CARMA: Novel Bayesian model for fine-mapping in meta-analysis studies

Zikun Yang, Ph.D

Department of Biostatistics, Columbia University

Self introduction

- Graduated from the Department of Statistics in Indiana University (Hoosiers!).
- Thesis advisor: Andrew Womack, Ph.D.
- Doctor thesis: Model selection in high-dimensional regime with Bayesian statistics.
- Currently working in the Department of Biostatistics in Columbia University
- PI: Iuliana Ionita-Laza, Ph.D.
- Researches: statistical genetics with applications on genomic data, e.g. GWAS, MPRA etc..

Bayesian statistics

- Proposed a new Bayesian shrinkage prior, Heavy-tailed Horseshoe prior. Comparing to HS, HS+, D-L priors, showed better MSE, better KL risk bounds, better posterior concentration, also the asymptotically minimax risk rate in L_2 norm.
- Showed posterior model selection consistency under the scenario of growing true model with Zellner-Siow and Poisson prior.

Bayesian statistical genetics

- Proposed CARMA fine-mapping method.
- Proposed PO-EN model, which is tailored to the data structure of the massively parallel reporter assays (MPRAs). Using positive and unlabeled/background data together with epigenetic features to build presence-only prediction models of regulatory effects of variants.

Content of today’s talk

Content

- Briefly review the background story of genetics research (GWAS)
- Motivate for the fine-mapping methods
- Challenges of the new method and how we address the challenges
- Simulation and real-data analysis
- **Remark:** More focusing on the features and challenges of genetic data instead of statistical properties or details of the newly proposed model.

Genome-wide association studies (GWAS)

From genome-wide associations to candidate causal variants

- Common complex human traits, quantitative traits (BMI) or diseases (T2D), often result from multiple environmental and genetic causes.
- GWAS have been widely used to identify the genomic regions on chromosomes that harbour genetic determinants of complex traits.
- Many putative loci (genomic regions) of genetic disease has discovered based on GWAS
- The natural next step is to identify putative causal genetic variants, i.e. single-nucleotide polymorphisms (SNPs), at these loci.

GWAS pipeline

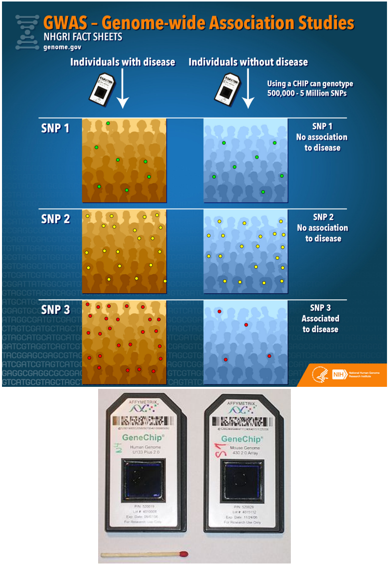

GWAS – Genome-wide Association Studies
NHGRI FACT SHEETS
genome.gov
Individuals with disease
Individuals without disease
Using a CHIP can genotype 500,000 - 5 Million SNPs
SNP 1
SNP 1 No association to disease
SNP 2
SNP 2 No association to disease
SNP 3
SNP 3 Associated to disease
NIH
AFFYMETRIX
GeneChip
Human Genome U133 Plus 2.0
GeneChip
Mouse Genome 430 2.0 Array

Figure: GWAS study and chips.

GWAS

- Recruit subjects.
- Collect trait (binary or quantitative)
- Collect covariates of subjects, e.g. age, gender etc.
- Collect genotypes (imputed) through genotyping techniques (chips)

GWAS pipeline

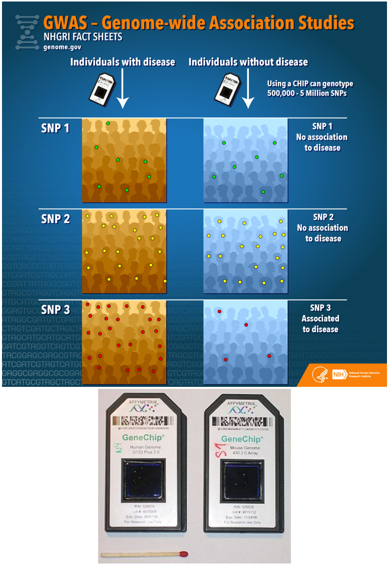

GWAS – Genome-wide Association Studies
NHGRI FACT SHEETS
genome.gov
Individuals with disease
Individuals without disease
Using a CHIP can genotype 500,000 - 5 Million SNPs
SNP 1
SNP 1 No association to disease
SNP 2
SNP 2 No association to disease
SNP 3
SNP 3 Associated to disease
AFFYMETRIX
GeneChip
GeneChip

Figure: GWAS study and chips.

GWAS

- Recruit subjects.
- Collect trait (binary or quantitative)
- Collect covariates of subjects, e.g. age, gender etc.
- Collect genotypes (imputed) through genotyping techniques (chips)

Marginal association of SNPs

- Run linear mixed models or generalized mixed models
- The result of LMM is the marginal association of testing SNP to the complex trait.

GWAS pipeline

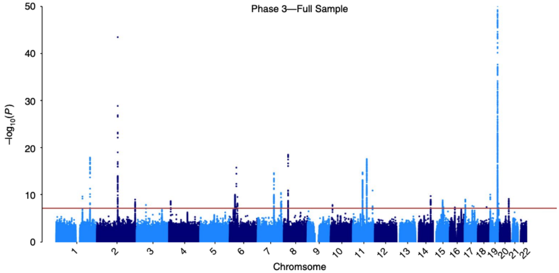


Figure: Manhattan figure of Alzheimer's disease study[Jansen et al., 2019]. The y axis is $-\log_{10}$(P-values), and the commonly-used genome-wide statistical significance threshold of P value is $< 5 \times 10^{-8}$ for a reliable GWAS results.

GWAS

- Collect trait and genotypes (imputed)

Marginal association of SNPs

- Run linear mixed models or generalized mixed models
- Summarize results in Manhattan plot

GWAS pipeline

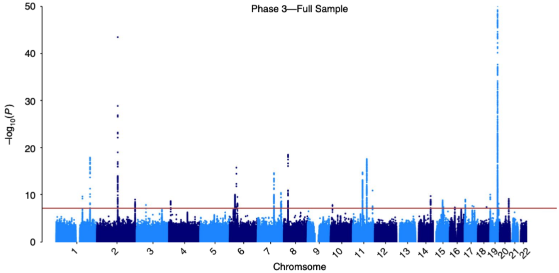


Figure: Manhattan figure of Alzheimer's disease study[Jansen et al., 2019]. The y axis is $-\log_{10}$(P-values), and the commonly-used genome-wide statistical significance threshold of P value is $< 5 \times 10^{-8}$ for a reliable GWAS results.

GWAS

- Collect trait and genotypes (imputed)

Marginal association of SNPs

- Run linear mixed models or generalized mixed models
- Summarize results in Manhattan plot

Investigate on independent genomic region (locus)

- List of associated SNPs
- Explore each independent regions

GWAS pipeline

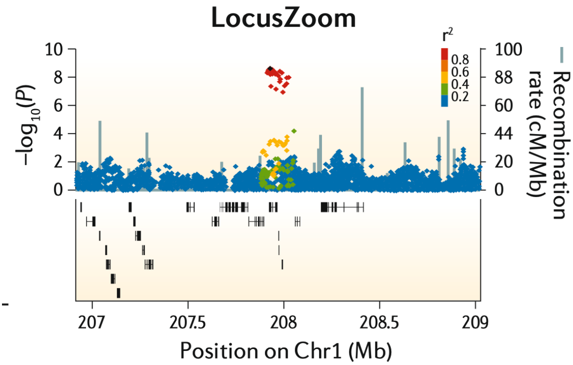


Figure: This figure illustrates the patterns of association of each SNP with the lead SNP, as well as the annotation of genes in the region. Source: [Schaid et al., 2018]

GWAS

- Collect trait and genotypes (imputed)

Marginal association of SNPs

- Run linear mixed models or generalized mixed models
- Summarize results in Manhattan plot

Investigate on independent locus

- List of associated SNPs
- Explore each independent regions

Linkage disequilibrium (LD)

- The leading SNPs are correlated to neighboring SNPs through LD
- Association does not imply causation

GWAS pipeline

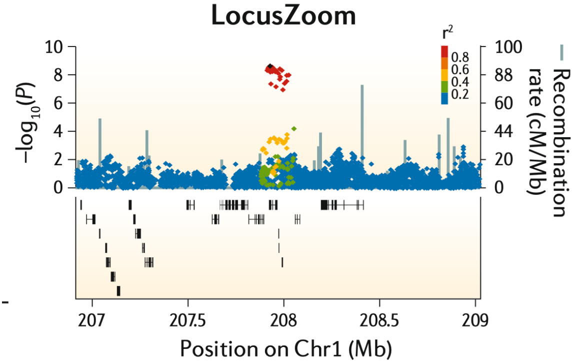


Figure: This figure illustrates the patterns of association of each SNP with the lead SNP. Source: [Schaid et al., 2018]

Goal of Bayesian fine-mapping

Fine-mapping methods utilize the results of the **marginal association test** (between individual genotypes and phenotype) to select and **prioritize genetic variants** accounting for the **complex LD structure** among variants.

GWAS

- Collect trait and genotypes (imputed)

Marginal association of SNPs

- Run linear mixed models or generalized mixed models
- Summarize results in Manhattan plot

Investigate on independent locus

- List of associated SNPs
- Explore each independent regions

Linkage disequilibrium (LD)

- The leading SNPs are correlated to neighboring SNPs through LD
- Association does not imply causation

Data structure of fine-mapping methods

Due to logistic concerns and the availability of meta-analysis, researchers use and share summary statistics and LD matrix instead of directly using phenotype and genotype (large file size).

The marginal association test

- In a given genomic region (Locus), there are p variants and n subjects.
- Let $\boldsymbol{Z}$ denote a p-dimensional vector, where Z_i is the summary statistics of the marginal test between the ith variant, $i = 1, \ldots, p$ and the phenotype $(\boldsymbol{y})$.
- The sampling distribution of $\boldsymbol{Z}$ can be written as:

  $$\boldsymbol{Z}|\boldsymbol{\lambda}, \sigma_y^2, \boldsymbol{\Sigma} \sim \mathsf{MVN}(\boldsymbol{\Sigma}\boldsymbol{\lambda}, \sigma_y^2\boldsymbol{\Sigma}),$$

  where $\boldsymbol{\Sigma}$ is the LD correlation matrix of the variants in the given region.
- We assume $\boldsymbol{\lambda}$ is a sparse vector, and want to identify non-zero entries of $\boldsymbol{\lambda}$ associated with the causal variants.

Challenges of genetic data

Complex LD structure

High and complex correlations among variants (i.e., high linkage disequilibrium (LD)).

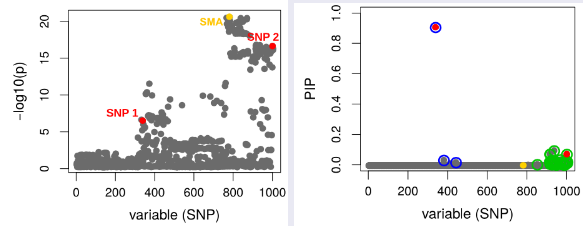


Figure: The SNP (●) is not causal but moderately correlated to the causal SNPs (●). [Wang et al., 2020]

Challenges of genetic data

Complex LD structure

High and complex correlations among variants (i.e., high linkage disequilibrium (LD)).

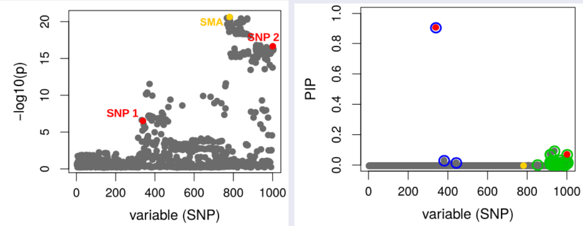


Figure: The SNP (●) is not causal but moderately correlated to the causal SNPs (●). [Wang et al., 2020]

Highly correlated variants

The causal variants could be highly correlated up to tens or even hundreds of other variants with very similar Z-scores. How to distinguish causal SNP from other highly correlated ones.

Challenges of genetic data

Complex LD structure

High and complex correlations among variants (i.e., high linkage disequilibrium (LD)).

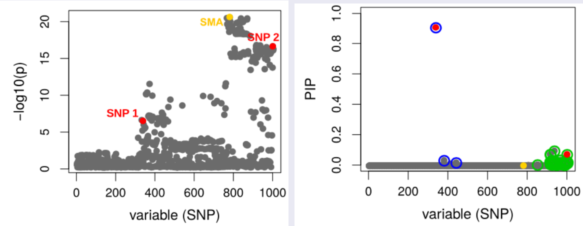


Figure: The SNP (●) is not causal but moderately correlated to the causal SNPs (●). [Wang et al., 2020]

Highly correlated variants

The causal variants could be highly correlated up to tens or even hundreds of other variants with very similar Z-scores. How to distinguish causal SNP from other highly correlated ones.

Mismatch between $\boldsymbol{Z}$/LD due to meta-analysis

- $\boldsymbol{Z}$ To increase power, $\boldsymbol{Z}$ is often generated by the meta-analysis, where the sample size of generating individual Z-score can be dramatically different.
- Σ To avoid transform the in-sample LD matrix of very large file size, Σ is usually extracted from reference panels, e.g., 1000G Genomes.
- This creates inconsistencies between $\boldsymbol{Z}$/LD.

CARMA

Prior on effect size (coefficient)

The sampling distribution of $\boldsymbol{Z}$ is:

$$\boldsymbol{Z}|\boldsymbol{\lambda}, \sigma_y^2, \boldsymbol{\Sigma} \quad \sim \quad \text{MVN}(\boldsymbol{\Sigma}\boldsymbol{\lambda}, \sigma_y^2\boldsymbol{\Sigma}).$$

- Let $\boldsymbol{\gamma}' = \{0,1\}^p$ denote an indicator vector, such that $\gamma_i = 1$ iff $\lambda_i \neq 0$.
- Let S denote an index set such that $i \in S$ if $\gamma_i = 1$.
- $\boldsymbol{\gamma}_S$ and M_S uniquely define a candidate model.
- Given S, the prior distribution of the assumed non-zero effect sizes $\boldsymbol{\lambda}_{\boldsymbol{\gamma}_S}$ that is associated with $\boldsymbol{\gamma}_S$ is

$$\begin{aligned}
\boldsymbol{\lambda}_{\boldsymbol{\gamma}_S}|\sigma_y^2, \tau, \boldsymbol{\gamma}_S \quad &\sim \quad \text{MVN}(0, \frac{\sigma_y^2}{\tau}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_S}^{-1}), \\
\tau \quad &\sim \quad \text{Gamma}(\frac{1}{2}, \frac{1}{2}), \\
\sigma_y^2 \quad &\sim \quad \frac{1}{\sigma_y^2}. \\
\lambda_i \quad &= \quad 0 \text{ if } \gamma_i = 0.
\end{aligned}$$

Remark: CARMA is the first model introduce heavy-tailed prior distribution on the effect size (coefficients) in the fine-mapping setting, higher the power and smaller the size of the identified causal variants.

PIP

We are interested in the posterior inclusion probability (PIP), i.e. $\Pr(\gamma_i = 1|\boldsymbol{Z}, \boldsymbol{\Sigma})$.

Dimensional penalization

Model space

- Fine-mapping is intrinsically a model selection problem in an ultra-sparse scenario.
- Require dimensional penalization from the model space to control FDR.
- Surprisingly, it has not been formally addressed by the previous fine-mapping methods, i.e. using the naive prior $\Pr\left(\gamma_i = 1\right) = \frac{1}{p}$.

Dimensional penalization

Model space

- Fine-mapping is intrinsically a model selection problem in an ultra-sparse scenario.
- Require dimensional penalization from the model space to control FDR.
- Surprisingly, it has not been formally addressed by the previous fine-mapping methods, i.e. using the naive prior $\Pr(\gamma_i = 1) = \frac{1}{p}$.

Prior on Model space in CARMA

- We introduce a prior distribution on model space to control the total number of causal SNPs that any candidate model assumes
- For a given model M_S, let $|S| = \sum_{\gamma_i \in \boldsymbol{\gamma}_S} \gamma_i$ denote the total number of causal SNPs for a given $\boldsymbol{\gamma}_S$ (dimension of M_S), we assign

  $$|S||\eta \sim \text{Truncated Poisson}(\eta), \text{ for } |S| \in \{0, \ldots, p\},$$

  which is first proposed in [Womack et al., 2015].
- Then, for a specific model $\boldsymbol{\gamma}_S$ or M_S, the prior probability of $\boldsymbol{\gamma}_S$ is

  $$\Pr(\boldsymbol{\gamma}_S|\eta) = \frac{\Pr(|S||\eta)}{\binom{p}{|S|}}.$$

- Another goal is to identify the true model ($\boldsymbol{\gamma}_T$ or M_T) that generated the summary statistics, through posterior inference within a Bayesian paradigm. The model selection consistency is shown in [Castillo et al., 2012, Womack et al., 2015].

Computation of PIP

- Let $\mathcal{M}$ denote the model set that contains all candidate models.
- Then
  - the posterior probability of any non-null model $\boldsymbol{\gamma}_S$
  - the posterior probability of γ_i being equal to 1 (PIP)

  can be computed as

$$\Pr\left(\boldsymbol{\gamma}_S|\boldsymbol{Z}\right) = \frac{PO_{\boldsymbol{\gamma}_S:\boldsymbol{\gamma}_0}}{\sum_{\boldsymbol{\gamma}_A\in\mathcal{M}} PO_{\boldsymbol{\gamma}_A:\boldsymbol{\gamma}_0}},$$

$$\Pr\left(\gamma_i = 1|\boldsymbol{Z}\right) = \sum_{\boldsymbol{\gamma}_S:i\in S} \Pr\left(\boldsymbol{\gamma}_S|\boldsymbol{Z}\right),$$

where the posterior odds $(PO_{\boldsymbol{\gamma}_S:\boldsymbol{\gamma}_0})$ is defined as the product of the Bayes factor $\left(\frac{f(\boldsymbol{Z}|\boldsymbol{\gamma}_S)}{f(\boldsymbol{Z}|\boldsymbol{\gamma}_0)}\right)$ and the prior odds $\left(\frac{\Pr(\boldsymbol{\gamma}_S|\eta)}{\Pr(\boldsymbol{\gamma}_0|\eta)}\right)$:

$$\begin{aligned} PO_{\boldsymbol{\gamma}_S:\boldsymbol{\gamma}_0} &= \frac{f(\boldsymbol{Z}|\boldsymbol{\gamma}_S)}{f(\boldsymbol{Z}|\boldsymbol{\gamma}_0)} \frac{\Pr\left(\boldsymbol{\gamma}_S|\eta\right)}{\Pr\left(\boldsymbol{\gamma}_0|\eta\right)} \\ &= \frac{\eta^{|S|}(p-|S|)!}{p!} \int_0^\infty \left[1 - \frac{\boldsymbol{Z}_S'\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_S}^{-1}\boldsymbol{Z}_S}{\boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}\left(1+\tau\right)}\right]^{-\frac{p}{2}} \left(\frac{1+\tau}{\tau}\right)^{-\frac{|S|}{2}} f(\tau)\mathrm{d}\tau. \end{aligned}$$

Computation algorithm

Drawbacks of previous methods

- There are 2^p candidate models. Impossible to going over all models.
- Exhaustively screening $\Pr(\boldsymbol{Z}|\Sigma, \boldsymbol{\gamma})$ for all $\{\boldsymbol{\gamma}; \sum \boldsymbol{\gamma} < \#\}$ with a restriction on $\#$ (e.g.$\# = 2$) is very slow and restrictive.
- Stochastic search (MCMC) is also slow and requiring restriction on $\sum \boldsymbol{\gamma}$, and exploring posterior model space unevenly for highly correlated variants.

Shotgun algorithm [?]

- Given a specific model denoted by an index set S, say $S = \{1, 2, 3\}$, the Shotgun algorithm is an iterative procedure that exhaustively examines the neighborhood of the current model, defined as:

$$\Gamma_{-}(S) := \{A : A \subset S, |S| - |A| = 1\} \text{ (one less variable than } S\text{)},$$

$$\Gamma_{+}(S) := \{A : A \supset A, |A| - |S| = 1\} \text{ (one more variable than } S\text{)},$$

$$\Gamma_{\Leftrightarrow}(S) := \{A : |S| - |A \cap S| = 1, |A| = |S|\} \text{ (models that replaces one variable in } S\text{)}.$$

- To update the current model, the algorithm selects one candidate model from $\Gamma_{-}(S)$, $\Gamma_{+}(S)$, and $\Gamma_{\Leftrightarrow}(S)$ according to the corresponding posterior probabilities, i.e. $\Pr(M_A|\boldsymbol{Z})$.

Advantages of Shotgun

- Semi-exhaustive searching, evenly exploring the group of highly correlated variables.
- Stochastically moves towards the high posterior area in the model space.

Incorporating functional annotations

Motivation

- There are abundant information of the causality of SNPs from the external resource, i.e. functional annotations (gene expression (eQTL)).
- We want to use annotations to distinguish causal variants from highly correlated non-causal variants.

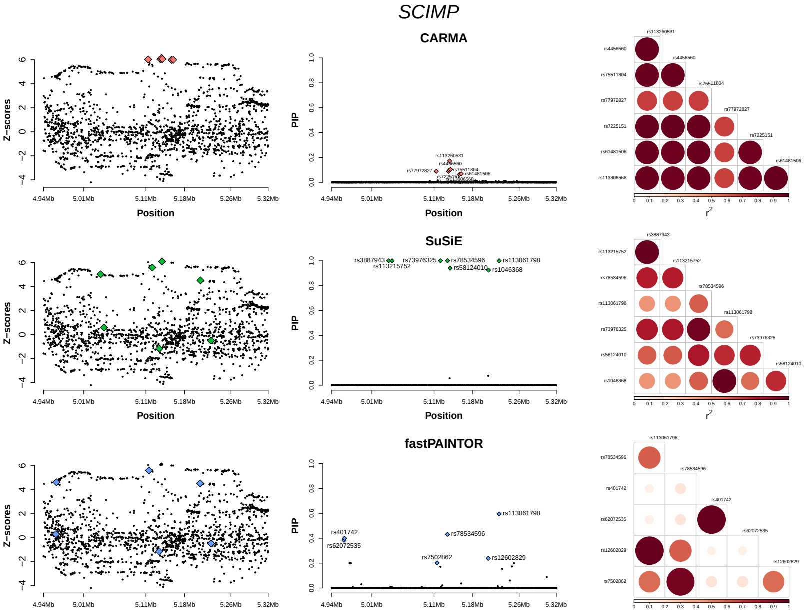
Strategy of EM-algorithm

- By adding the functional annotations, the likelihood can be written as:

  $$L(\boldsymbol{\theta}; \boldsymbol{Z}, \boldsymbol{\gamma}, W) = f(\boldsymbol{Z}|\boldsymbol{\gamma}) \Pr(\boldsymbol{\gamma}|W, \boldsymbol{\theta}),$$

  where W is the matrix of annotations and $\boldsymbol{\theta}$ is the corresponding coefficients.
- Using EM algorithm to maximize likelihood, which associates $\boldsymbol{\gamma}$ and W with a Poisson regression.
- **Feature selection on functional annotations.** By introducing the elastic net penalty, CARMA can perform variable selection on the potentially high-dimensional functional annotation data.
- **Multiplicity control.** CARMA associates the Poisson regression to the Poisson prior on model space.

Outliers motivation



Outliers in fine-mapping

Outlier detection

- Let $\tilde{\boldsymbol{Z}} = \left\{Z_1, \ldots, Z_{|D|}\right\}'$ denote a group of highly correlated SNPs with cor $(Z_i, Z_j) \geq r_{\text{outlier}}$ $(r_{\text{outlier}} > 0.9)$, for $\forall i, j \in D$
- We can assume that the random variable vector $\tilde{\boldsymbol{Z}}$ follow a MVN$\left(\mathbf{1}_{|D|}\beta, \boldsymbol{\Sigma}_D\right)$.
- In a loop algorithm, we test whether Z_i is generated by a different distribution, i.e. $N(\beta, c)$:

$$H_0 : c = 1;\ Z_i \text{ is not an outlier}$$

$$H_1 : c \neq 1;\ Z_i \text{ is an outlier.}$$

- We assume $\tilde{\boldsymbol{Z}}_{-i}$ are not outliers and follow a MVN$\left(\mathbf{1}_{|D|-i}\beta, \boldsymbol{\Sigma}_{D_{-i}}\right)$.
- Computing the Bayes factor between the two hypothesis and dropping the variants if reject H_0.

Credible set and credible model

Credible set

- In [Wang et al., 2020] the authors define a credible set, and can be simplified as

Definition

Given $\rho = 0.99$ and a correlation threshold r, S is a credible set of variants if

- $\sum_{i \in S} \Pr\left(\gamma_i | \boldsymbol{Z}\right) \geq \rho$
- min $\left\{cor(i,j) \geq r\right\}$, for all $i, j \in S$
- $|S|$ is minimal.

- Credible sets can identify groups of highly correlated variants for further experimental validations.

Credible Model

- Instead of credible set, we propose the concept of credible model.
- Let $\boldsymbol{\gamma}_{(b)}$, $b = 1, \ldots, B$, denote the ranked candidate models, such as $\boldsymbol{\gamma}_{(1)}$ receives the largest marginal likelihood.
- We use $\boldsymbol{\gamma}_{(1)}$ as the reference model to select all other candidate models.
- Including any candidate models into the credible model such as

$$\text{PO}_{\boldsymbol{\gamma}_{(1)}:\boldsymbol{\gamma}_{(b)}} = \frac{\Pr\left(\boldsymbol{\gamma}_{(1)} | \boldsymbol{Z}, \eta\right)}{\Pr\left(\boldsymbol{\gamma}_{(b)} | \boldsymbol{Z}, \eta\right)} = \frac{\Pr\left(\boldsymbol{Z} | \boldsymbol{\gamma}_{(1)}\right) \Pr\left(\boldsymbol{\gamma}_{(1)} | \eta\right)}{\Pr\left(\boldsymbol{Z} | \boldsymbol{\gamma}_{(b)}\right) \Pr\left(\boldsymbol{\gamma}_{(b)} | \eta\right)} < 10.$$

Simulating genotype based on real data

Simulation settings

- We use the R package 'sim1000G' [Dimitromanolakis et al., 2019] to simulate genotypes based on the 1000 Genomes Project data.
- We focus on 94 loci identified as risk regions in a recent GWAS on breast cancer [Fachal et al., 2020].
- The number of variants in each region ranges between $\sim 1,500 - 4,000$.
- We simulate genotype data for $n = 10,000$ individuals.

Prior probability

- We use functional annotations (randomly select 200 chromatin features out of 919 DeepSEA chromatin features [Zhou and Troyanskaya, 2015]) to determine the causalities of variants, such as $\Pr\left(\gamma_i = 1 | \boldsymbol{\theta}, \boldsymbol{w}_i\right) = \frac{\exp\left\{\boldsymbol{w}_i' \boldsymbol{\theta}\right\}}{1+\exp\left\{\boldsymbol{w}_i' \boldsymbol{\theta}\right\}}$.
- Let T denote the index set of the true causals selected, and $|T| = 3$.

Phenotype and summary statistics

- For each $i \in T$, $\gamma_i = 1$ and $\beta_i \sim N(0, 0.5^2)$.
- The phenotypic variance σ_y^2 is computed such that $\phi = 0.0075$, where $\phi = \frac{\text{Var}(\boldsymbol{X}\boldsymbol{\beta})}{\sigma_y^2 + \text{Var}(\boldsymbol{X}\boldsymbol{\beta})}$.
- Then we sample $\boldsymbol{y}$ such that $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$; $\boldsymbol{\epsilon} \sim$ N(0, $\sigma_y^2 I_{n \times n}$).
- Compute $\boldsymbol{Z}$ and $\boldsymbol{\Sigma}$.
- Two other very popular testing fine-mapping models, SuSiE[Wang et al., 2020] and fastPAINTOR[Kichaev et al., 2017].
- We assume two scenarios (1) no functional annotation and (2) with functional annotations (919 DeepSEA chromatin features).

Simulation results (no outlier)

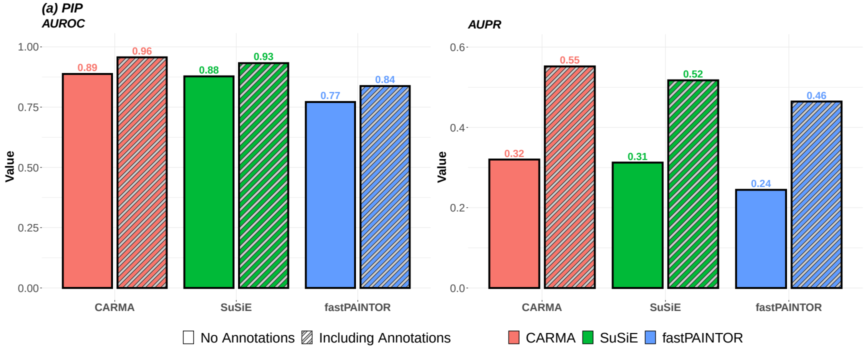


Figure: AUROC and AUPR of the testing models

Simulation results (no outlier)

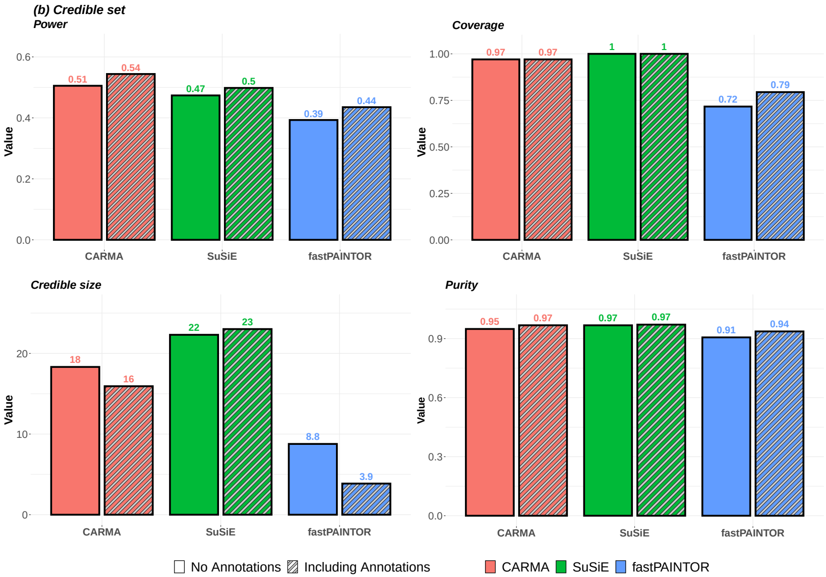


Figure: Performances associated with the credible set.

Simulation results (no outlier)

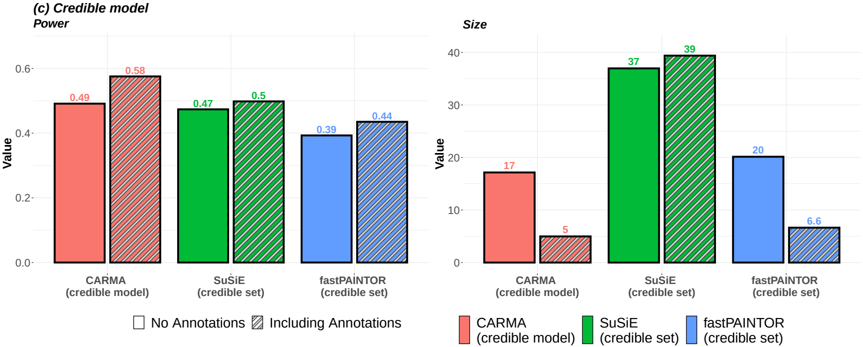
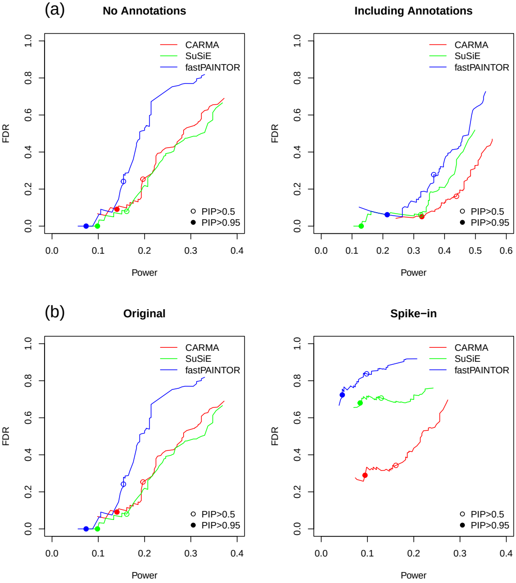


Figure: Credible models and counterparts of SuSiE and fastPAINTOR.

Simulation results (with outlier)



Real data (AD study [Jansen et al., 2019])

Data process

- We present fine-mapping results at 30 GWAS loci identified in a large meta-analysis of clinically diagnosed AD and AD-by-proxy with 71,880 cases and 383,378 controls of European ancestry [Jansen et al., 2019].
- For the CARMA model, we include 924 functional annotations including DeepSEA [Zhou and Troyanskaya, 2015], CADD [Kircher et al., 2014], PO-EN [Yang et al., 2021], and PolyFun [Weissbrod et al., 2020].
- For each model, we consider two scenarios:
    1. no functional annotation
    2. including functional annotations

Challenges

- The sample sizes can vary from 9,703 to 444,006 depending on which datasets are included in the meta-analyses.
- The LD matrix is extracted from AD-by-proxy dataset (UK Biobank).
- Severe discrepancies between Z/LD.

Real data (AD study [Jansen et al., 2019])

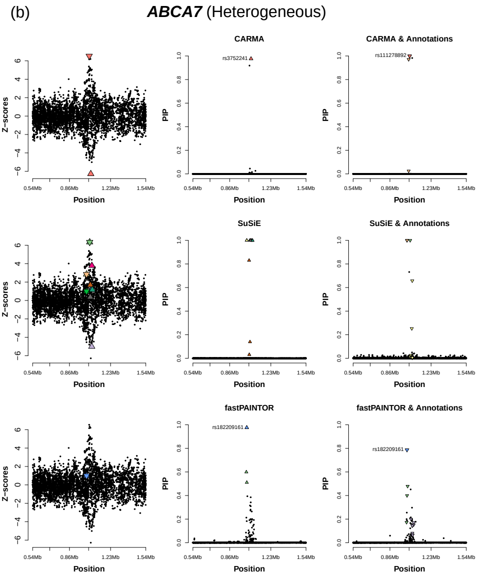


Figure: Locus*ABCA7*.

Real data (AD study [Jansen et al., 2019])

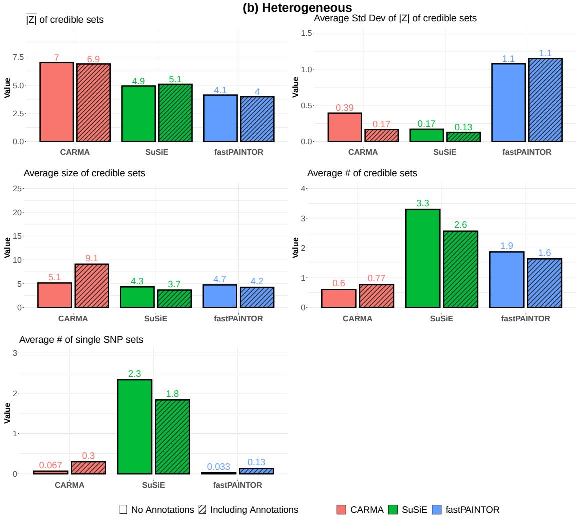


Figure: Summary of the all 30 loci.

Future researches

CARMA

- The paper has been submitted to NATURE Genetics, under review now.
- The R package is on GitHub with user manual.
- The authors are Zikun Yang, Chen Wang, Atlas Khan, Badri Vardarajan, Richard Mayeux, Krzysztof Kiryluk, Iuliana Ionita-Laza.

Statistical genetics

- Currently, I am working on a better solution of the outlier detection with extra information of the meta-analysis.
- Working on developing multi-ethnics fine-mapping method, i.e. combining European, African, East Asian, etc.. Considering structure of variational Bayes that I am totally not familiar with :).
- Working on a real data of Alzheimer's disease based on the subjects from Dominican, Mexican, Peru, and ethnics associated with Caribbean are. Largest datasets of such cohorts to date.

Bayesian Statistic

- Working on the Heavy-tailed Horseshoe prior, finishing paper.
- Trying to replace LMM model in genetics with Horseshoe prior.

Special thank

Special thanks to

- Dr. Ionita-Laza
- Dr. Womack

Thank you all!

THANK YOU!

Castillo, I., van der Vaart, A., et al. (2012).
Needles and straw in a haystack: Posterior concentration for possibly sparse sequences.
The Annals of Statistics, 40(4):2069–2101.

Dimitromanolakis, A., Xu, J., Krol, A., and Briollais, L. (2019).
sim1000g: a user-friendly genetic variant simulator in r for unrelated individuals and family-based designs.
BMC bioinformatics, 20(1):26.

Fachal, L., Aschard, H., Beesley, J., Barnes, D. R., Allen, J., Kar, S., Pooley, K. A., Dennis, J., Michailidou, K., Turman, C., et al. (2020).
Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes.
Nature genetics, 52(1):56–73.

Jansen, I. E., Savage, J. E., Watanabe, K., Bryois, J., Williams, D. M., Steinberg, S., Sealock, J., Karlsson, I. K., Hägg, S., Athanasiu, L., et al. (2019).
Genome-wide meta-analysis identifies new loci and functional pathways influencing alzheimer's disease risk.
Nature genetics, 51(3):404–413.

Kichaev, G., Roytman, M., Johnson, R., Eskin, E., Lindstroem, S., Kraft, P., and Pasaniuc, B. (2017).
Improved methods for multi-trait fine mapping of pleiotropic risk loci.
Bioinformatics, 33(2):248–255.

Kircher, M., Witten, D. M., Jain, P., O'roak, B. J., Cooper, G. M., and Shendure, J. (2014).
A general framework for estimating the relative pathogenicity of human genetic variants.
Nature genetics, 46(3):310–315.

Schaid, D. J., Chen, W., and Larson, N. B. (2018).

EM-algorithm

Details of EM-algorithm

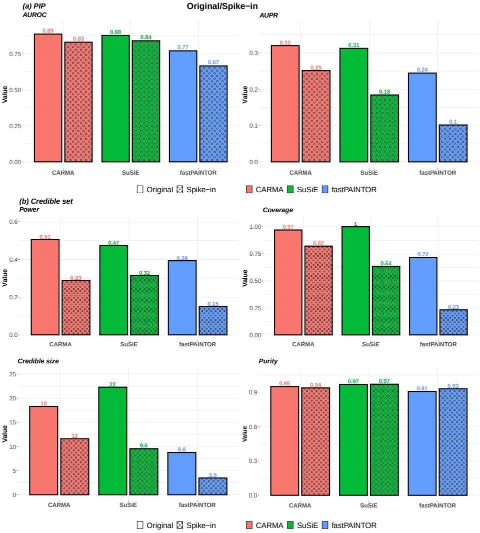
- Suppose that the truncated model space Γ of the top B models with the largest posterior probabilities.
- Let $\boldsymbol{G}' = \{G_1, \ldots, G_p\}$ denote the count vector associated with Γ, where $G_i \in \{0, 1, \ldots, B\}$ is the count of $\gamma_i = 1$ appearing in Γ.
- $\boldsymbol{G}$ is the missing value. In EM algorithm, we use Poisson regression to model it.
- Let $g_i^{(s)}$ denote the actual count of γ_i appearing in $\Gamma^{(s)}$ after running Shotgun algorithm at step (s) of the EM algorithm.
- We approximate $\mathbf{E}\left[G_i | \boldsymbol{Z}, \boldsymbol{w}_i, \boldsymbol{\theta}^{(s)}\right]$ by $g_i^{(s)}$ in EM algorithm.

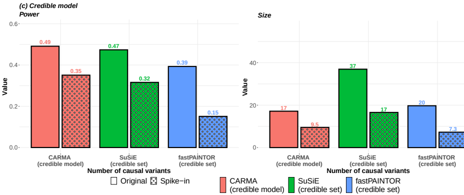
EM-algorithm

```
Input: Summary statistics $\boldsymbol{Z}$, functional annotations $W$, hyperparameter $\eta$ of the Poisson prior distribution, and $B$.
Initialization: Run Shotgun algorithm with the prior distribution Poisson$(\eta)$ to generate $\Gamma^{(0)}$ and $\boldsymbol{g}^{(0)}$.
for $s = 0, 1, \ldots$ do
        - EM
        E-step  Replace $G_i$ by $\mathbf{E}\left[G_i|\boldsymbol{Z}, \boldsymbol{w}_i, \boldsymbol{\theta}^{(s)}\right]$, which is approximated by $g_i^{(s)}$, $i = 1, \ldots, p$.
        M-step  Maximize the penalized log-likelihood as,
                $$\boldsymbol{\theta}^{(s+1)} := \underset{\boldsymbol{\theta} \in R^{q+1}}{\text{argmax}} \sum_{i=1}^{p} \left[g_i^{(s)} \boldsymbol{w}_i' \boldsymbol{\theta} - \exp\left\{\boldsymbol{w}_i' \boldsymbol{\theta}\right\}\right] - \frac{(1-\alpha)}{2} ||\boldsymbol{\theta}||^2 - \alpha ||\boldsymbol{\theta}||.$$
        Adjust the prior probability to introduce the multiplicity control (see details below),
                $$\hat{\theta}_1^{(s+1)} = \log\left(\frac{\eta B^{(s)}}{\eta + p}\right). \quad (1$$
        Then, compute the prior probability of the $(s+1)$ step:
                $$\hat{\Pr}\left(\gamma_i = 1|\boldsymbol{w}_i, \boldsymbol{\theta}^{(s+1)}\right) = \frac{\exp\left\{\boldsymbol{w}_i' \boldsymbol{\theta}^{(s+1)}\right\}}{B^{(s)}}, \quad (2$$
        where $B^{(s)}$ is the minimum between $B$ and the total number of models visited by the Shotgun algorithm in step $(s)$.
        - Shotgun
        Initiate Shotgun algorithm with the estimated prior probability vector $\left\{\hat{\Pr}(\gamma_1), \ldots, \hat{\Pr}(\gamma_p)\right\}'$. After running Shotgun algorithm, acquire $\Gamma^{(s+1)}$ and $\boldsymbol{g}^{(s+1)}$, which depends on $\boldsymbol{Z}$, $W$, and $\boldsymbol{\theta}^{(s+1)}$.
end
```

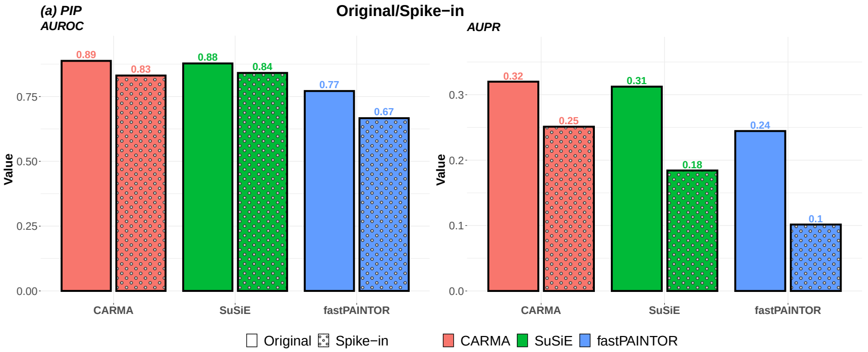
Algorithm 1: EM algorithm with functional annotations.

Simulation results (with outlier)





Simulation results (with outlier)



Outlier algorithm

At any step of the Shotgun algorithm, suppose that the current model is γ_S.

Input: The index set $S = \left\{s_1, \ldots, s_{|S|}\right\}$ for the current model γ_S, the threshold on the Bayes factor δ, and the threshold on the correlation ρ_{outlier}.

```
for $s = 1, \ldots, |S|$ do
    - Given $s_s \in S$, identify the group of highly correlated SNPs indicated by the index set
      $D = \{i;\ \text{cor}\left(Z_{s_s}, Z_i\right) \geq r_{\text{outlier}} \text{ for } \forall i \in \{1, \ldots, p\}\}$.
    - Define $\tilde{\boldsymbol{Z}} = \left\{Z_1, \ldots, Z_{|D|}\right\}'$ as the summary statistics vector of the set $D$ .
    repeat
        for $d = 1, \ldots, |D|$ do
            - Define the hypothesis test for $Z_d$, such that
                  $H_0 : Z_d \sim N(\beta, 1);$              $Z_d$ is not an outlier
                  $H_1 : Z_d \sim N(\beta, c),\ c \neq 1;$   $Z_d$ is an outlier.
            - Compute the corresponding Bayes factor $\hat{B}_d$ conditional on $\tilde{\boldsymbol{Z}}_{D_{-d}}$ and $\boldsymbol{\Sigma}_{D_{-d}}$.
            =
        end
        if $\exists d \in \{1, \ldots, |D|\}, \hat{B}_d < \delta$ then
            - Drop $Z_d$, where $\hat{B}_d = \min\left(\left\{\hat{B}_1, \ldots, \hat{B}_{|D|}\right\}\right)$, from the fine-mapping computation.
            - Drop $Z_d$ from $\tilde{\boldsymbol{Z}}$, i.e., $\tilde{\boldsymbol{Z}} = \left\{Z_1, \ldots, Z_{d-1}, Z_{d+1}, \ldots, Z_{|D|}\right\}'$.
            - Drop $d$th index from the index set $D$.
    until $\hat{B}_d \geq \delta$, for $\forall d \in \{1, \ldots, |D|\}$;
end
```

Algorithm 2: The outlier detection procedure implemented in Shotgun algorithm.