

Bayesian variable selection logistic regression: multivariate metaanalysis in GWAS

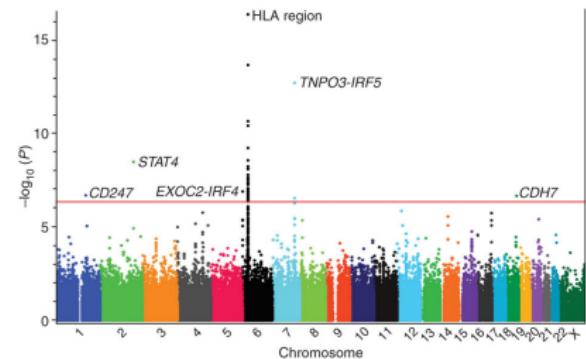
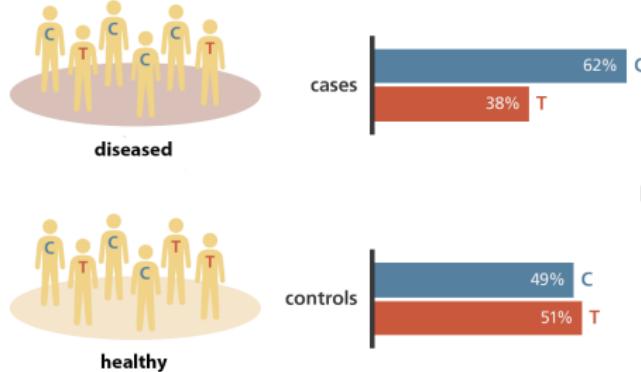
Saikat Banerjee

Max Planck Institute for Biophysical Chemistry

FEBRUARY 3, 2017



Genome-wide association studies (GWAS)



- ▶ Discovered thousands of variants associated with complex diseases

Association tests in GWAS

| # of samples | Genotype (x) | | | | | | Phenotype |
|--------------|--------------|----|----|----|----|----|-----------|
| | .. | GA | TT | AT | GC | AC | .. |
| .. | GG | TT | AT | GG | AA | .. | y_2 |
| .. | AA | TT | AA | GC | AA | .. | y_3 |
| .. | GA | TC | AT | GC | CC | .. | y_4 |
| .. | GG | TT | AT | CC | AC | .. | y_5 |
| .. | AA | TC | TT | CC | AC | .. | y_6 |

Association tests in GWAS

| # of samples | Genotype (x) | | | | | Phenotype | |
|--------------|--------------|----|----|----|----|-----------|-------|
| | GA | TT | AT | GC | AC | .. | y_1 |
| .. | GG | TT | AT | GG | AA | .. | y_2 |
| .. | AA | TT | AA | GC | AA | .. | y_3 |
| .. | GA | TC | AT | GC | CC | .. | y_4 |
| .. | GG | TT | AT | CC | AC | .. | y_5 |
| .. | AA | TC | TT | CC | AC | .. | y_6 |



Association tests in GWAS

| # of samples | Genotype (x) | | | | | Phenotype | |
|--------------|--------------|----|----|----|----|-----------|-------|
| | .. | GA | TT | AT | GC | AC | .. |
| .. | GG | TT | AT | GG | AA | .. | y_2 |
| .. | AA | TT | AA | GC | AA | .. | y_3 |
| .. | GA | TC | AT | GC | CC | .. | y_4 |
| .. | GG | TT | AT | CC | AC | .. | y_5 |
| .. | AA | TC | TT | CC | AC | .. | y_6 |



Association tests in GWAS

| | Genotype (x) | | | | | | Phenotype | |
|--------------|--------------|----|----|----|----|----|-----------|-------|
| # of samples | .. | GA | TT | AT | GC | AC | .. | y_1 |
| | .. | GG | TT | AT | GG | AA | .. | y_2 |
| | .. | AA | TT | AA | GC | AA | .. | y_3 |
| | .. | GA | TC | AT | GC | CC | .. | y_4 |
| | .. | GG | TT | AT | CC | AC | .. | y_5 |
| | .. | AA | TC | TT | CC | AC | .. | y_6 |



Association tests in GWAS

| # of samples | Genotype (x) | | | | | | Phenotype | |
|--------------|--------------|----|----|----|----|----|-----------|-------|
| | .. | GA | TT | AT | GC | AC | .. | y_1 |
| | .. | GG | TT | AT | GG | AA | .. | y_2 |
| | .. | AA | TT | AA | GC | AA | .. | y_3 |
| | .. | GA | TC | AT | GC | CC | .. | y_4 |
| | .. | GG | TT | AT | CC | AC | .. | y_5 |
| | .. | AA | TC | TT | CC | AC | .. | y_6 |



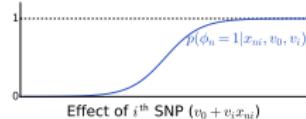
$$y_n = v_0 + v_i x_{ni} + \epsilon, \quad \text{with} \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \qquad \text{Quantitative phenotype}$$

Association tests in GWAS

| | Genotype (x) | | | | | | Phenotype | |
|--------------|--------------|----|----|----|----|----|-----------|-------|
| # of samples | .. | GA | TT | AT | GC | AC | .. | y_1 |
| | .. | GG | TT | AT | GG | AA | .. | y_2 |
| | .. | AA | TT | AA | GC | AA | .. | y_3 |
| | .. | GA | TC | AT | GC | CC | .. | y_4 |
| | .. | GG | TT | AT | CC | AC | .. | y_5 |
| | .. | AA | TC | TT | CC | AC | .. | y_6 |

$$y_n = v_0 + v_i x_{ni} + \epsilon, \quad \text{with} \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \quad \text{Quantitative phenotype}$$

$$p(y_n = 1 | x_{ni}, v_0, v_i) = \text{lf}(v_0 + v_i x_{ni}) \quad \text{Binary phenotype}$$



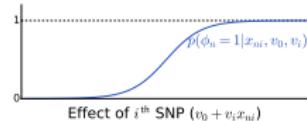
Association tests in GWAS

| # of samples | Genotype (x) | | | | | | Phenotype | |
|--------------|--------------|----|----|----|----|----|-----------|-------|
| | .. | GA | TT | AT | GC | AC | .. | y_1 |
| | .. | GG | TT | AT | GG | AA | .. | y_2 |
| | .. | AA | TT | AA | GC | AA | .. | y_3 |
| | .. | GA | TC | AT | GC | CC | .. | y_4 |
| | .. | GG | TT | AT | CC | AC | .. | y_5 |
| | .. | AA | TC | TT | CC | AC | .. | y_6 |

$$y_n = v_0 + v_i x_{ni} + \epsilon, \quad \text{with} \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \quad \text{Quantitative phenotype}$$

$$p(y_n = 1 | x_{ni}, v_0, v_i) = \text{lf}(v_0 + v_i x_{ni}) \quad \text{Binary phenotype}$$

- ▶ Is the coefficient v_i significantly different from 0? \Rightarrow P-values



Strengths

- Straightforward
- Computationally fast
- Conservative
- Easy to interpret

Strengths

- Straightforward
- Computationally fast
- Conservative
- Easy to interpret

Challenges

- Linkage disequilibrium
- Genetic networks
- Low effect sizes

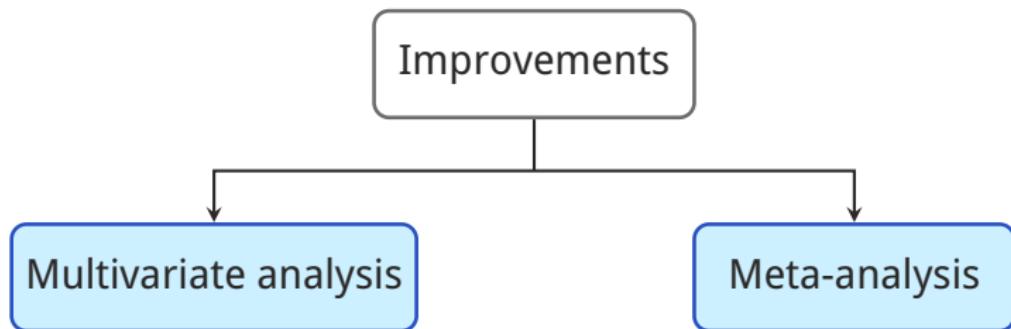
Univariate methods

Strengths

- Straightforward
- Computationally fast
- Conservative
- Easy to interpret

Challenges

- Linkage disequilibrium
- Genetic networks
- Low effect sizes

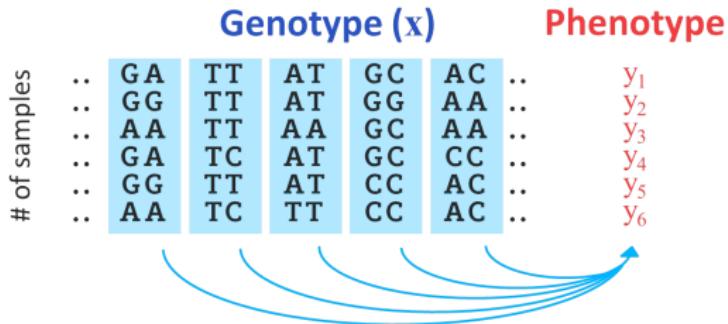


Multivariate methods

| # of samples | Genotype (x) | | | | | Phenotype | |
|--------------|--------------|----|----|----|----|-----------|-------|
| | .. | GA | TT | AT | GC | AC | .. |
| | .. | GG | TT | AT | GG | AA | y_1 |
| | .. | AA | TT | AA | GC | AA | y_2 |
| | .. | GA | TC | AT | GC | CC | y_3 |
| | .. | GG | TT | AT | CC | AC | y_4 |
| | .. | AA | TC | TT | CC | AC | y_5 |
| | | | | | | | y_6 |



Multivariate methods



Multivariate methods

| # of samples | Genotype (x) | | | | | Phenotype | | |
|--------------|--------------|----|----|----|----|-----------|----|-------|
| | .. | GA | TT | AT | GC | AC | .. | y_1 |
| | .. | GG | TT | AT | GG | AA | .. | y_2 |
| | .. | AA | TT | AA | GC | AA | .. | y_3 |
| | .. | GA | TC | AT | GC | CC | .. | y_4 |
| | .. | GG | TT | AT | CC | AC | .. | y_5 |
| | .. | AA | TC | TT | CC | AC | .. | y_6 |



The diagram illustrates the relationship between genotype and phenotype. It shows a grid of genotype data (x) with 6 rows and 6 columns. The first column is labeled '# of samples' and contains six dots. The subsequent five columns represent different loci or genes, with each cell containing a two-letter sequence (e.g., GA, TT, AT, GC, AC). To the right of the grid, the word 'Phenotype' is written in red, followed by six red labels: y_1 , y_2 , y_3 , y_4 , y_5 , and y_6 . Below the grid, six blue curved arrows originate from the bottom of each row and point upwards towards the corresponding y_i label, indicating a mapping or regression relationship.

$$y_n = v_0 + \sum_i v_i x_{ni} + \epsilon, \quad \text{with} \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \quad \text{Quantitative phenotype}$$

Multivariate methods

| | Genotype (x) | | | | | Phenotype | | |
|--------------|--------------|----|----|----|----|-----------|----|-------|
| # of samples | .. | GA | TT | AT | GC | AC | .. | y_1 |
| .. | GG | TT | AT | GG | AA | AA | .. | y_2 |
| .. | AA | TT | AA | GC | AA | AA | .. | y_3 |
| .. | GA | TC | AT | GC | CC | CC | .. | y_4 |
| .. | GG | TT | AT | CC | AC | AC | .. | y_5 |
| .. | AA | TC | TT | CC | AC | AC | .. | y_6 |



$$y_n = v_0 + \sum_i v_i x_{ni} + \epsilon, \quad \text{with} \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \quad \text{Quantitative phenotype}$$

$$p(y_n = 1 | x_{ni}, v_0, v_i) = \text{lf}\left(v_0 + \sum_i v_i x_{ni}\right) \quad \text{Binary phenotype}$$

Multivariate methods

| | Genotype (x) | | | | | Phenotype | | |
|--------------|--------------|----|----|----|----|-----------|----|-------|
| # of samples | .. | GA | TT | AT | GC | AC | .. | y_1 |
| .. | GG | TT | AT | GG | AA | AA | .. | y_2 |
| .. | AA | TT | AA | GC | AA | AA | .. | y_3 |
| .. | GA | TC | AT | GC | CC | CC | .. | y_4 |
| .. | GG | TT | AT | CC | AC | AC | .. | y_5 |
| .. | AA | TC | TT | CC | AC | AC | .. | y_6 |



$$y_n = v_0 + \sum_i v_i x_{ni} + \epsilon, \quad \text{with} \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \quad \text{Quantitative phenotype}$$

$$p(y_n = 1 | x_{ni}, v_0, v_i) = \text{lf}\left(v_0 + \sum_i v_i x_{ni}\right) \quad \text{Binary phenotype}$$

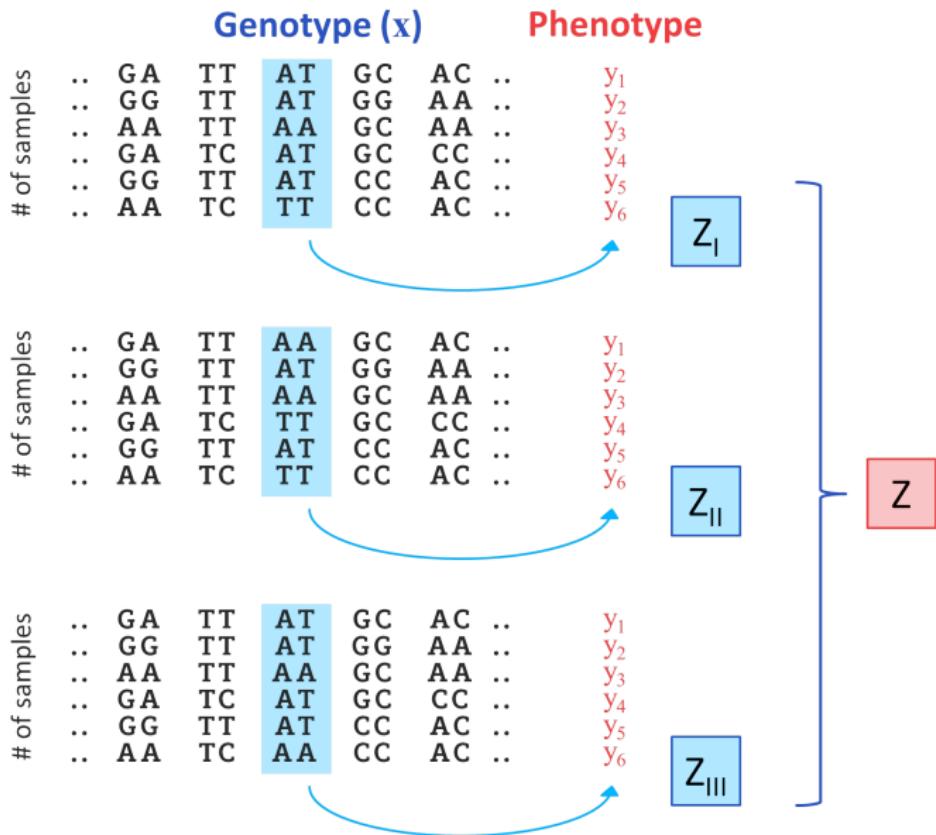
- ▶ Multivariate methods perform better than univariate methods

Meta-analysis

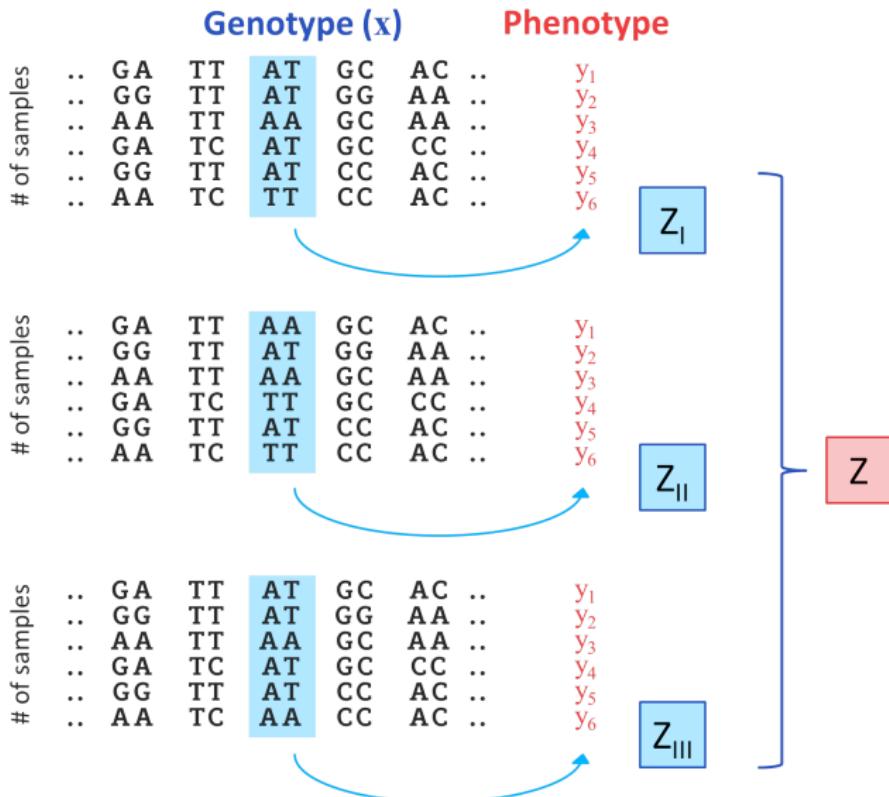
| # of samples | Genotype (x) | | | | | Phenotype | |
|--------------|--------------|----|----|----|----|-----------|-------|
| | .. | GA | TT | AT | GC | AC | .. |
| .. | GG | TT | AT | GG | AA | .. | y_2 |
| .. | AA | TT | AA | GC | AA | .. | y_3 |
| .. | GA | TC | AT | GC | CC | .. | y_4 |
| .. | GG | TT | AT | CC | AC | .. | y_5 |
| .. | AA | TC | TT | CC | AC | .. | y_6 |



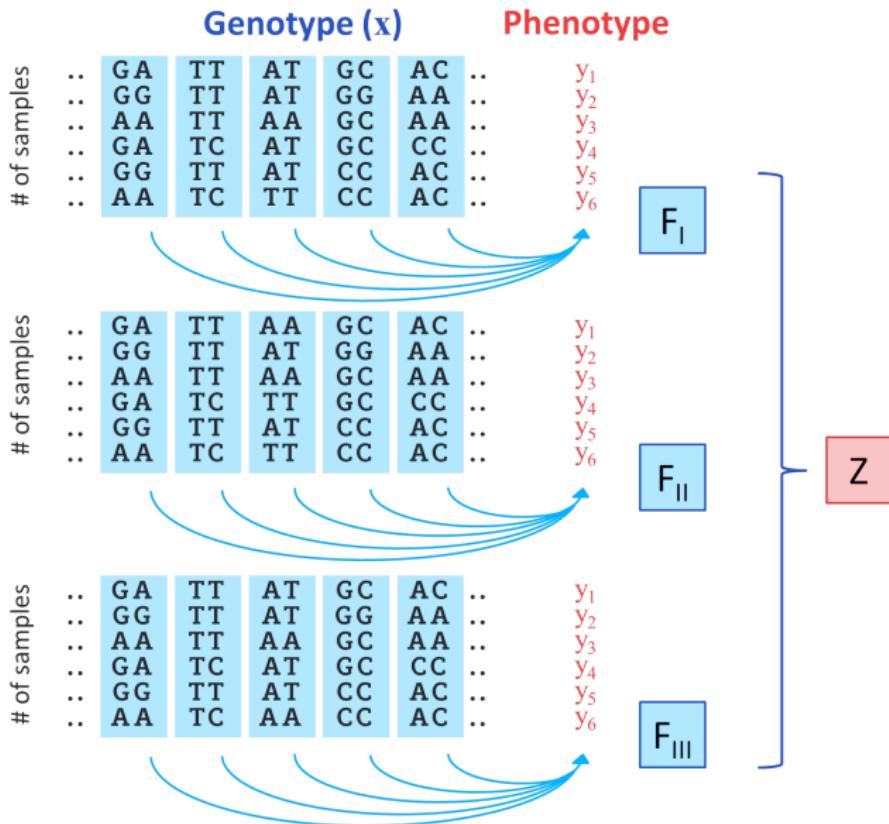
Meta-analysis



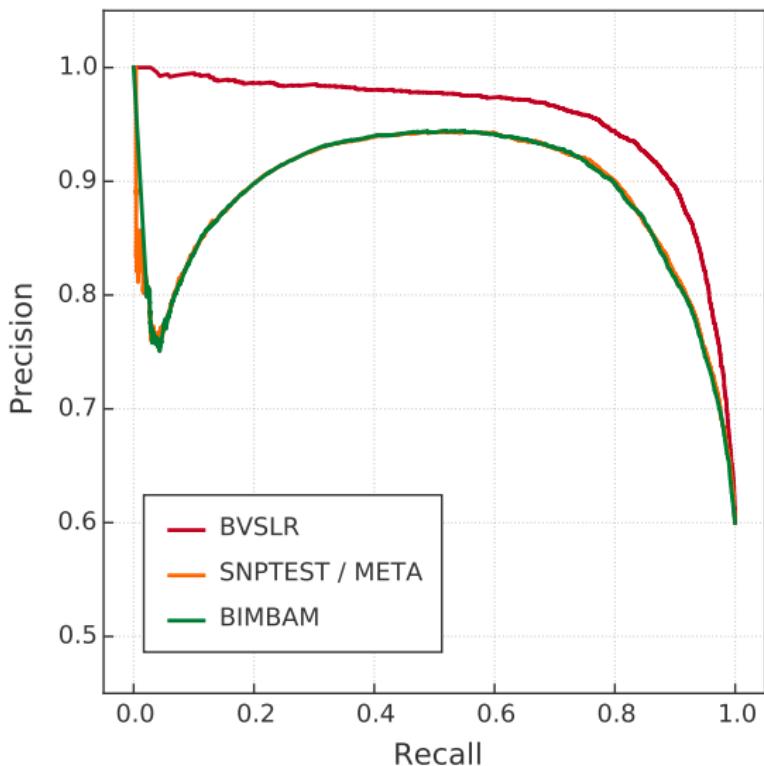
Goal of our method



Goal of our method



Preview of BVSLR



$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Bayesian variable selection regression (BVSR)

BIMBAM

Servin and Stephens, *PLoS Genetics* 2007
Guan and Stephens, *Ann. Appl. Stats.* 2011

$$y_n = v_0 + \sum_i v_i x_{ni} + \epsilon, \quad \text{with} \quad \epsilon \sim \mathcal{N}(0, \tau^{-1}) \quad \text{Quantitative phenotype}$$

Bayesian variable selection regression (BVSR)

BIMBAM

Servin and Stephens, *PLoS Genetics* 2007
Guan and Stephens, *Ann. Appl. Stats.* 2011

$$y_n = v_0 + \sum_i v_i x_{ni} + \epsilon, \quad \text{with} \quad \epsilon \sim \mathcal{N}(0, \tau^{-1}) \quad \text{Quantitative phenotype}$$

- ▶ Likelihood for N patients:

$$p(\mathbf{y} | \mathbf{x}, \mathbf{v}, \tau) = \mathcal{N}(\mathbf{y} | \mathbf{x}^\top \mathbf{v}, \tau^{-1} \mathbb{I})$$

Bayesian variable selection regression (BVSR)

BIMBAM

Servin and Stephens, *PLoS Genetics* 2007
Guan and Stephens, *Ann. Appl. Stats.* 2011

$$y_n = v_0 + \sum_i v_i x_{ni} + \epsilon, \quad \text{with} \quad \epsilon \sim \mathcal{N}(0, \tau^{-1}) \quad \text{Quantitative phenotype}$$

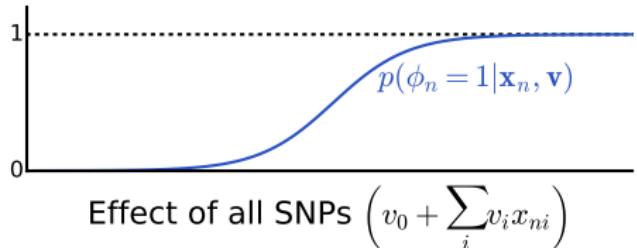
- ▶ Likelihood for N patients:

$$p(\mathbf{y} | \mathbf{x}, \mathbf{v}, \tau) = \mathcal{N}(\mathbf{y} | \mathbf{x}^T \mathbf{v}, \tau^{-1} \mathbb{I})$$

- ▶ Number of SNPs \gg samples → Overfitting
- ▶ Effective priors on \mathbf{v} for sparsity

Bayesian variable selection logistic regression (BVSLR)

$$p(\phi_n = 1 | \mathbf{x}_n, \mathbf{v}) = \text{lf}\left(v_0 + \sum_i v_i x_{ni}\right)$$

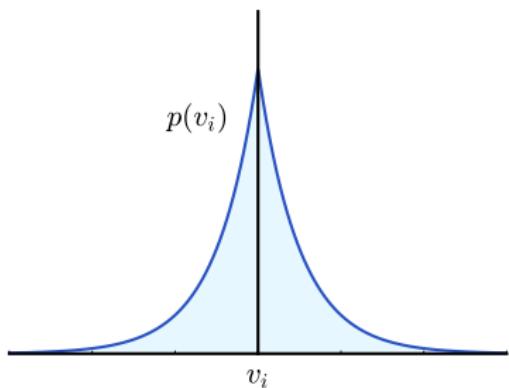


- Likelihood for N patients:

$$p(\boldsymbol{\phi} | \mathbf{x}, \mathbf{v}) = \prod_{n=1}^N p(\phi_n | \mathbf{x}_n, \mathbf{v}) = \prod_{n=1}^N \frac{\exp(\phi_n \mathbf{v}^\top \mathbf{x}_n)}{1 + \exp(\mathbf{v}^\top \mathbf{x}_n)}$$

LASSO penalisation

- ▶ Constraint: $\sum_i |v_i| \leq t$, where $t(> 0)$ is a *tuning parameter*.
- ▶ Applied as a Lagrangian penalty to the joint log-likelihood.



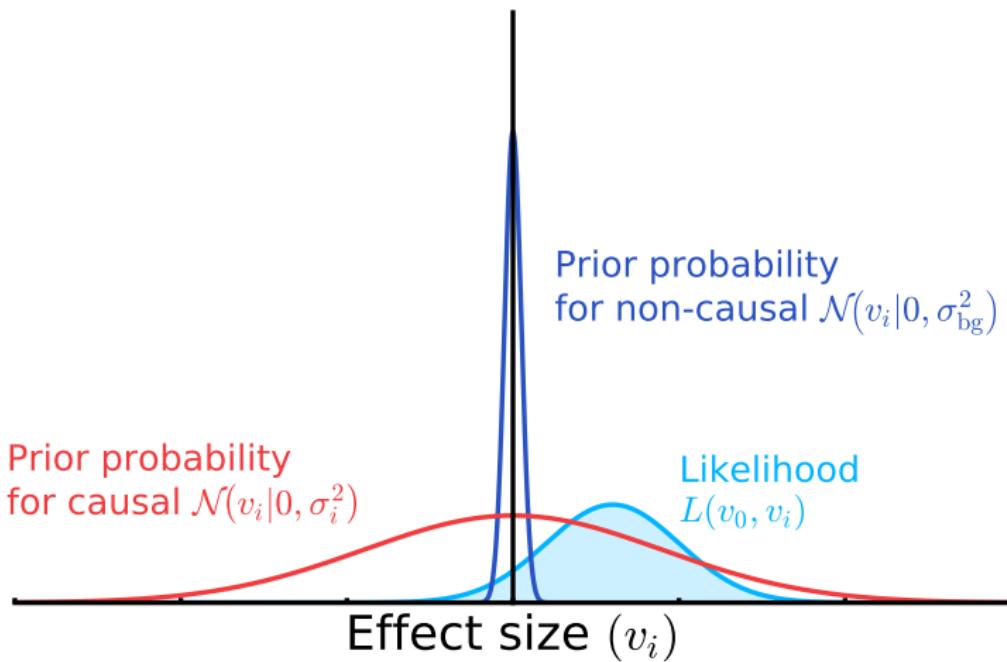
LASSO penalty assumes implicit Laplace prior \rightarrow more zero-valued coefficients.

Problems:

- ▶ Multicollinearity
- ▶ Variability in penalty parameter

Sparsity in BVSR / BVSLR

Looks at both **null hypothesis** and **alternate hypothesis**



Priors in BVSR

$$p(\mathbf{y} | \mathbf{x}, \mathbf{v}, \tau) = \mathcal{N}(\mathbf{y} | \mathbf{x}^\top \mathbf{v}, \tau^{-1} \mathbb{I})$$

$$p(\tau) = \text{Gamma}\left(\frac{\lambda}{2}, \frac{\kappa}{2}\right)$$

$$p(v_0 | \tau) = \mathcal{N}(v_0 | 0, \sigma_\mu^2 / \tau)$$

$$p(v_i | \tau) = \underbrace{(1 - \pi) \delta_0}_{\text{Non-causal}} + \underbrace{\pi \mathcal{N}(v_i | 0, \sigma_a^2 / \tau)}_{\text{Causal}}$$

Hyperparameters $\Rightarrow \lambda, \kappa, \pi, \sigma_\mu, \sigma_a$

Priors in BVSLR

$$p(\boldsymbol{\phi} | \mathbf{x}, \mathbf{v}) = \prod_{n=1}^N \frac{\exp(\phi_n \mathbf{v}^\top \mathbf{x}_n)}{1 + \exp(\mathbf{v}^\top \mathbf{x}_n)}$$

$$p(v_i | \boldsymbol{\theta}) = \underbrace{(1 - \pi) \mathcal{N}(v_i | 0, \sigma_{bg}^2)}_{\text{Non-causal}} + \underbrace{\pi \mathcal{N}(v_i | \mu, \sigma^2)}_{\text{Causal}}$$

Hyperparameters $\boldsymbol{\theta}$ $\Rightarrow \pi, \mu, \sigma_{bg}, \sigma$

Causality configurations

$$\begin{aligned} p(v_i | \boldsymbol{\theta}) &= \underbrace{(1 - \pi) \mathcal{N}(v_i | 0, \sigma_{\text{bg}}^2)}_{\text{Non-causal}} + \underbrace{\pi \mathcal{N}(v_i | \mu, \sigma^2)}_{\text{Causal}} \\ &= \sum_{z_i=0,1} \pi^{z_i} (1 - \pi)^{(1-z_i)} \mathcal{N}(v_i | \mu_{\mathbf{z},i}, \sigma_{\mathbf{z},i}^2) \\ \mu_{\mathbf{z},i} &= z_i \mu \quad \text{and} \quad \sigma_{\mathbf{z},i}^2 = \sigma_{\text{bg}}^2 + z_i [\sigma^2 - \sigma_{\text{bg}}^2] \end{aligned}$$

$\mathbf{z} \in \{0,1\}^I \Rightarrow \text{Causality configurations}$

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

Optimising the hyperparameters

$$p(v_i | \theta) = \sum_{z_i=0,1} \pi^{z_i} (1-\pi)^{(1-z_i)} \mathcal{N}(v_i | \mu_{\mathbf{z},i}, \sigma_{\mathbf{z},i}^2)$$

Optimising the hyperparameters

$$p(v_i | \boldsymbol{\theta}) = \sum_{z_i=0,1} \pi^{z_i} (1-\pi)^{(1-z_i)} \mathcal{N}(v_i | \mu_{\mathbf{z},i}, \sigma_{\mathbf{z},i}^2)$$

$$p(\mathbf{v} | \boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\mathbf{z} | \boldsymbol{\theta}) \mathcal{N}(\mathbf{v} | \boldsymbol{\mu}_{\mathbf{z}}, \text{diag}(\boldsymbol{\sigma}_{\mathbf{z}}^2))$$

Optimising the hyperparameters

$$p(\mathbf{v} | \boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\mathbf{z} | \boldsymbol{\theta}) \mathcal{N}(\mathbf{v} | \boldsymbol{\mu}_{\mathbf{z}}, \text{diag}(\boldsymbol{\sigma}_{\mathbf{z}}^2))$$

Optimising the hyperparameters

$$p(\mathbf{v} | \boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\mathbf{z} | \boldsymbol{\theta}) \mathcal{N}(\mathbf{v} | \boldsymbol{\mu}_{\mathbf{z}}, \text{diag}(\boldsymbol{\sigma}_{\mathbf{z}}^2))$$

$$p(\boldsymbol{\phi} | \mathbf{x}, \mathbf{v}) = \prod_{n=1}^N \frac{\exp(\phi_n \mathbf{v}^\top \mathbf{x}_n)}{1 + \exp(\mathbf{v}^\top \mathbf{x}_n)}$$

Optimising the hyperparameters

$$p(\mathbf{v} | \boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\mathbf{z} | \boldsymbol{\theta}) \mathcal{N}(\mathbf{v} | \boldsymbol{\mu}_{\mathbf{z}}, \text{diag}(\boldsymbol{\sigma}_{\mathbf{z}}^2))$$

$$p(\boldsymbol{\phi} | \mathbf{x}, \mathbf{v}) = \prod_{n=1}^N \frac{\exp(\phi_n \mathbf{v}^\top \mathbf{x}_n)}{1 + \exp(\mathbf{v}^\top \mathbf{x}_n)}$$

Evidence approximation: Maximising the marginal likelihood

$$\begin{aligned} mL(\boldsymbol{\theta}) &= p(\boldsymbol{\phi} | \mathbf{x}, \boldsymbol{\theta}) \\ &= \int p(\boldsymbol{\phi} | \mathbf{x}, \mathbf{v}) p(\mathbf{v} | \boldsymbol{\theta}) d\mathbf{v} \rightarrow \max \end{aligned}$$

$$mL(\boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\mathbf{z} | \boldsymbol{\theta}) \int p(\boldsymbol{\phi} | \mathbf{x}, \mathbf{v}) \mathcal{N}(\mathbf{v} | \boldsymbol{\mu}_{\mathbf{z}}, \text{diag}(\boldsymbol{\sigma}_{\mathbf{z}}^2)) d\mathbf{v}$$

Optimising the hyperparameters

$$p(\mathbf{v} | \boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\mathbf{z} | \boldsymbol{\theta}) \mathcal{N}(\mathbf{v} | \boldsymbol{\mu}_{\mathbf{z}}, \text{diag}(\boldsymbol{\sigma}_{\mathbf{z}}^2))$$

$$p(\boldsymbol{\phi} | \mathbf{x}, \mathbf{v}) = \prod_{n=1}^N \frac{\exp(\phi_n \mathbf{v}^\top \mathbf{x}_n)}{1 + \exp(\mathbf{v}^\top \mathbf{x}_n)}$$

Evidence approximation: Maximising the marginal likelihood

$$\begin{aligned} mL(\boldsymbol{\theta}) &= p(\boldsymbol{\phi} | \mathbf{x}, \boldsymbol{\theta}) \\ &= \int p(\boldsymbol{\phi} | \mathbf{x}, \mathbf{v}) p(\mathbf{v} | \boldsymbol{\theta}) d\mathbf{v} \rightarrow \max \end{aligned}$$

$$mL(\boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\mathbf{z} | \boldsymbol{\theta}) \int p(\boldsymbol{\phi} | \mathbf{x}, \mathbf{v}) \mathcal{N}(\mathbf{v} | \boldsymbol{\mu}_{\mathbf{z}}, \text{diag}(\boldsymbol{\sigma}_{\mathbf{z}}^2)) d\mathbf{v}$$

Laplace approximation (?)

Optimising the hyperparameters

$$p(\mathbf{v} | \boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\mathbf{z} | \boldsymbol{\theta}) \mathcal{N}(\mathbf{v} | \boldsymbol{\mu}_{\mathbf{z}}, \text{diag}(\boldsymbol{\sigma}_{\mathbf{z}}^2))$$

$$p(\boldsymbol{\phi} | \mathbf{x}, \mathbf{v}) = \prod_{n=1}^N \frac{\exp(\phi_n \mathbf{v}^\top \mathbf{x}_n)}{1 + \exp(\mathbf{v}^\top \mathbf{x}_n)}$$

Evidence approximation: Maximising the marginal likelihood

$$\begin{aligned} mL(\boldsymbol{\theta}) &= p(\boldsymbol{\phi} | \mathbf{x}, \boldsymbol{\theta}) \\ &= \int p(\boldsymbol{\phi} | \mathbf{x}, \mathbf{v}) p(\mathbf{v} | \boldsymbol{\theta}) d\mathbf{v} \rightarrow \max \end{aligned}$$

$$mL(\boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\mathbf{z} | \boldsymbol{\theta}) \int p(\boldsymbol{\phi} | \mathbf{x}, \mathbf{v}) \mathcal{N}(\mathbf{v} | \boldsymbol{\mu}_{\mathbf{z}}, \text{diag}(\boldsymbol{\sigma}_{\mathbf{z}}^2)) d\mathbf{v}$$

Laplace approximation (?)

Quasi-Laplace approximation

$$mL(\boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\mathbf{z} | \boldsymbol{\theta}) \int p(\boldsymbol{\phi} | \mathbf{x}, \mathbf{v}) \mathcal{N}(\mathbf{v} | \boldsymbol{\mu}_{\mathbf{z}}, \text{diag}(\boldsymbol{\sigma}_{\mathbf{z}}^2)) d\mathbf{v}$$
$$p(\boldsymbol{\phi} | \mathbf{x}, \mathbf{v}) \mathcal{N}(\mathbf{v} | \boldsymbol{\mu}_{\mathbf{z}}, \text{diag}(\boldsymbol{\sigma}_{\mathbf{z}}^2))$$
$$= p(\boldsymbol{\phi} | \mathbf{x}, \mathbf{v}) \mathcal{N}(\mathbf{v} | \tilde{\boldsymbol{\mu}}, \text{diag}(\tilde{\boldsymbol{\sigma}}^2)) \times \frac{\mathcal{N}(\mathbf{v} | \boldsymbol{\mu}_{\mathbf{z}}, \text{diag}(\boldsymbol{\sigma}_{\mathbf{z}}^2))}{\mathcal{N}(\mathbf{v} | \tilde{\boldsymbol{\mu}}, \text{diag}(\tilde{\boldsymbol{\sigma}}^2))}$$
$$\propto \mathcal{N}(\mathbf{v} | \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\sigma}}^2)$$

Quasi-Laplace approximation

$$mL(\boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\mathbf{z} | \boldsymbol{\theta}) \int p(\boldsymbol{\phi} | \mathbf{x}, \mathbf{v}) \mathcal{N}(\mathbf{v} | \boldsymbol{\mu}_{\mathbf{z}}, \text{diag}(\boldsymbol{\sigma}_{\mathbf{z}}^2)) d\mathbf{v}$$
$$p(\boldsymbol{\phi} | \mathbf{x}, \mathbf{v}) \mathcal{N}(\mathbf{v} | \boldsymbol{\mu}_{\mathbf{z}}, \text{diag}(\boldsymbol{\sigma}_{\mathbf{z}}^2))$$
$$= \underbrace{p(\boldsymbol{\phi} | \mathbf{x}, \mathbf{v}) \mathcal{N}(\mathbf{v} | \tilde{\boldsymbol{\mu}}, \text{diag}(\tilde{\boldsymbol{\sigma}}^2))}_{\propto \mathcal{N}(\mathbf{v} | \tilde{\mathbf{v}}, \tilde{\boldsymbol{\Lambda}}^{-1})} \times \frac{\mathcal{N}(\mathbf{v} | \boldsymbol{\mu}_{\mathbf{z}}, \text{diag}(\boldsymbol{\sigma}_{\mathbf{z}}^2))}{\mathcal{N}(\mathbf{v} | \tilde{\boldsymbol{\mu}}, \text{diag}(\tilde{\boldsymbol{\sigma}}^2))}$$

Analytical solution

$$mL(\boldsymbol{\theta}) = p(\boldsymbol{\phi} \mid \mathbf{x}, \boldsymbol{\theta})$$

$$\approx D' \sum_{\mathbf{z}} p(\mathbf{z} \mid \boldsymbol{\theta}) \frac{|\text{diag}(\boldsymbol{\lambda}_{\mathbf{z}})|^{\frac{1}{2}}}{|\boldsymbol{\Lambda}_{\mathbf{z}}|^{\frac{1}{2}}} \exp \left(-\frac{1}{2} \boldsymbol{\mu}_{\mathbf{z}}^\top \text{diag}(\boldsymbol{\lambda}_{\mathbf{z}}) \boldsymbol{\mu}_{\mathbf{z}} + \frac{1}{2} \mathbf{v}_{\mathbf{z}}^\top \boldsymbol{\Lambda}_{\mathbf{z}} \mathbf{v}_{\mathbf{z}} \right)$$

where

$$\boldsymbol{\Lambda}_{\mathbf{z}} := \tilde{\boldsymbol{\Lambda}} - \text{diag}(\tilde{\boldsymbol{\lambda}}) + \text{diag}(\boldsymbol{\lambda}_{\mathbf{z}})$$

$$\mathbf{v}_{\mathbf{z}} := \boldsymbol{\Lambda}_{\mathbf{z}}^{-1} [\tilde{\boldsymbol{\Lambda}} \tilde{\mathbf{v}} + \text{diag}(\boldsymbol{\lambda}_{\mathbf{z}}) \boldsymbol{\mu}_{\mathbf{z}} - \text{diag}(\tilde{\boldsymbol{\lambda}}) \tilde{\boldsymbol{\mu}}]$$

- ▶ Optimization can be done by gradient descent methods (e.g. L-BFGS).

Inference of causality in BVSLR

Using the definition of conditional probability,

$$p(\mathbf{z} | \boldsymbol{\phi}, \mathbf{x}, \boldsymbol{\theta}) = \frac{p(\boldsymbol{\phi}, \mathbf{z} | \mathbf{x}, \boldsymbol{\theta})}{p(\boldsymbol{\phi} | \mathbf{x}, \boldsymbol{\theta})}$$

The posterior probability for SNP i to be causal is

$$p(z_i = 1 | \boldsymbol{\phi}, \mathbf{x}, \boldsymbol{\theta}) = \sum_{\mathbf{z}: z_i = 1} p(\mathbf{z} | \boldsymbol{\phi}, \mathbf{x}, \boldsymbol{\theta})$$

Inference of causality in BVSLR

Using the definition of conditional probability,

$$p(\mathbf{z} | \boldsymbol{\phi}, \mathbf{x}, \boldsymbol{\theta}) = \frac{p(\boldsymbol{\phi}, \mathbf{z} | \mathbf{x}, \boldsymbol{\theta})}{p(\boldsymbol{\phi} | \mathbf{x}, \boldsymbol{\theta})}$$

The posterior probability for SNP i to be causal is

$$p(z_i = 1 | \boldsymbol{\phi}, \mathbf{x}, \boldsymbol{\theta}) = \sum_{\mathbf{z}: z_i = 1} p(\mathbf{z} | \boldsymbol{\phi}, \mathbf{x}, \boldsymbol{\theta})$$

The probability for a locus to be causal = 1 – probability of *not* having a single causal SNP

$$p(\text{locus is causal} | \boldsymbol{\phi}, \mathbf{x}, \boldsymbol{\theta}) = 1 - p(\mathbf{z} = \mathbf{0} | \boldsymbol{\phi}, \mathbf{x}, \boldsymbol{\theta})$$

BVSLR can be extended to multiple studies

$$\begin{aligned} mL(\boldsymbol{\theta}) &= p(\boldsymbol{\phi} | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_S, \boldsymbol{\theta}) \\ &= \int p(\boldsymbol{\phi} | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_S, \mathbf{v}) p(\mathbf{v} | \boldsymbol{\theta}) d\mathbf{v} \rightarrow \max \end{aligned}$$

Assuming the quasi-Laplace approximation holds for each study,

$$\begin{aligned} \prod_{s=1}^S [p(\boldsymbol{\phi} | \mathbf{x}_s, \mathbf{v}) \mathcal{N}(\mathbf{v} | \tilde{\boldsymbol{\mu}}_{\mathbf{z},s}, \text{diag}(\tilde{\sigma}_{\mathbf{z},s}^2))] &\propto \prod_{s=1}^S \mathcal{N}(\mathbf{v} | \tilde{\mathbf{v}}_s, \tilde{\boldsymbol{\Lambda}}_s^{-1}) \\ &= \mathcal{N}(\mathbf{v} | \tilde{\mathbf{v}}, \tilde{\boldsymbol{\Lambda}}^{-1}) \end{aligned}$$

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

Simulation details

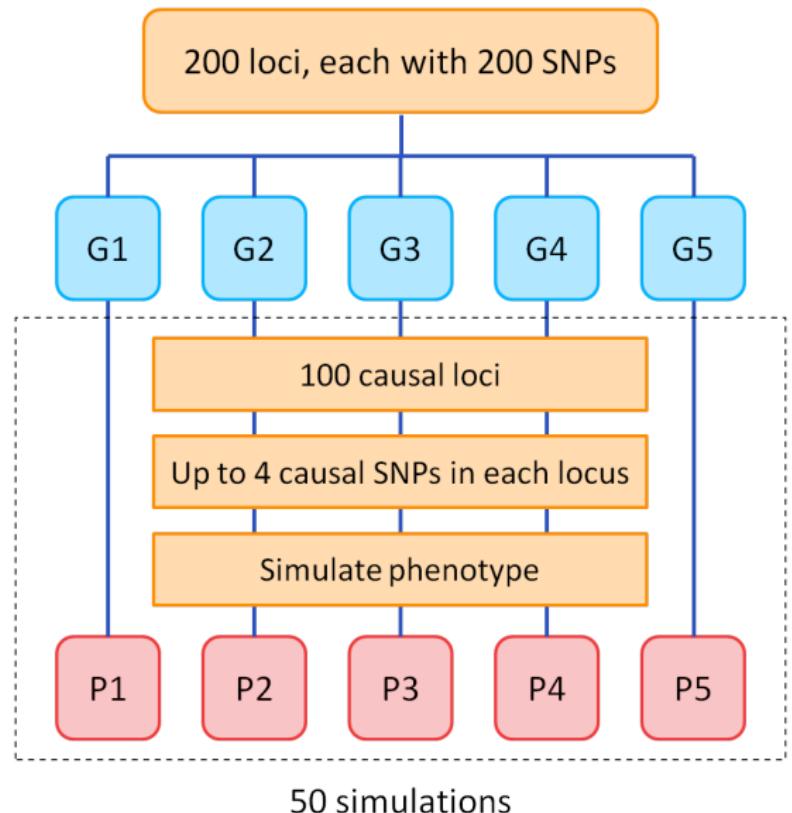
- ▶ Genotype: German Myocardial Infarction Family Study (GERMIFS)
- ▶ Five cohorts : G1, G2, G3, G4, G5
- ▶ Phenotype simulation:

$$\text{Disease liability} \quad Y_n = \sum_i v_i x_{ni} + \varepsilon_n$$

$$\text{Var}\left(\sum_i v_i x_{ni}\right) = h_g^2 = 0.4 \text{ and } \text{Var}(\varepsilon) = 0.6$$

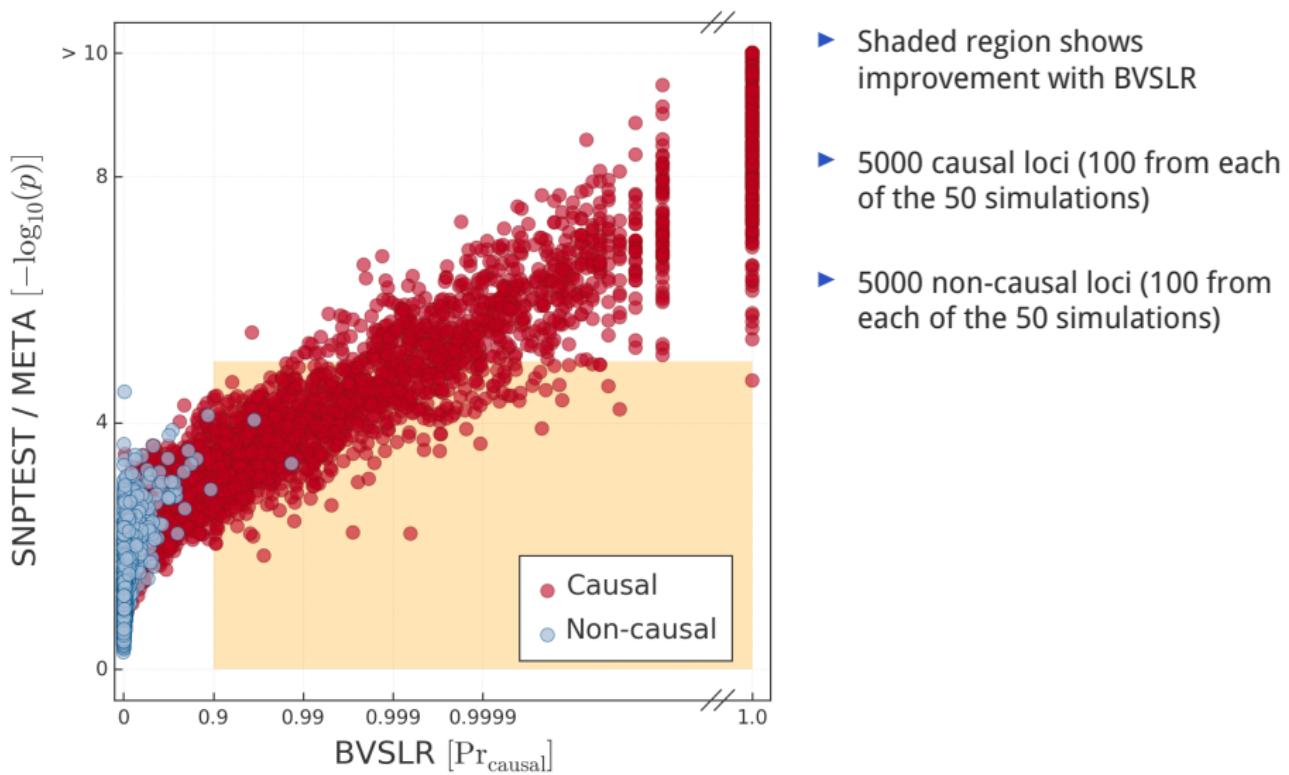
Cases are sampled from Y_n exceeding the threshold of normal distribution truncating the proportion of k (disease prevalence)

Simulation details

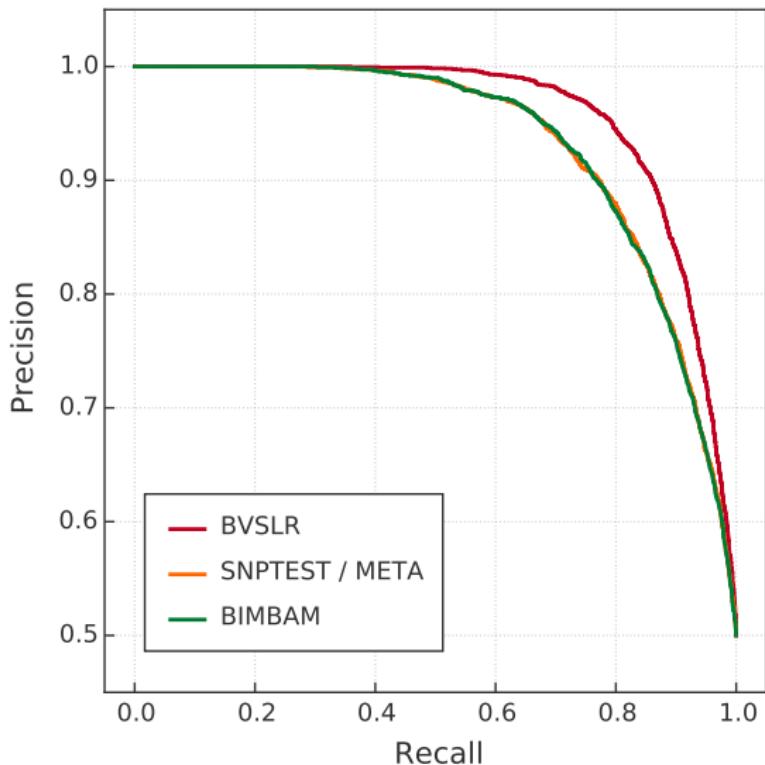


- ▶ BVSLR
- ▶ BIMBAM
- ▶ SNPTEST / META
- ▶ PAINTOR

Prediction of causal loci

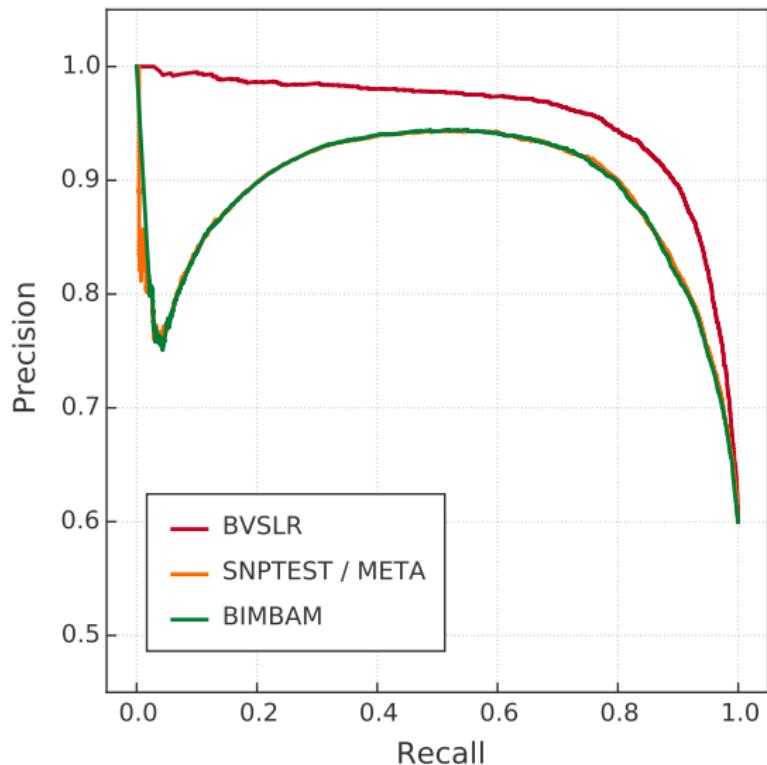


Prediction of causal loci



$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

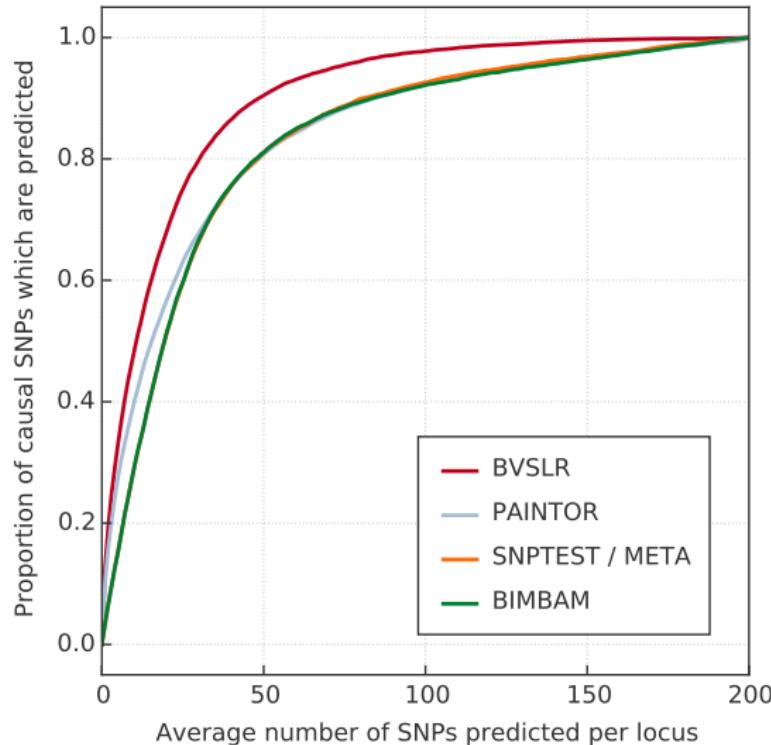
If there are non-causal loci in LD with causal regions



$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

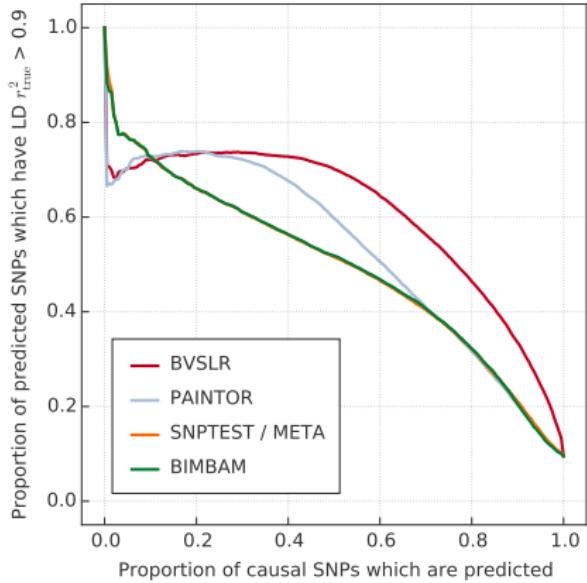
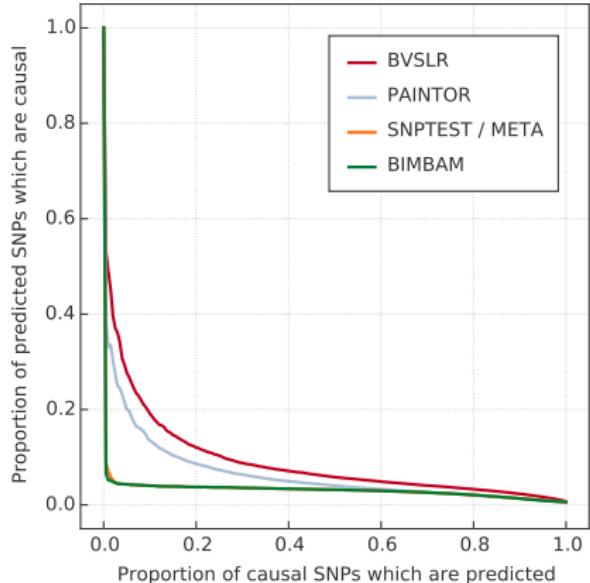
- ▶ 8 loci (out of 200) in LD with each other were introduced in the simulation

Finemapping causal variants



► Comparable to PAINTOR up to 20% recall

BVSLR predicts SNPs in strong LD with actual ones



Challenges of BVSLR

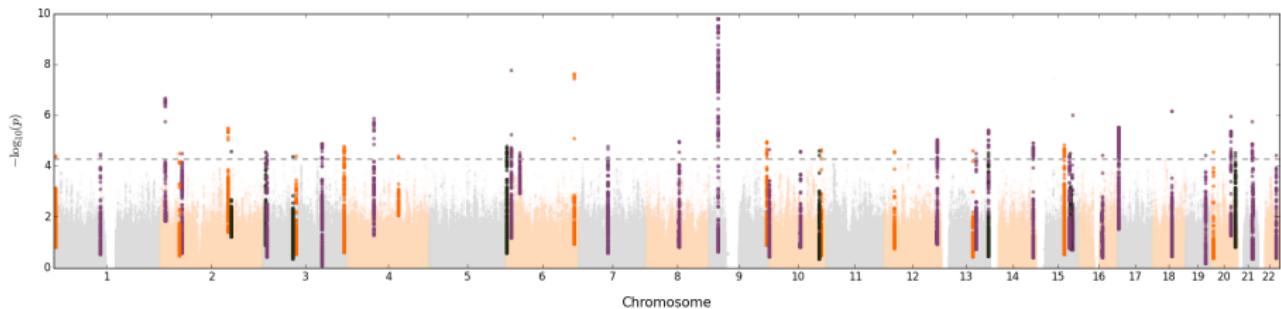
- ▶ Computational time (summing over z-states)
- ▶ New summary statistics
- ▶ Prior assumptions

Association with coronary artery diseases (CAD)

- ▶ 5 GERMIFS cohorts
- ▶ 6228 cases, 6854 controls
- ▶ Imputed with 1000G Phase 1

Association with coronary artery diseases (CAD)

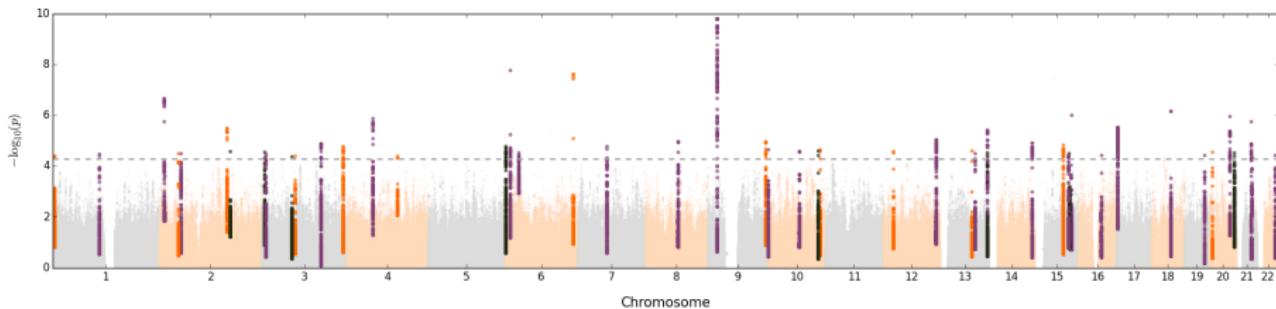
- ▶ 5 GERMIFS cohorts
- ▶ 6228 cases, 6854 controls
- ▶ Imputed with 1000G Phase 1



GWAS using SNPTEST / META

Association with coronary artery diseases (CAD)

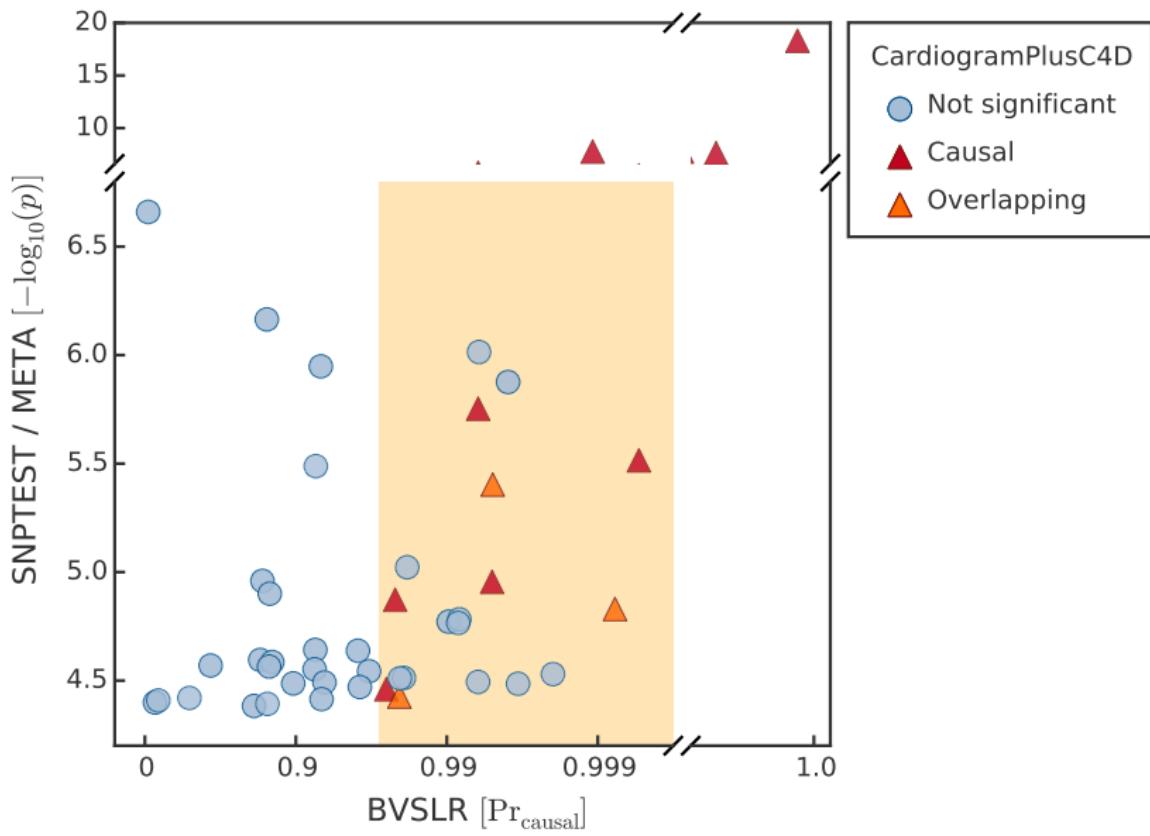
- ▶ 5 GERMIFS cohorts
- ▶ 6228 cases, 6854 controls
- ▶ Imputed with 1000G Phase 1



GWAS using SNPTEST / META

- ▶ Applied BVSLR on these 45 loci, selecting 400 SNPs at each locus.

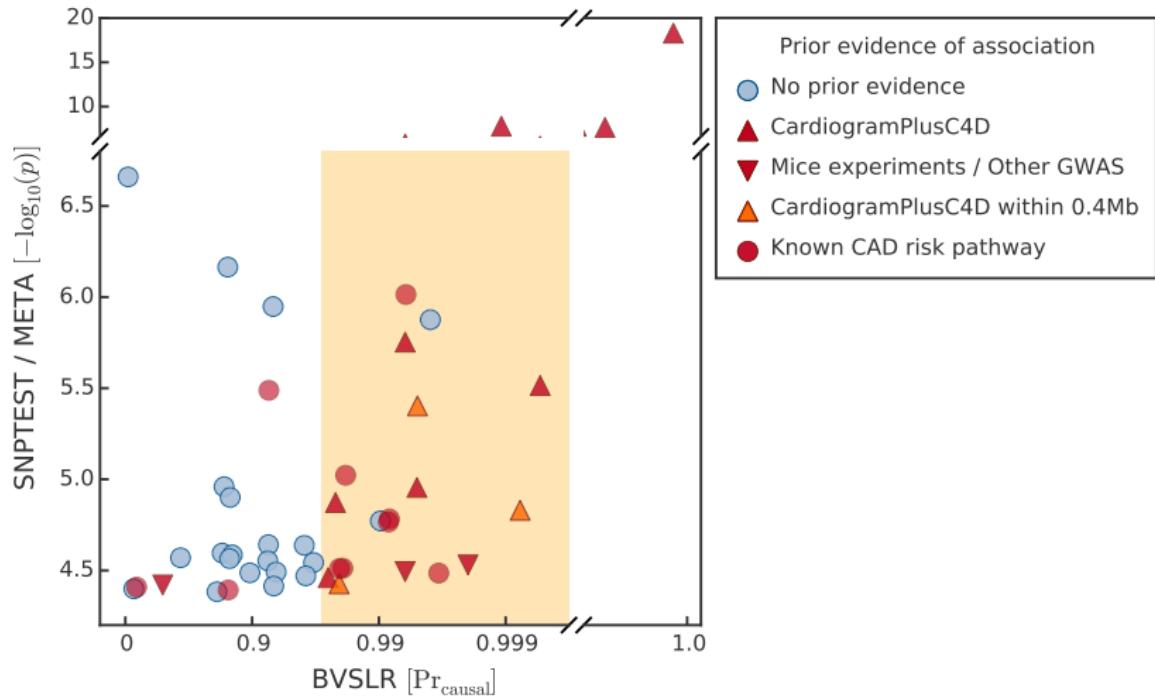
BVSLR predictions



Top BVSLR predicted loci (not discovered in CardiogramPlusC4D)

| Region | Pr | Gene | Comments |
|---------|-------|---------------|--|
| 6p21.3 | 0.998 | C6orf10-BTNL2 | GWAS for CAD in Han Chinese, 2012 |
| 15q25 | 0.997 | IL-16 | GWAS for CAD in Han Chinese, 2012 |
| 4q13.1 | 0.996 | desert | - |
| 15q25 | 0.994 | AKAP13 | C. hypertrophy (mice) / GWAS (BP in Koreans, 2011) |
| 2p16 | 0.994 | NRXN1 | GWAS for CAD in OHGS1 + WTCCC2 |
| 3q28 | 0.992 | IL1RAP | Involved in risk pathway |
| 6p25 | 0.992 | SERPINB | patented as biomarker for CVD |
| 7q11.22 | 0.990 | AUTS2 | - |
| 12q24 | 0.982 | ZNF664 | GWAS hit for HDL-C, TG |
| 13q21.1 | 0.981 | ARHGEF1 | Controls vascular tone and BP |

Literature-based classification



- ▶ Novel Bayesian method for GWAS
- ▶ Multivariate analysis in meta studies
- ▶ Precision
- ▶ Predicts new associations in CAD

Many thanks to ...



Johannes Söding



Heribert Schunkert



Jeanette Erdmann

Many thanks to ...



Johannes Söding



Heribert Schunkert

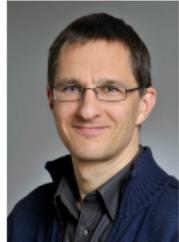


Jeanette Erdmann



AG Söding

Many thanks to ...



Johannes Söding



Heribert Schunkert



Jeanette Erdmann



AG Söding

Thank you!