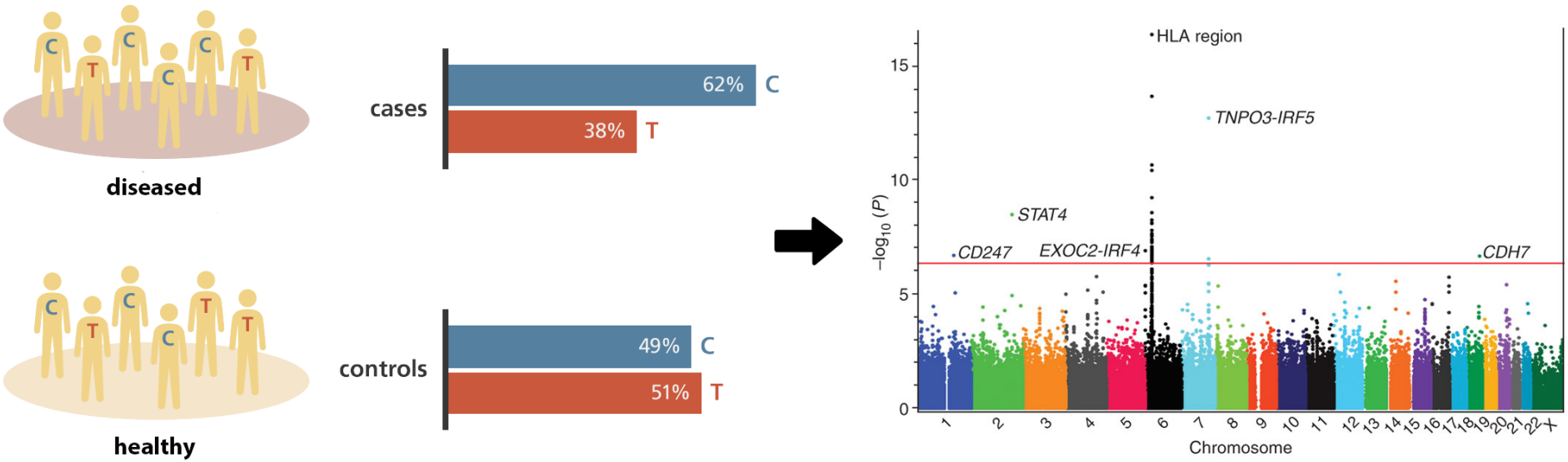
Bayesian variable selection logistic regression: multivariate metaanalysis in GWAS

Saikat Banerjee

Max Planck Institute for Biophysical Chemistry

FEBRUARY 2, 2017

Genome-wide association studies (GWAS)



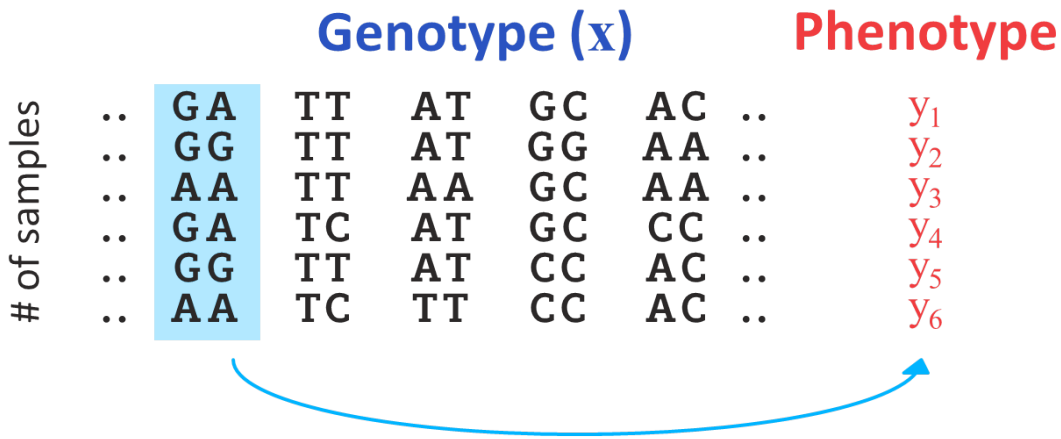
- Discovered thousands of variants associated with complex diseases

Radstake *et al.* , *Nat. Gen.* 2010

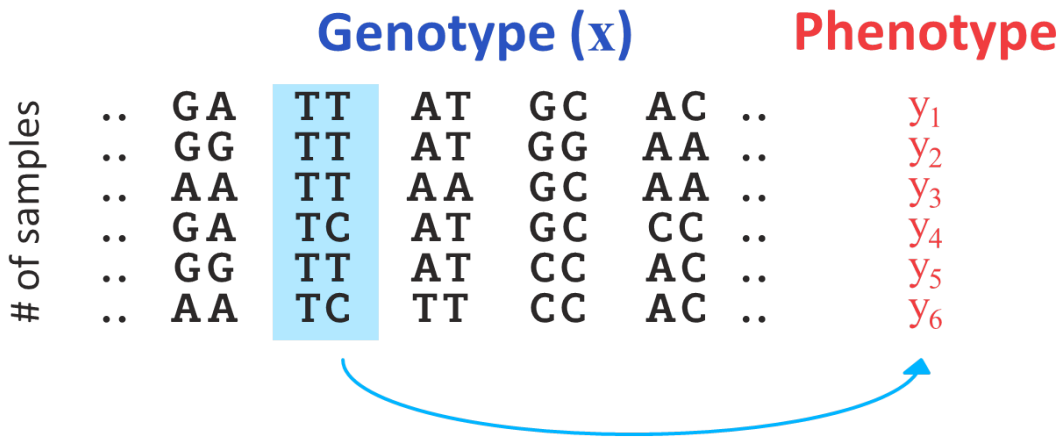
Association tests in GWAS

	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

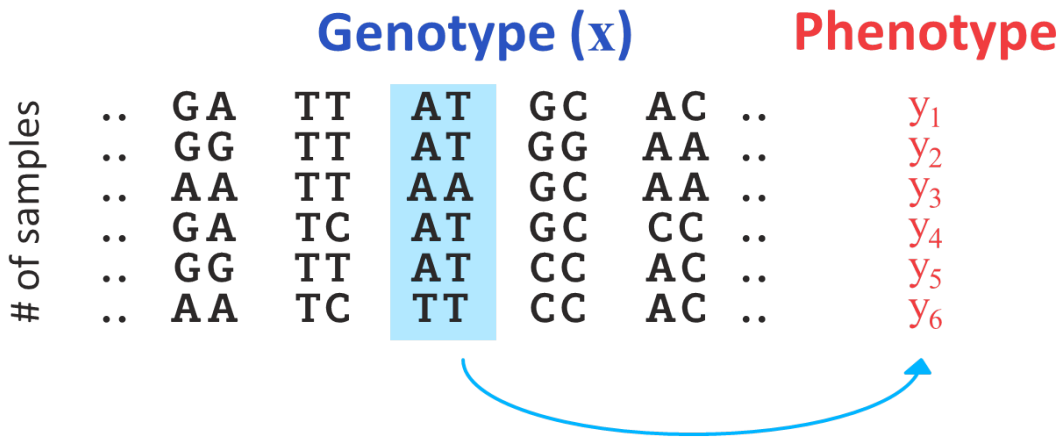
Association tests in GWAS



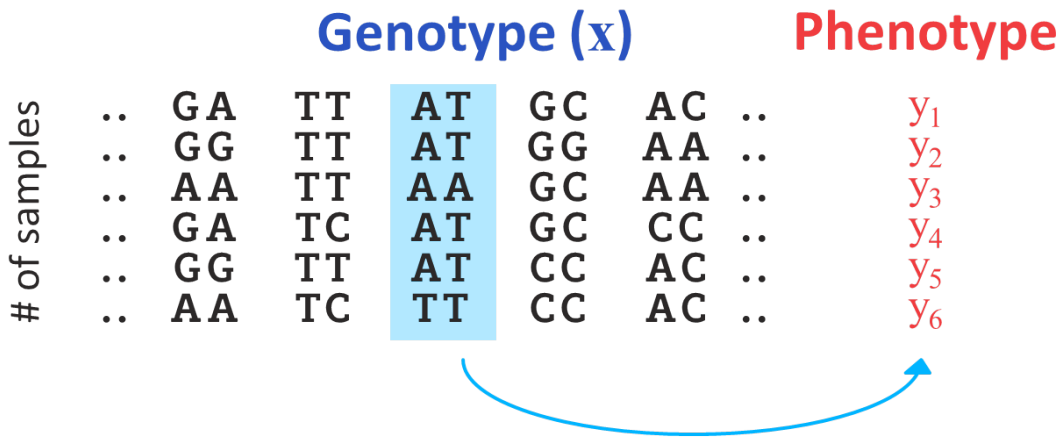
Association tests in GWAS



Association tests in GWAS

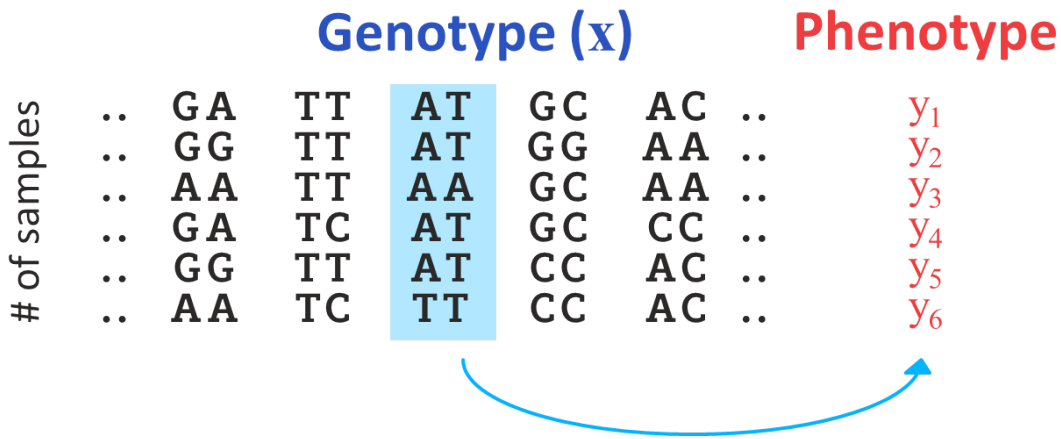


Association tests in GWAS



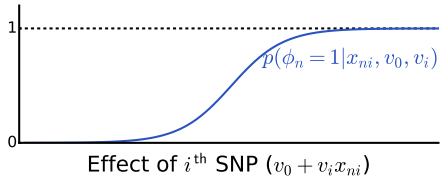
$$y_n = v_0 + v_i x_{ni} + \epsilon, \quad \text{with} \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$ Quantitative phenotype

Association tests in GWAS

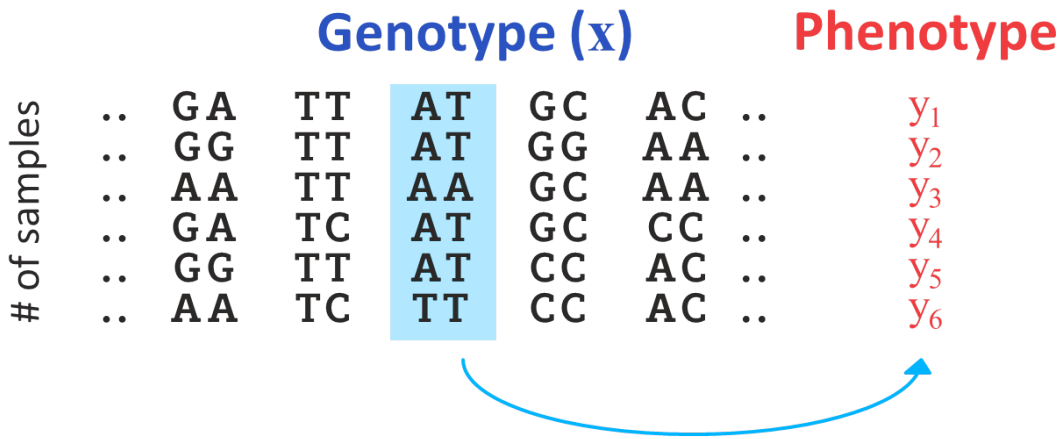


$$y_n = v_0 + v_i x_{ni} + \epsilon, \quad \text{with} \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$ Quantitative phenotype

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



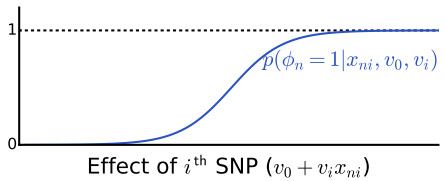
Association tests in GWAS



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

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

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



Univariate methods

Strengths

- Straightforward
- Computationally fast
- Conservative
- Easy to interpret

Univariate methods

Strengths

- Straightforward
- Computationally fast
- Conservative
- Easy to interpret

Challenges

- Linkage disequilibrium
- Genetic networks
- Low effect sizes

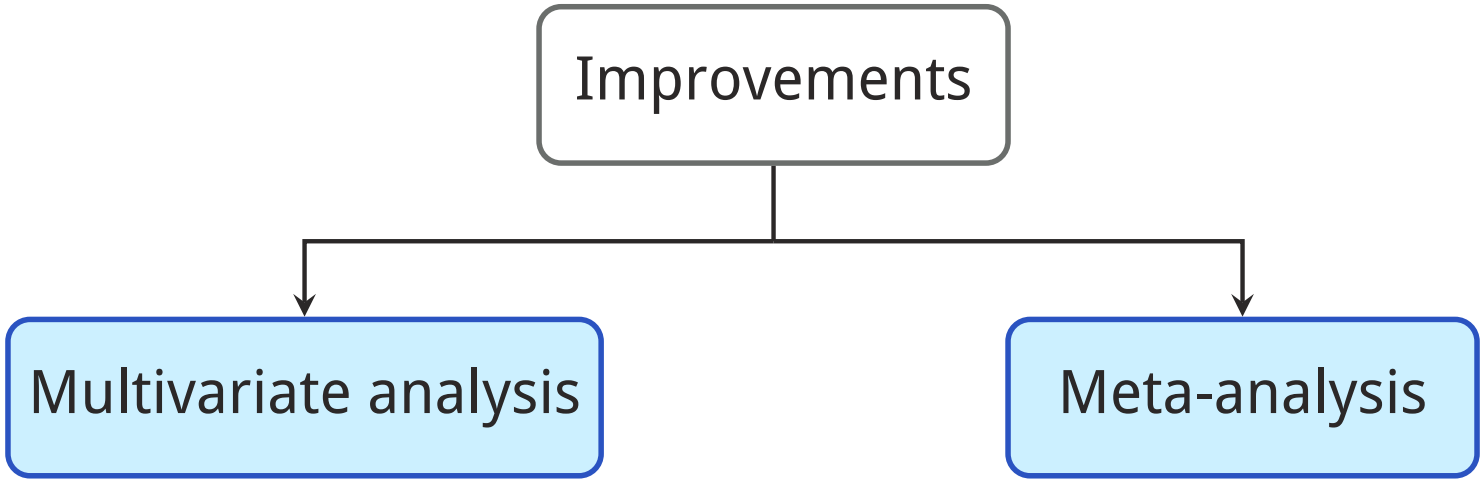
Univariate methods

Strengths

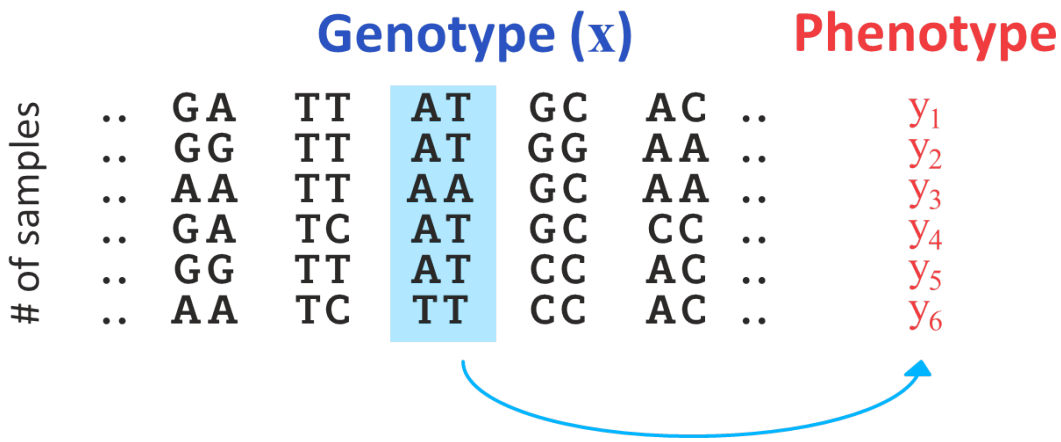
- Straightforward
- Computationally fast
- Conservative
- Easy to interpret

Challenges

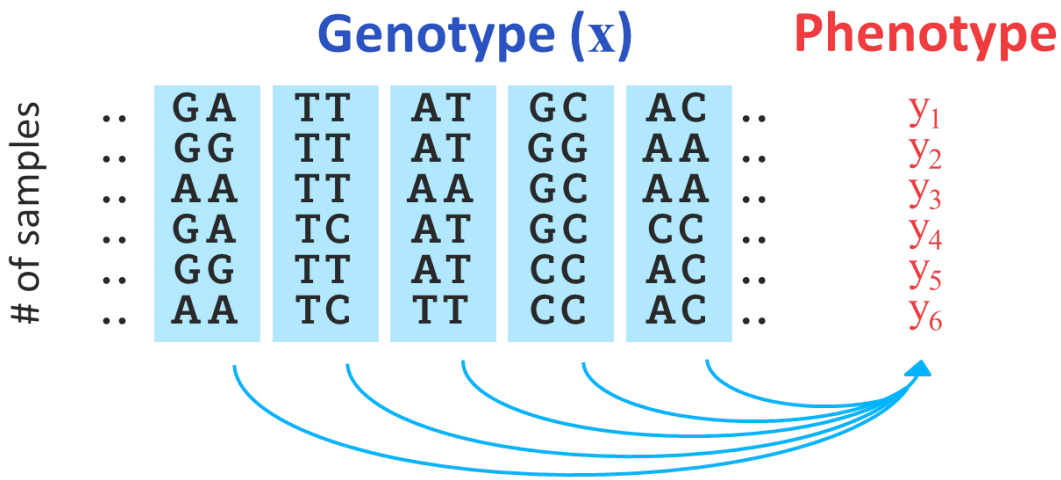
- Linkage disequilibrium
- Genetic networks
- Low effect sizes



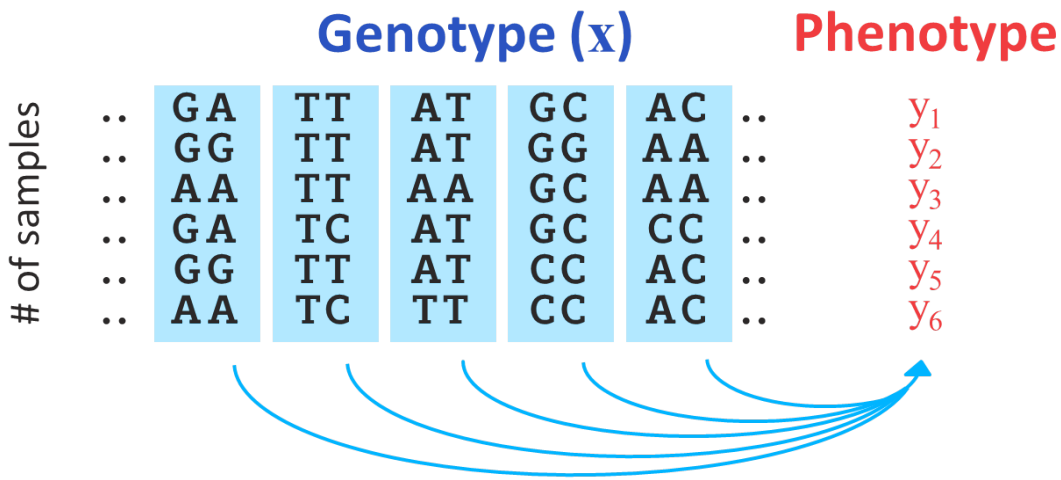
Multivariate methods



Multivariate methods



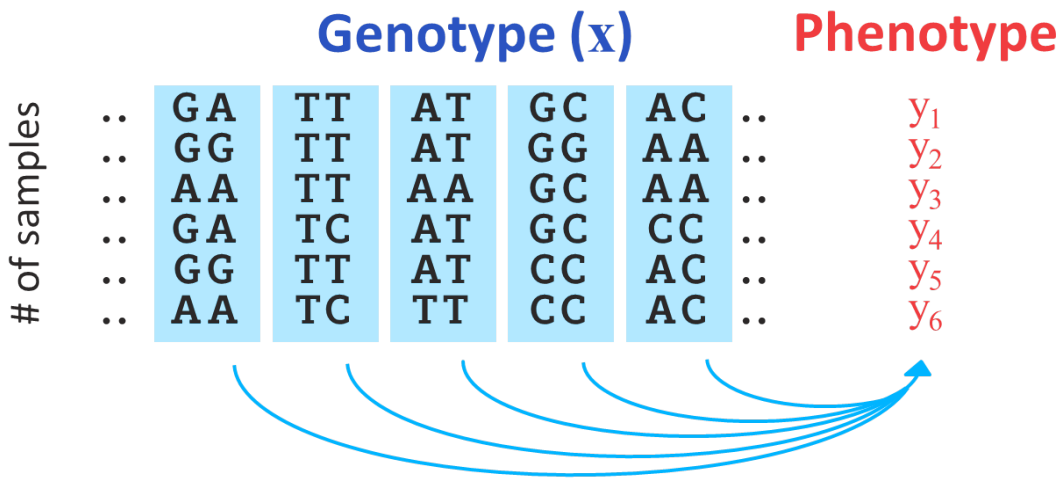
Multivariate methods



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

Quantitative phenotype

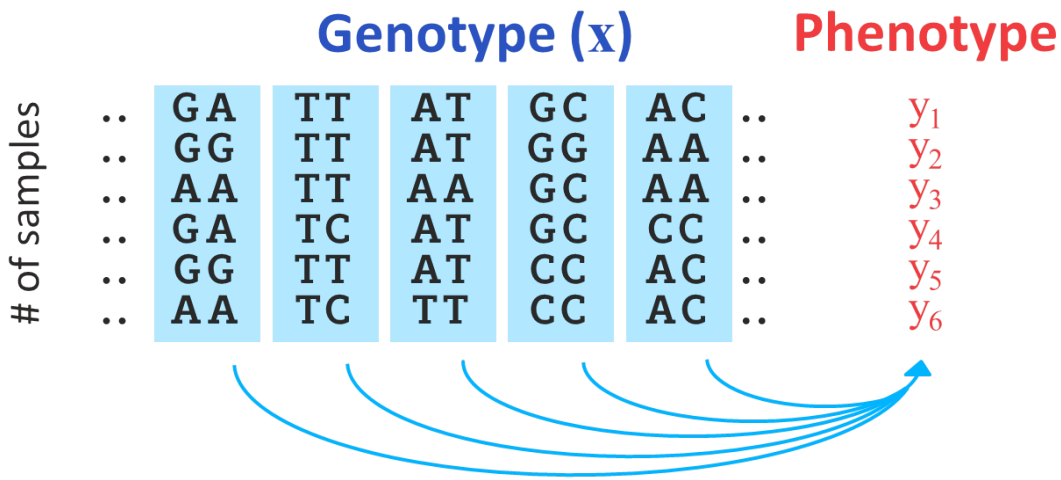
Multivariate methods



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

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

Multivariate methods

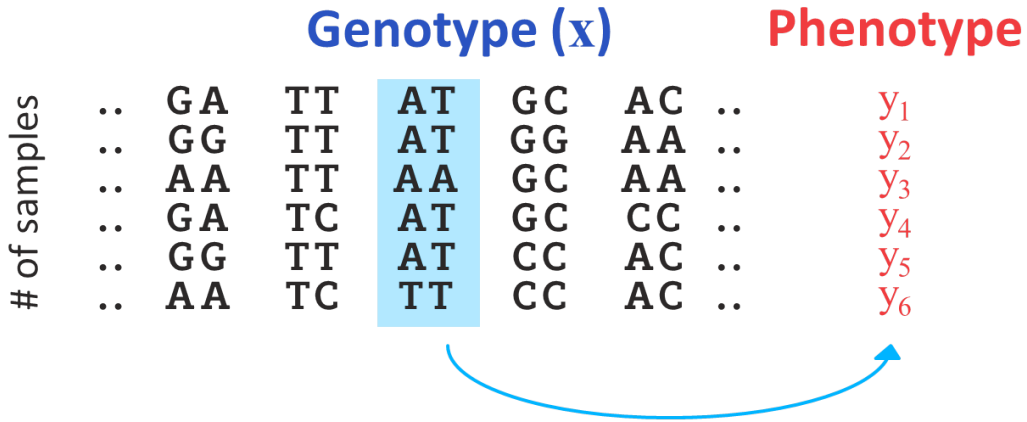


$$y_n = v_0 + \sum_i v_i x_{ni} + \epsilon, \quad \text{with} \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$ Quantitative phenotype

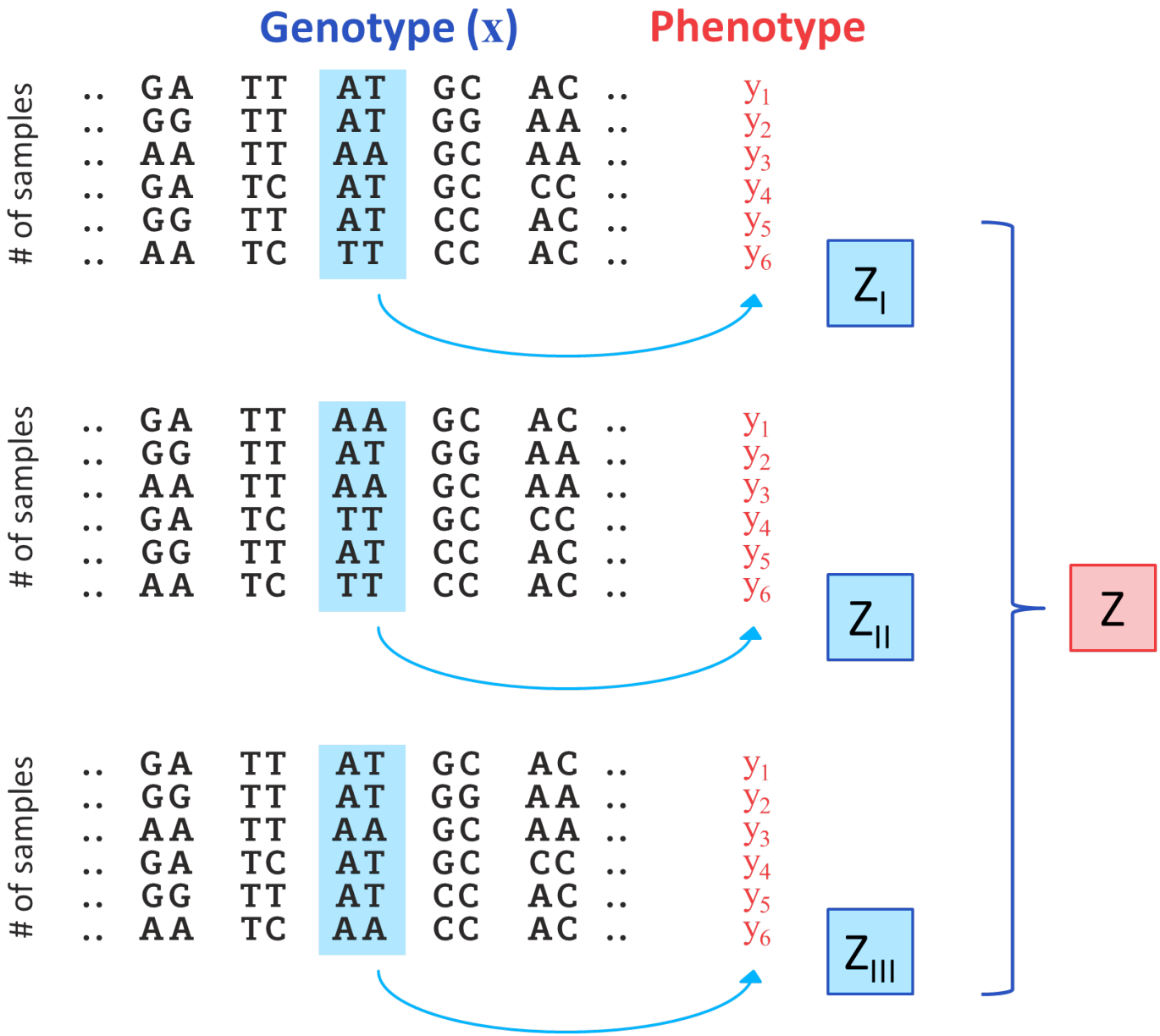
$$p(y_n = 1 | x_{ni}, v_0, v_i) = \mathrm{lf}\left(v_0 + \sum_i v_i x_{ni} \right)$$ Binary phenotype

- Multivariate methods perform better than univariate methods

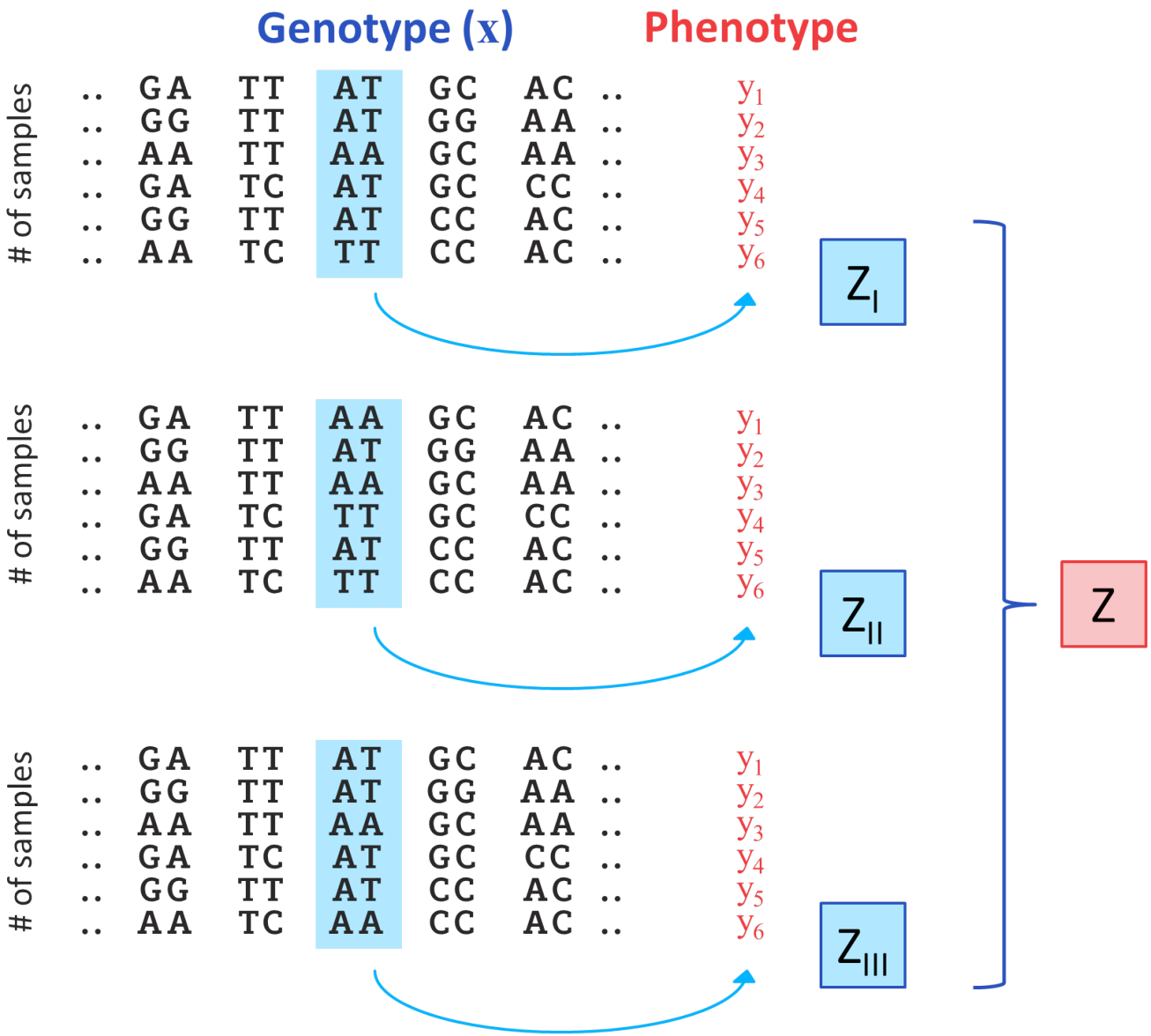
Meta-analysis



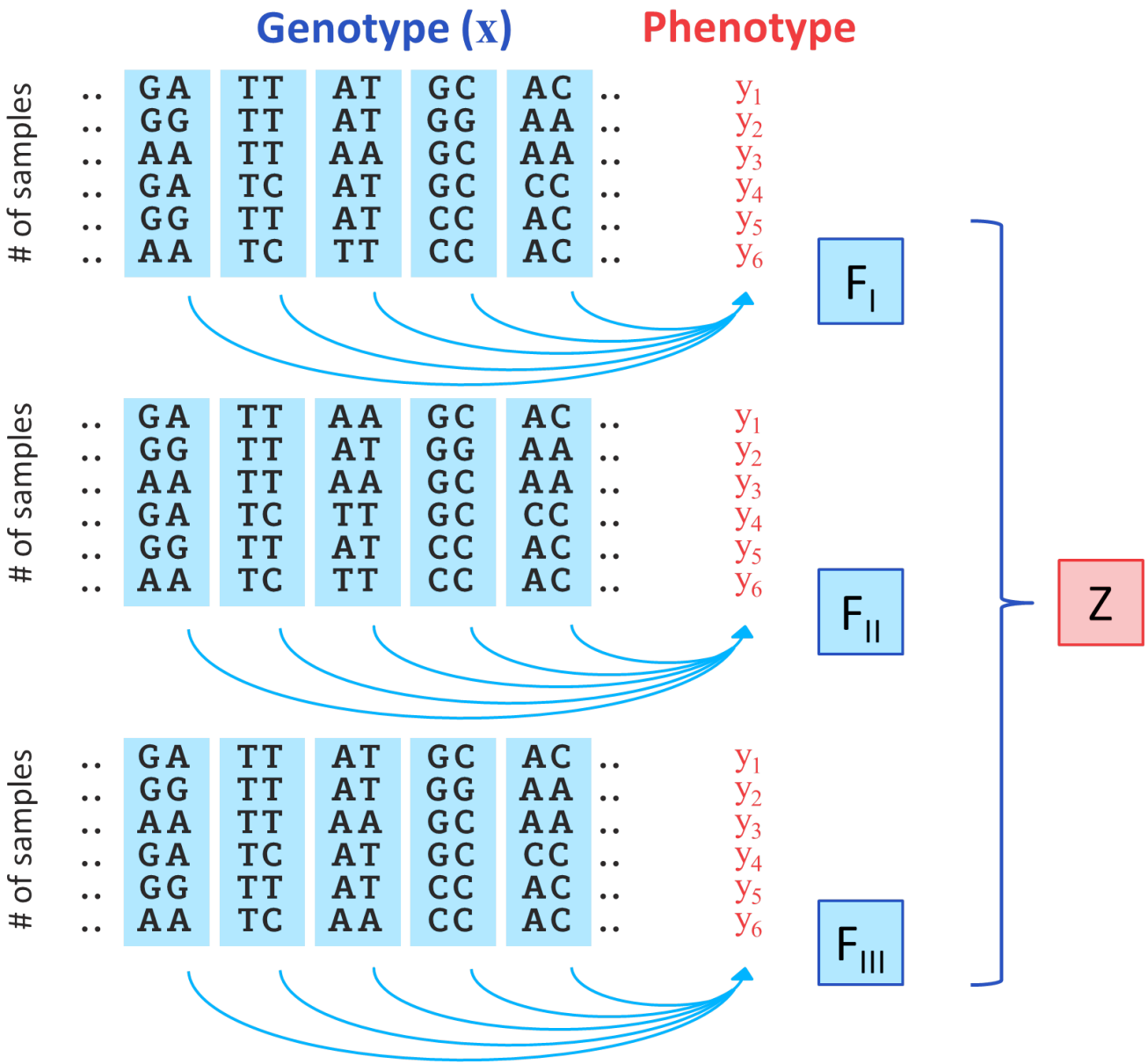
Meta-analysis



Goal of our method



Goal of our method



Bayesian variable selection regression (BVSR)

BIMBAM

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

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

Quantitative phenotype

Bayesian variable selection regression (BVSR)

BIMBAM

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

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

- Likelihood for N patients:

$$p\left(\mathbf{y} \mid \mathbf{x}, \mathbf{v}, \tau\right) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{x}^{\intercal}\mathbf{v}, \tau^{-1}\mathbb{I}\right)$$

Bayesian variable selection regression (BVSR)

BIMBAM

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

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

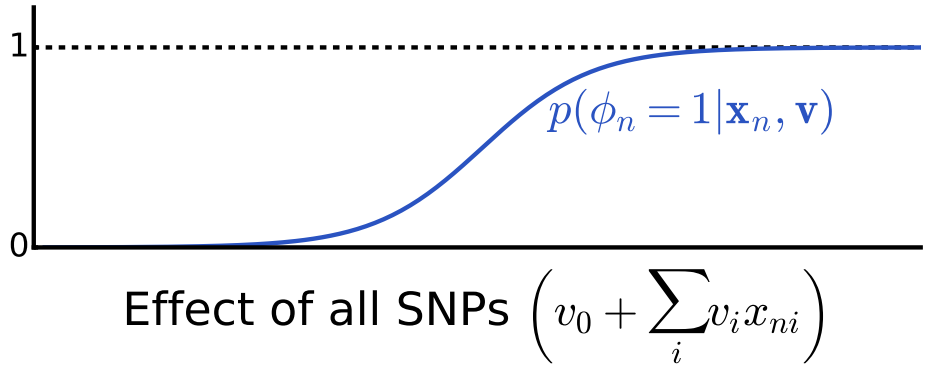
- Likelihood for N patients:

$$p\left(\mathbf{y} \mid \mathbf{x}, \mathbf{v}, \tau\right) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{x}^{\intercal}\mathbf{v}, \tau^{-1}\mathbb{I}\right)$$

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

Bayesian variable selection logistic regression (BVSLR)

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

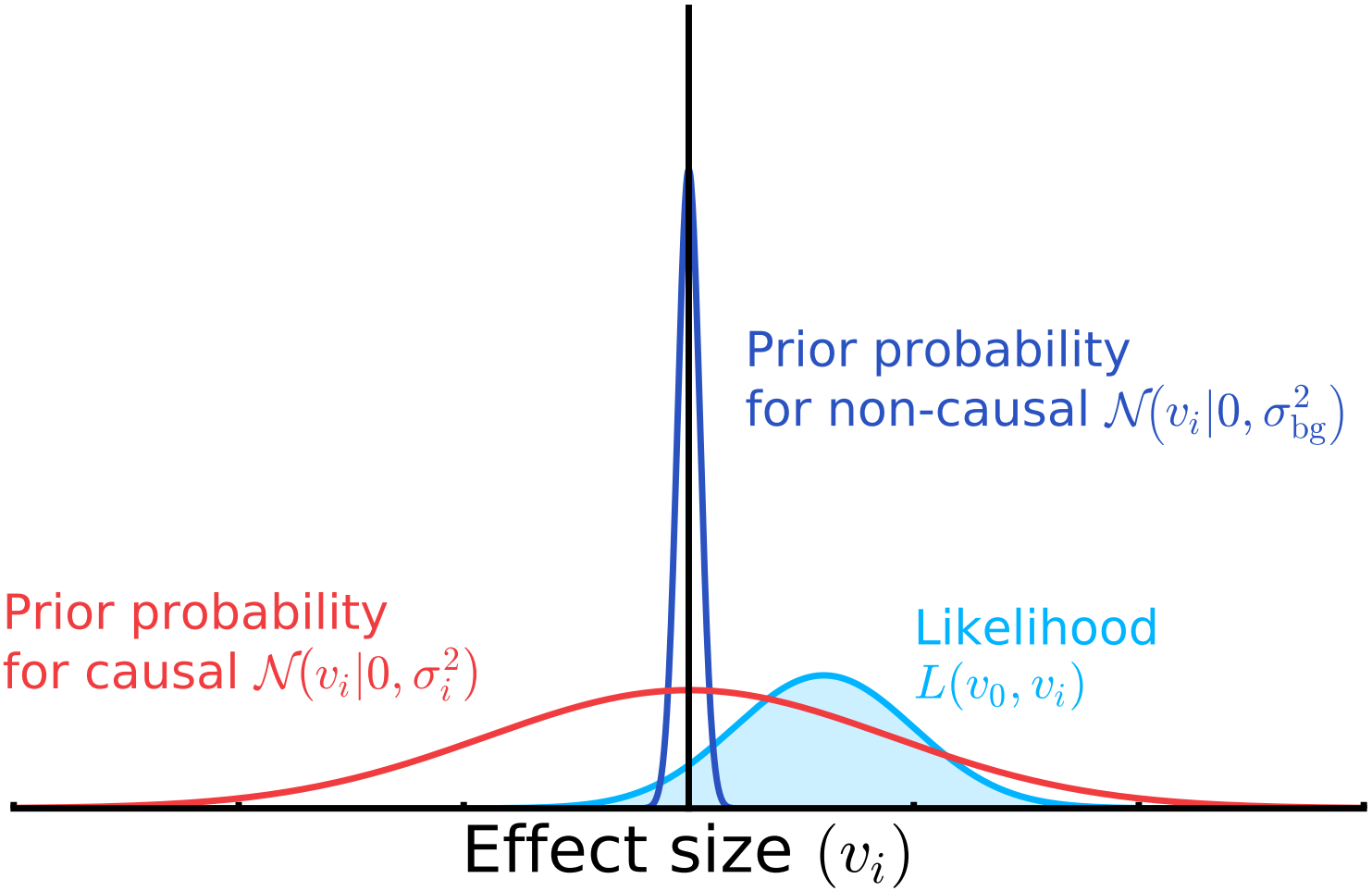


- Likelihood for N patients:

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

Sparsity in BVSR / BVSLR

Looks at both null hypothesis and alternate hypothesis



Simulation details

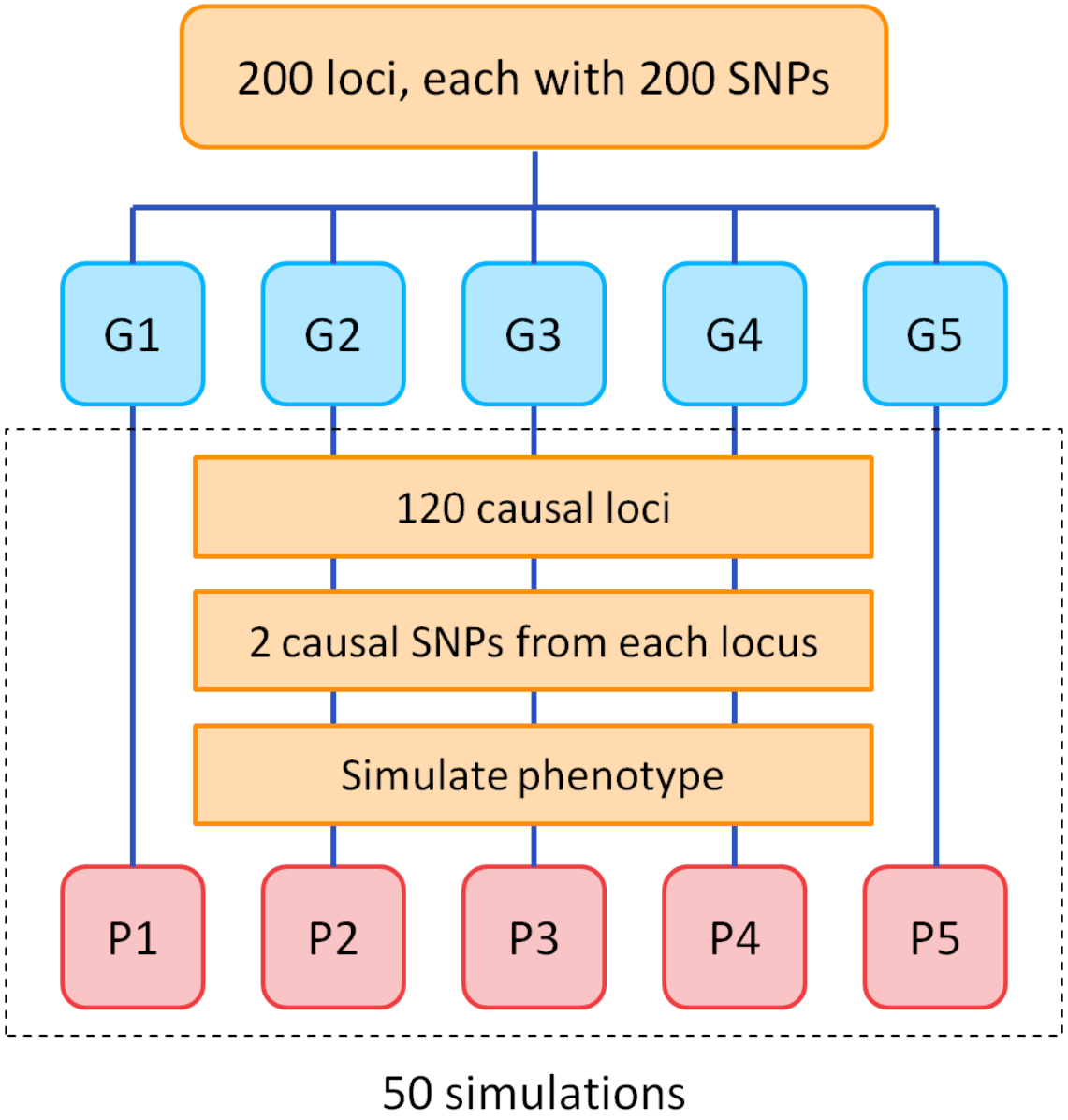
- Genotype: German Myocardial Infarction Family Study (GERMIFS)
- Five cohorts : G1, G2, G3, G4, G5
- Phenotype simulation:

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

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

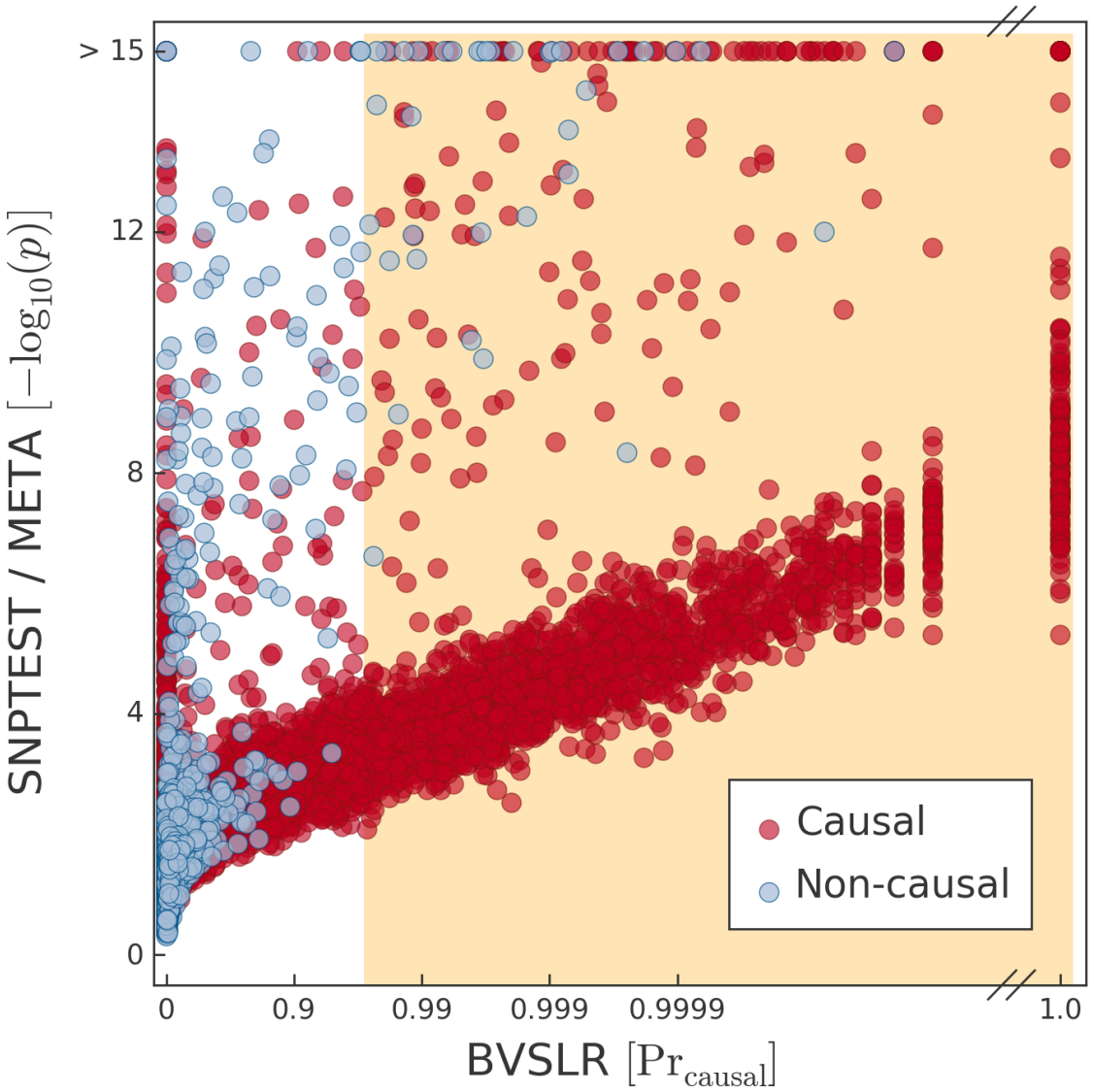
Cases are sampled from Y_n exceeding the threshold of normal distribution truncating the proportion of k (disease prevalence)

Simulation details



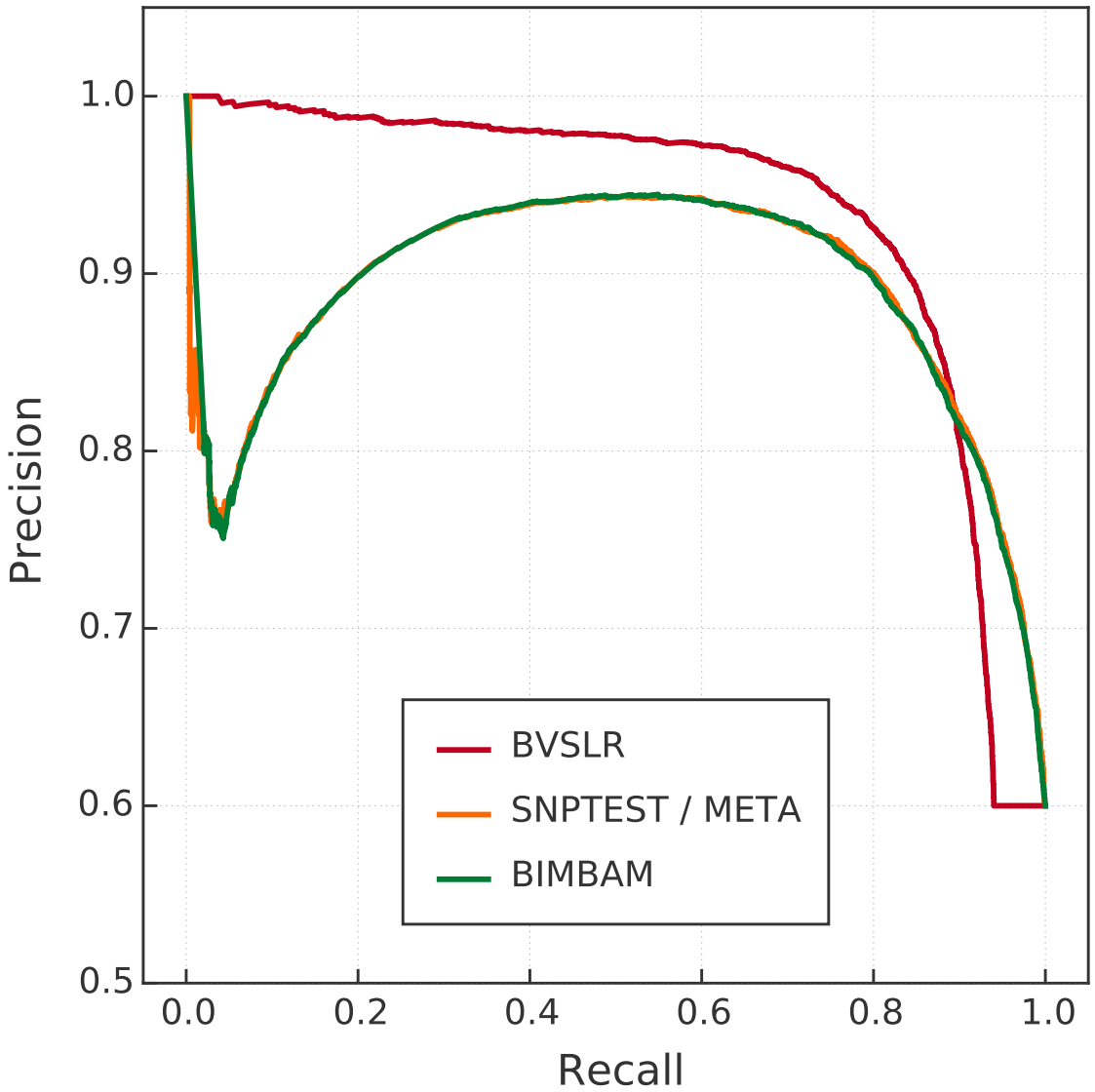
- BVSLR
- BIMBAM
- SNPTEST / META
- PAINTOR

Prediction of causal loci



- 6000 causal loci (120 from each of the 50 simulations)
- 4000 non-causal loci (80 from each of the 50 simulations)

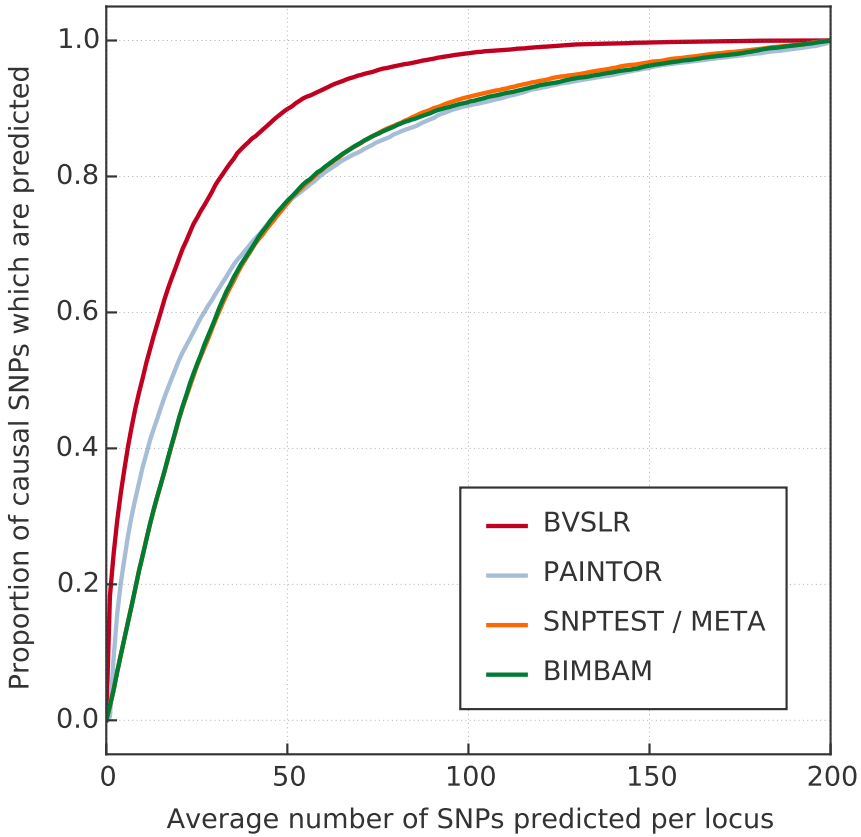
Prediction of causal loci

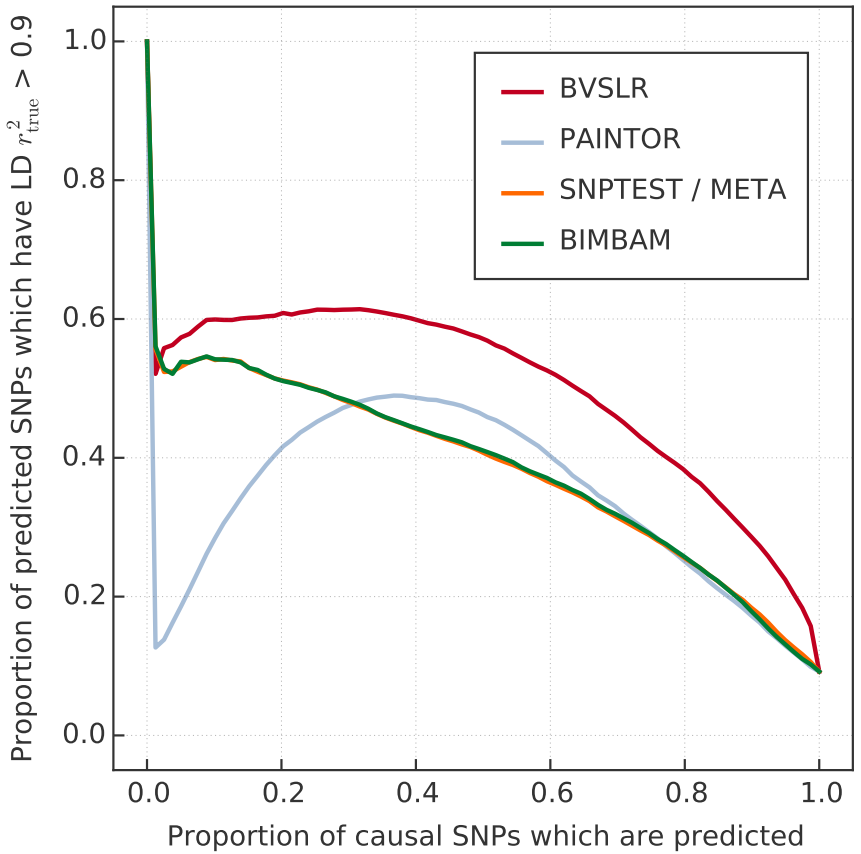


$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

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

Finemapping causal variants





Challenges of BVSLR

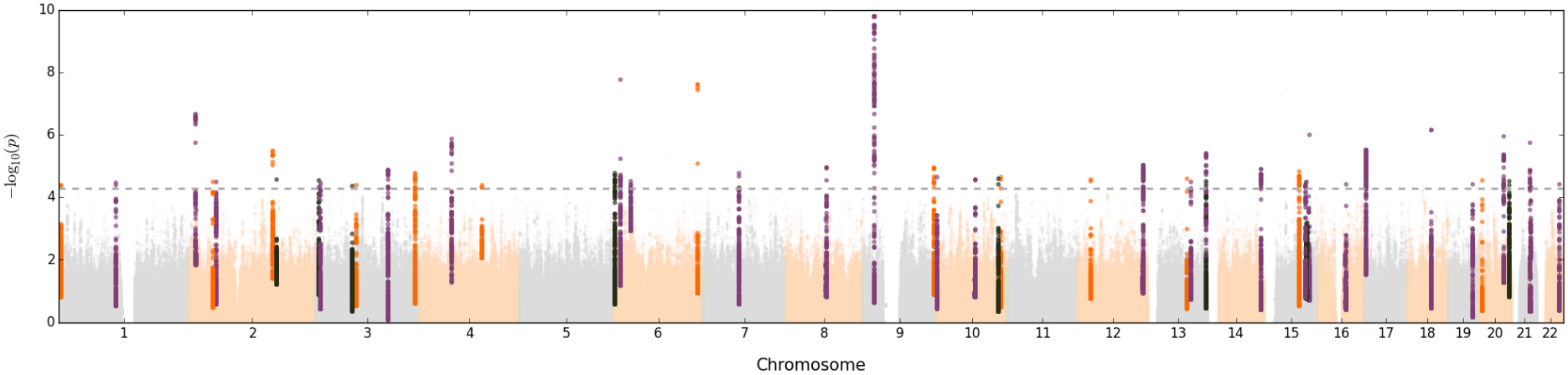
- Computational time (summing over $\mathbf{z}$-states)
- New summary statistics
- Prior assumptions

Association with coronary artery diseases (CAD)

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

Association with coronary artery diseases (CAD)

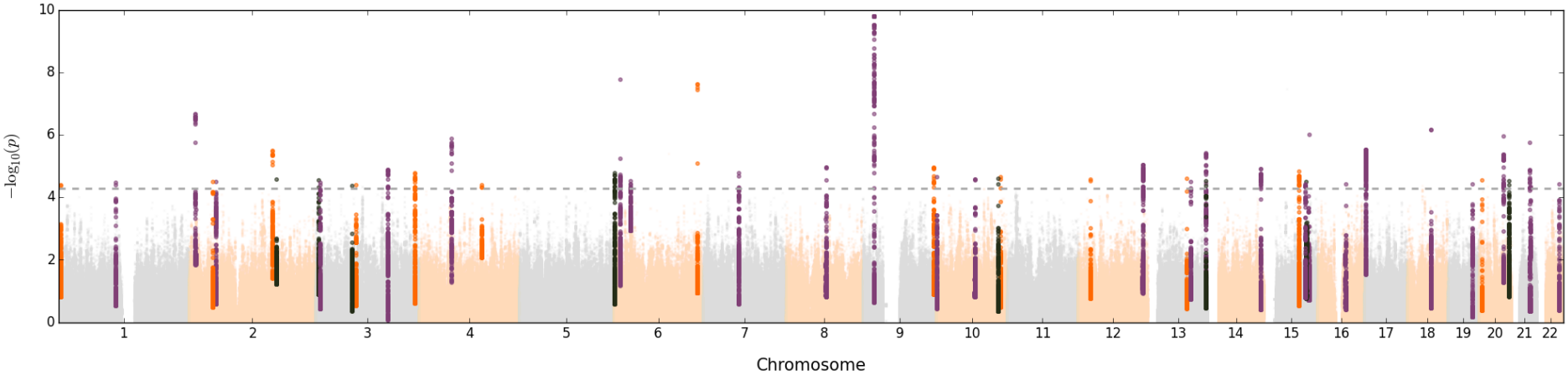
- 5 GERMIFS cohorts
- 6228 cases, 6854 controls
- Imputed with 1000G Phase 1



GWAS using SNPTEST / META

Association with coronary artery diseases (CAD)

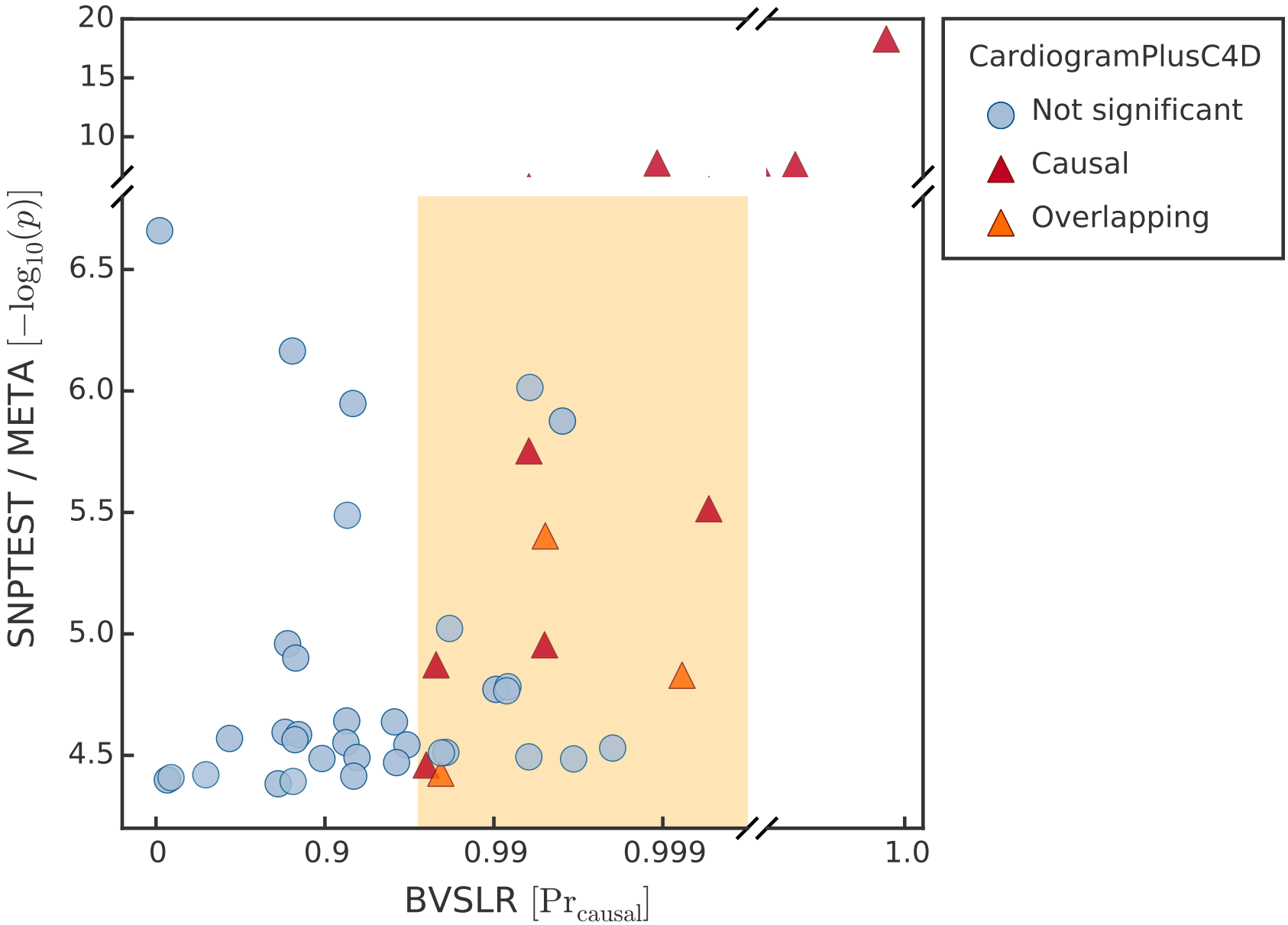
- 5 GERMIFS cohorts
- 6228 cases, 6854 controls
- Imputed with 1000G Phase 1



GWAS using SNPTEST / META

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

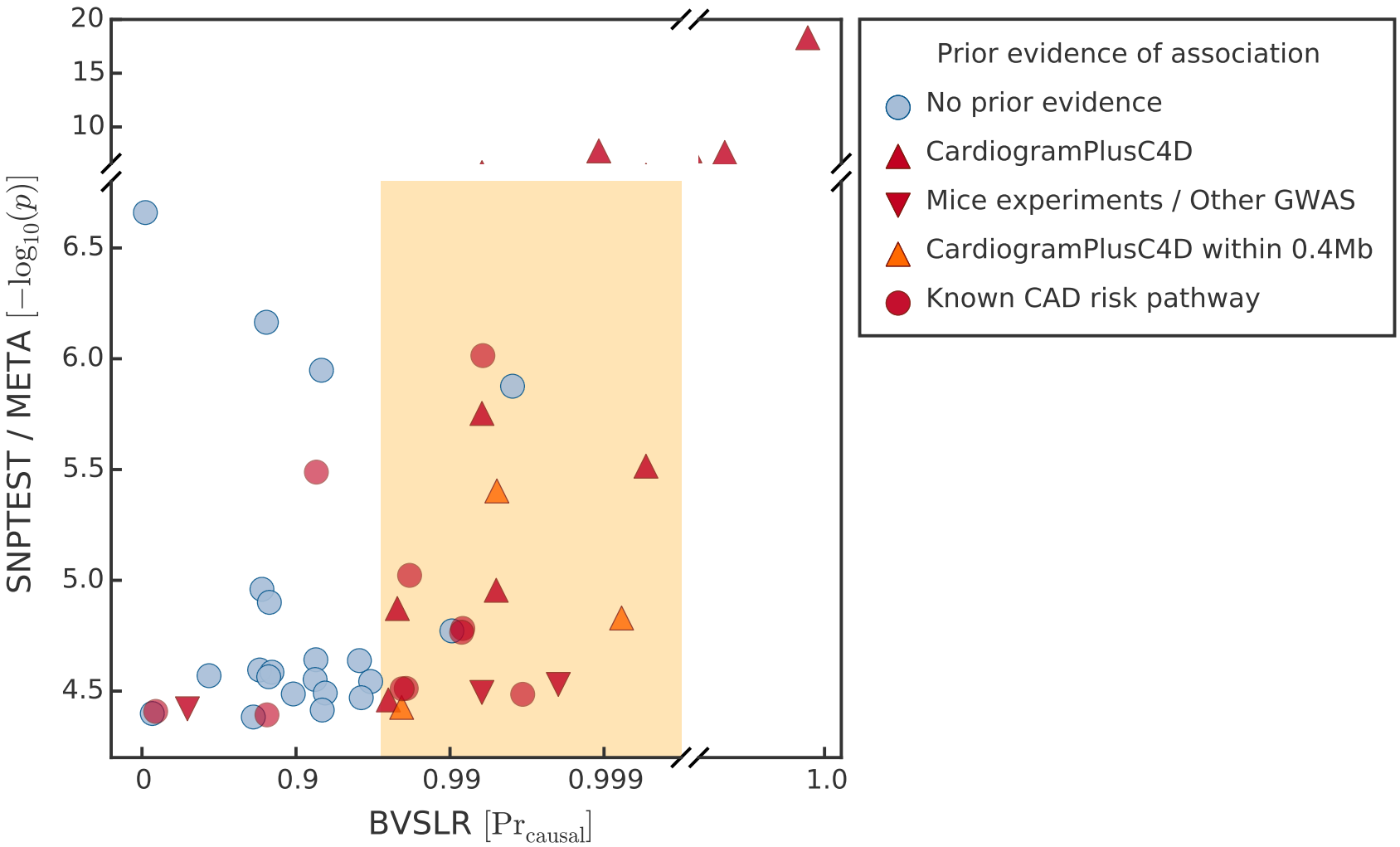
BVSLR predictions



Top BVSLR predicted loci

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

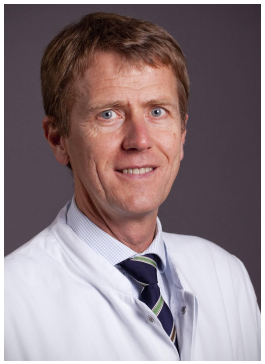


Essence

- 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

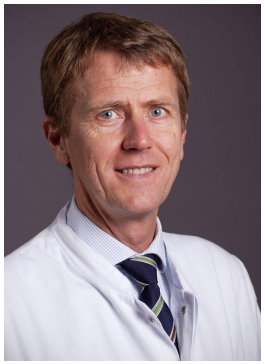
e:Med SYSTEMS MEDICINE

e:AtheroSysMed

Contact me:
saikat.banerjee@mpibpc.mpg.de

Many thanks to . . .


Johannes Söding


Heribert Schunkert


Jeanette Erdmann


AG Söding

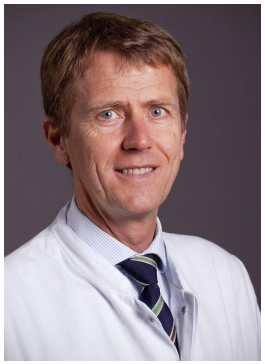
e:Med SYSTEMS MEDICINE

e:AtheroSysMed

Contact me:
saikat.banerjee@mpibpc.mpg.de

Many thanks to . . .


Johannes Söding


Heribert Schunkert


Jeanette Erdmann


AG Söding

Thank you!

e:Med SYSTEMS MEDICINE

e:AtheroSysMed

Contact me:
saikat.banerjee@mpibpc.mpg.de