myocardial Infarction Family Study, GerMIFS I-V) with a total of 6234 cases and 6848 controls from white European ancestry. We pre-selected the top 50 loci with SNPTEST / META, and applied B-LORE.

6 Meta-analysis with real genotype and simulated phenotype

We simulated 13082 phenotypes for 5 cohorts usings 100 loci of ∼200 SNPs, using one or more causal SNPs in each locus.



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. 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