A scalable method for genetic fine-mapping using summary statistics

Yuxin Zou[1], **Gao Wang**[2], **Peter Carbonetto**[2] **and Matthew Stephens**[1,2] [1]**Department of Statistics and** [2]**Department of Human Genetics, The University of Chicago, Chicago, IL 60637, USA**

Why fine-mapping

- **Genome-wide association studies** have successfully identified many genomic regions associated with complex diseases and traits.
- **Fine-mapping**
  - ▷ Pinpoint the causal variants contributing to diseases and traits.
- **Bayesian variable selection methods**
  - ▷ Designed to quantify uncertainty of genetic variables.
  - ▷ Takes into account Linkage Disequilibrium (LD).

Motivation

Challenges in fine-mapping:

- Most Bayesian variable selection methods are computationally intensive.
- Cannot directly infer at per-variable level resolution.

***SuSiE* [1]** provides a solution to the challenges above.

- It is computationally efficient, $O(npL)$
- It provides a simpler way to summarize fine-mapping results using "credible sets".

BUT it requires individual-level genotype and phenotype data. We present **SUm of SIngle Effects Regression with Summary Statistics (*SuSiE-RSS*)**.

SuSiE-RSS

The model is

$$\hat{\boldsymbol{z}} \sim N_p(\boldsymbol{R}\boldsymbol{z}, \sigma^2\boldsymbol{R}) \tag{1}$$

$$\boldsymbol{z} = \sum_{l=1}^{L} \boldsymbol{z}_l \tag{2}$$

$$\boldsymbol{z}_l = \boldsymbol{\gamma}_l z_l \tag{3}$$

$$\boldsymbol{\gamma}_l \sim \text{Multinom}(1, \boldsymbol{\pi}) \tag{4}$$

$$z_l \sim N(0, \sigma_{0l}^2) \tag{5}$$

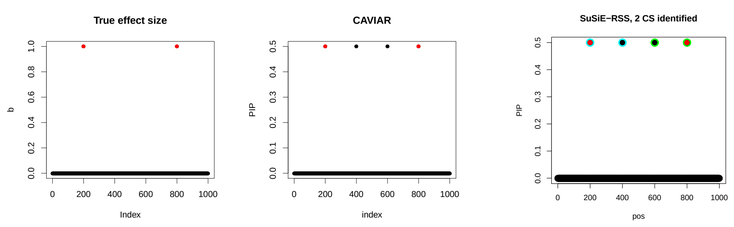
$$\sigma^2 \leq 1 \tag{6}$$

- z scores are $\hat{z}_j = \hat{b}_j/\hat{s}_j$, where $\hat{b}_j = \frac{\boldsymbol{x}^\intercal \boldsymbol{y}}{\boldsymbol{x}^\intercal \boldsymbol{x}}$, $\hat{s}_j = \frac{\hat{\sigma^2}_j}{\boldsymbol{x}^\intercal \boldsymbol{x}}$
- $\boldsymbol{R}$ is the LD matrix.
- $\boldsymbol{R}$ should be the sample correlation matrix from the original individual-level genotype data, $\boldsymbol{X}$
- Misspecification of LD can lead to unreliable inferences
- LD correction:
  - ▷ Suppose $\boldsymbol{X}_{\text{out}}$ is the $n' \times p$ misspecified genotype data (centered, scaled), we estimate LD as

$$\boldsymbol{R} = \frac{1}{n'}\left(\boldsymbol{X}_{\text{out}}^\intercal \boldsymbol{X}_{\text{out}} + \hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^\intercal\right) \tag{7}$$

Toy example

Compare *SuSiE-RSS* and CAVIAR results:

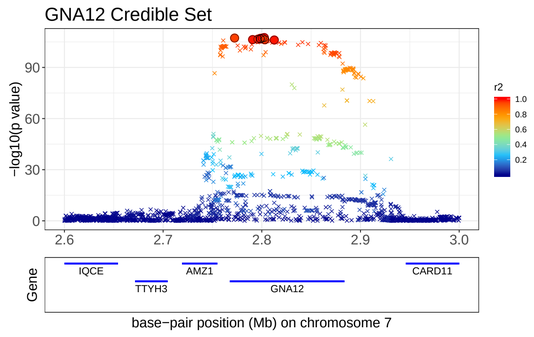


Possible combinations in CAVIAR:

200	400	600	800	Probability
1	0	1	0	0.25
1	0	0	1	0.25
0	1	1	0	0.25
0	1	0	1	0.25

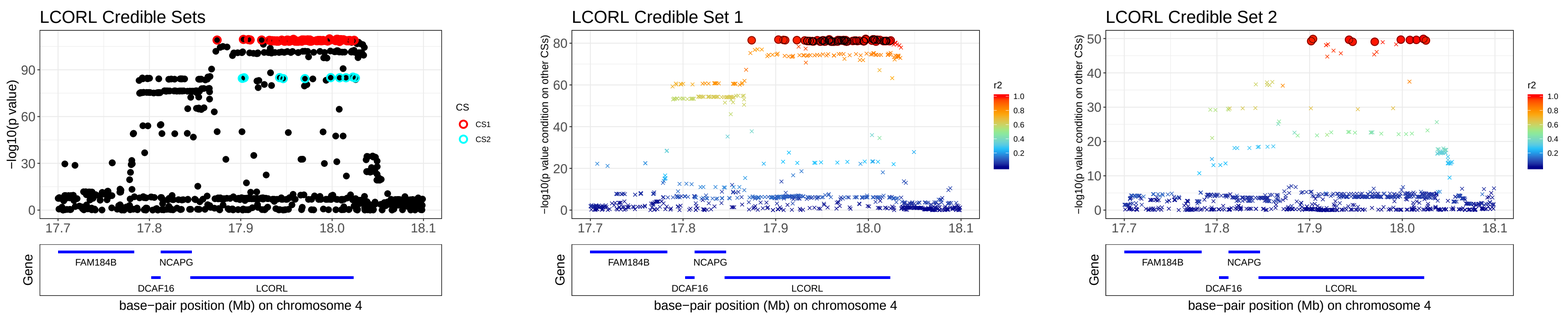
Fine-mapping standing height in UK Biobank data

- *GNA12* – *SuSiE-RSS* estimates 1 causal variant.

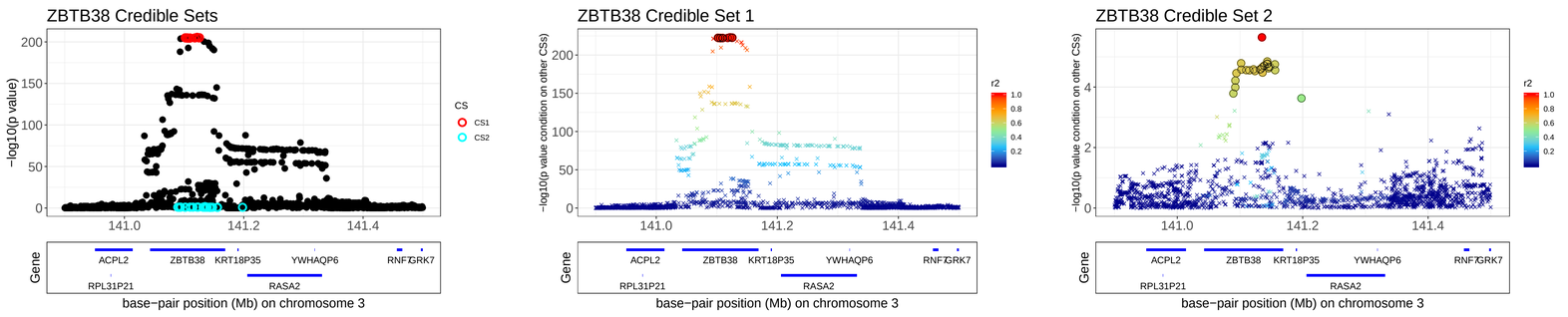


CHR	POS	ID	REF	ALT	MAF(%)	PIP(%)
7	2802522	rs798488	C	T	30.0	35.8
7	2800521	rs798491	G	A	30.0	20.0
7	2798294	rs798494	A	C	30.0	12.4
7	2772431	rs798528	C	A	30.8	11.7
7	2798731	rs798493	G	A	30.0	5.7
7	2795957	rs798497	G	A	30.0	4.3
7	2790685	rs798500	C	T	30.0	2.7
7	2803037	rs798486	G	A	29.5	1.6
7	2812632	rs35957220	G	C	30.0	1.6
7	2802943	rs798487	A	G	29.4	1.6

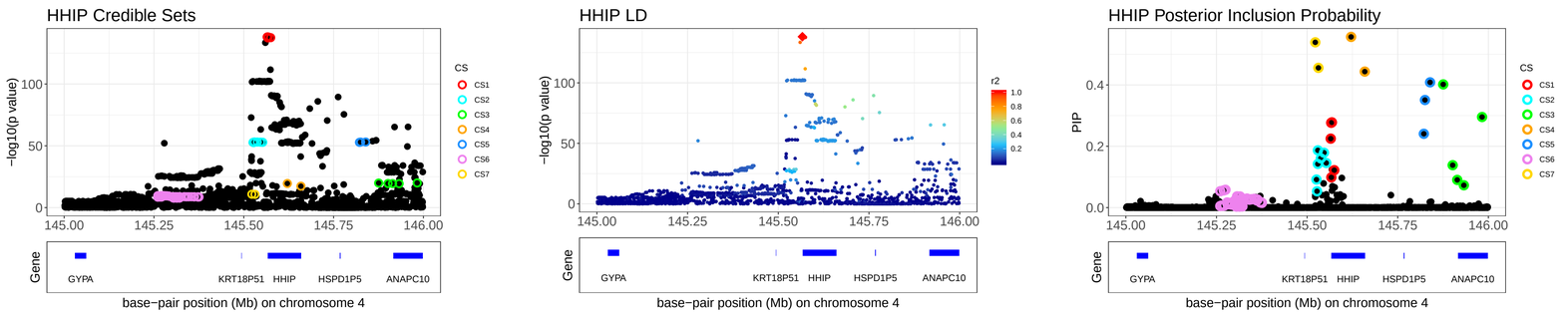
- *LCORL* – *SuSiE-RSS* estimates 2 causal variants.



- *ZBTB38* – *SuSiE-RSS* estimates 2 causal variants.



- *HHIP* – *SuSiE-RSS* estimates 7 causal variants.



Numerical studies of fine-mapping with CAVIAR and FINEMAP

The simulation is based on genotype data from the Genotype-Tissue Expression (GTEx) project. The individuals are randomly separated into 2 groups.

- ▷ The first group is used to compute the summary statistics, $\hat{\boldsymbol{z}}$ and $\boldsymbol{R}$.
- ▷ The second group is treated as $\boldsymbol{X}_{\text{out}} \rightarrow$ misspecified LD.

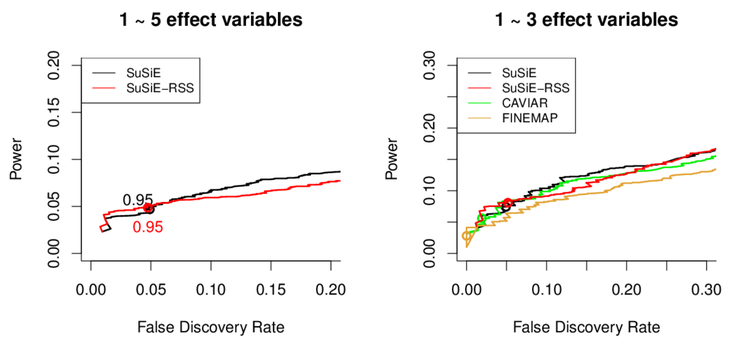
We compare our method with SuSiE, CAVIAR version 2.2 and FINEMAP version 1.1.

Computation speed (unit: second)

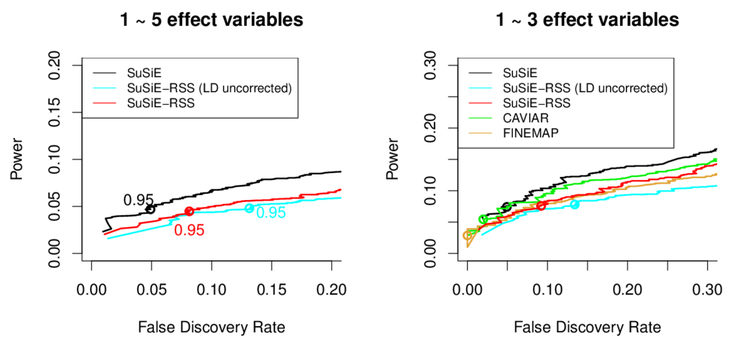
Method	Avg.	Min.	Max.
SuSiE-RSS †	2.82	0.37	13.81
FINEMAP	27.23	14.34	54.93
CAVIAR	1587.35	51.34	5043.72

† *SuSiE-RSS* is implemented in R.
Others are implemented in C++.

SNP-level power using correct LD



SNP-level power using misspecified LD with correction



References

[1] Gao Wang, Abhishek K Sarkar, Peter Carbonetto, and Matthew Stephens. A simple new approach to variable selection in regression, with application to genetic fine-mapping. *bioRxiv*, page 501114, 2018.

Yuxin Zou
yuxin@uchicago.edu
Matthew Stephens
mstephens@uchicago.edu

SuSiE R package
https://github.com/stephenslab/susieR