Fine Mapping With TWAS Results Across Multiple Tissues

Shuai Li, Xinyu (Brian) Guo

Johns Hopkins Bloomberg School of Public Health

May 14, 2021

1. Background and Methods
2. Data For Alzheimer's Disease
3. Analysis (SuSiE)
4. Reference

Section 1

Background and Methods

What is TWAS?

- TWAS: transcriptome-wide association study.
- To determine significant trait-expression associations. [1]
- This method increases the power of identifying functionally relevant loci by leveraging expression quantitative trait loci (eQTLs) from external references in relevant tissues. [2]

[1] Gusev et al. "Integrative approaches for large-scale transcriptome-wide association studies" 2016 Nature Genetics

[2] Bhattacharya et al. "A framework for transcriptome-wide association studies in breast cancer in diverse study populations" 2020 Genome Biology

TWAS/FUSION Software

- Functional Summary-based Imputation:

  FUSION is a suite of tools for performing a TWAS by predicting functional/molecular phenotypes into GWAS using only summary statistics (usually from GWAS). The goal is to identify associations between a GWAS phenotype and a functional phenotype that was only measured in reference data. [1]

[1] Gusev et al. "Integrative approaches for large-scale transcriptome-wide association studies" 2016 Nature Genetics

TWAS/FUSION Software

- Inputs:
  - GWAS summary statistics
  - Reference panels (i.e. precomputed functional weights (primarily gene expression) from multiple tissues)
  - Regerence LD data
- Outputs:
  - A data frame with corresponding z and p values for each SNPs.

[1] Gusev et al. "Integrative approaches for large-scale transcriptome-wide association studies" 2016 Nature Genetics

Bayesian Fine Mapping

- Why fine-map?
  - To find causal genes
  - To pinpoint variant
  - To understand genetic architecture
    - Gene enrichment
    - Cross-trait comparison, cross-tissue
- Bayesian fine-mapping outputs:
  - PIP: Posterior inclusion probability (the probability that a variant is causal)
  - 95% Credible Sets: Set of variants that contains $\geq 95\%$ probability

[3] ▸ Hilary Finucane, Broad Institute

Bayesian Fine Mapping

- Single-causal-variant PIPs:

$$\begin{aligned} PIP_j &= P(\text{j causal} \mid \text{data}) \\ &= \frac{P(\text{data} \mid \text{j causal})}{\sum_k P(\text{data} \mid \text{k causal})} \\ &= \frac{P(\text{data} \mid \text{j causal})/P(data|nocausal)}{\sum_k P(\text{data} \mid \text{k causal})/P(data|nocausal)} \\ &= \frac{\text{Bayesian Factor}_j}{\sum_k \text{Bayesian Factor}_k} \end{aligned}$$

[3] ▸ Hilary Finucane, Broad Institute

Bayesian Fine Mapping

- 95% Credible Sets (S):

$$P(\text{causal var is in S}) \geq 0.95$$

- Under Single-causal-variant assumption:

$$P(\text{causal var is in S}) = \sum_{j \in S} PIP_j$$

- To get the most compact credible set, add variant with highest PIPs untill sum to 0.95.

[3] ▸ Hilary Finucane, Broad Institute

Bayesian Fine Mapping

- Factors affecting Bayesian fine mapping power
  - LD
  - Sample Size
  - Effect size

[4] Schaid et al. Nat Rev Genet 2018

Bayesian Fine Mapping

- Multiple-causal-variant Fine-mapping (two approaches):
  - Divide the whole data into many pieces, and apply single-causal-variant fine-mapping in each piece
  - Jointly model Multiple-causal-variant

[3] Hilary Finucane, Broad Institute

Section 2

Data For Alzheimer's Disease

Overview

- Data
  - Gene Expression Matrix: Gene expression level in each tissue
    - Z-values
    - P-values
  - Correlation matrix: Correlation of expression in each tissue for each gene

Gene Expression matrix

```
dim(dat_ad_n[[1]])
```

```
## [1] 33960    49
```

```
dim(dat_ad_n[[2]])
```

```
## [1] 33960    49
```

```
dat_ad_n[[1]][1:5,1:5]
```

```
##       GENE Whole_Blood   Vagina    Uterus   Thyroid
## 1  EXOC3L2    23.14987       NA  9.249285        NA
## 2   CLASRP    12.86142       NA 30.900394 10.092551
## 3 TRAPPC6A    11.31097 1.770764        NA  6.203560
## 4    NKPD1    10.63722       NA        NA  2.276769
## 5 CEACAM19    10.52064       NA  6.806502 10.630916
```

Gene Expression matrix

```
dat_ad_n[[2]][1:5,1:5]
```

```
##       GENE Whole_Blood Vagina    Uterus  Thyroid
## 1  EXOC3L2   7.29e-119     NA  1.13e-20       NA
## 2   CLASRP    3.71e-38     NA 5.90e-210 2.98e-24
## 3 TRAPPC6A    5.79e-30 0.0383        NA 2.76e-10
## 4    NKPD1    1.00e-26     NA        NA 1.14e-02
## 5 CEACAM19    3.47e-26     NA  5.00e-12 1.07e-26
```

```
1 - pnorm(dat_ad_n[[1]][3,"Vagina"])
```

```
## [1] 0.0383
```

Correlation matrix

```
length(cov_matrix)
```

```
## [1] 33994
```

```
names(cov_matrix)[1:10]
```

```
##  [1] "MCF2L2"        "TRMT10C"       "CORO1A"        "C
##  [5] "THOP1"         "RP11-731K22.1" "DBNDD2"        "V
##  [9] "PDCD5"         "CTD-2026D20.2"
```

```
#Obtain correlation
gene = 'EXOC3L2'
cor_matrix <- cov2cor(cov_matrix[[gene]])
```

Correlation matrix

```
dim(cor_matrix)
```

```
## [1] 23 23
```

```
round(cor_matrix[15:18,15:18],3)
```

```
##           Liver   Lung Pancreas Pituitary
## Liver     1.000  0.401    0.053     0.003
## Lung      0.401  1.000    0.120    -0.143
## Pancreas  0.053  0.120    1.000    -0.137
## Pituitary 0.003 -0.143   -0.137     1.000
```

Section 3

Analysis (SuSiE)

SuSiE

```r
devtools::install_github("stephenslab/susieR")
library(susieR)
fitted_rss <- susie_rss(z-scores, R, L = 10)
```

- z-scores: A p-vector of z scores
- R: p by p correlation matrix
- L: Maximum number of components model (Credible Sets).

Implementation

- run_susie
  - Pre-process data
    - Drop NA in expression z-scores vector
    - Take out the common tissue information from expression vector and correlation matrix.
  - Fit model: susie_rss(z-scores, R, L=4)
    - Expression z-score matrix is of length p
    - Correlation matrix is p by p matrix
    - They contain same tissue information
    - L=4

Main logic

We loop through all genes. For each gene, we implement run_susie, and take out the significant tissues in Credible Sets (cs), as well as their Posterior inclusion probability (PIP) scores. We stored these information in a csv file.

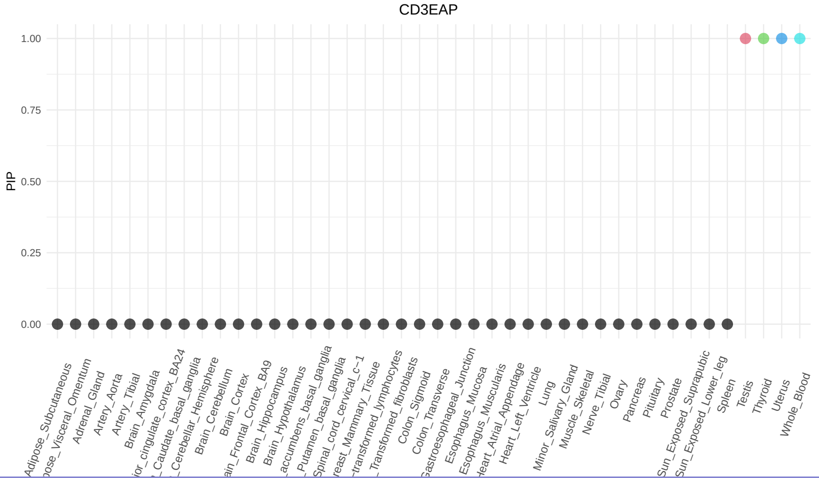
Result

```
head(all_res)
```

```
##   X variable_prob cs                tissues    GENE
## 1 1     1.0000000  1            Whole_Blood EXOC3L2
## 2 2     1.0000000  2 Adipose_Subcutaneous EXOC3L2
## 3 3     0.9999839  4                Testis EXOC3L2
## 4 4     0.9933120  3 Heart_Left_Ventricle EXOC3L2
## 5 5     1.0000000  1                Uterus  CLASRP
## 6 6     1.0000000  2               Thyroid  CLASRP
```

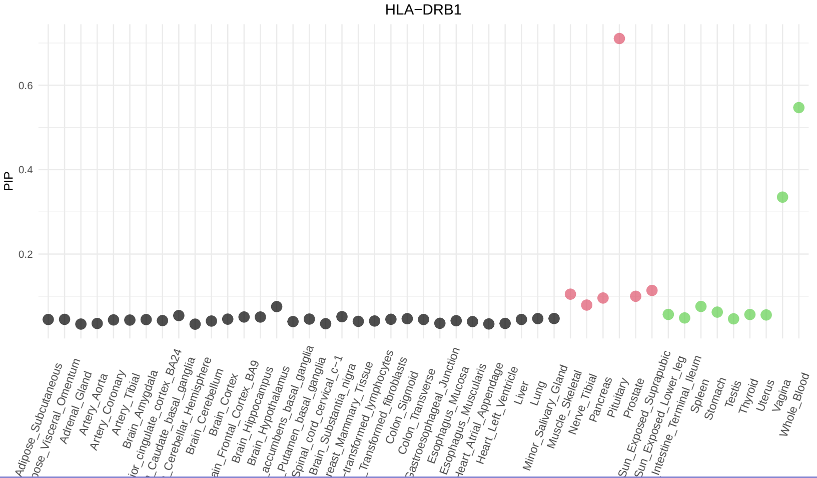
Result

```
plot_pip('CD3EAP',dat_ad,cov_matrix,all_res)
```



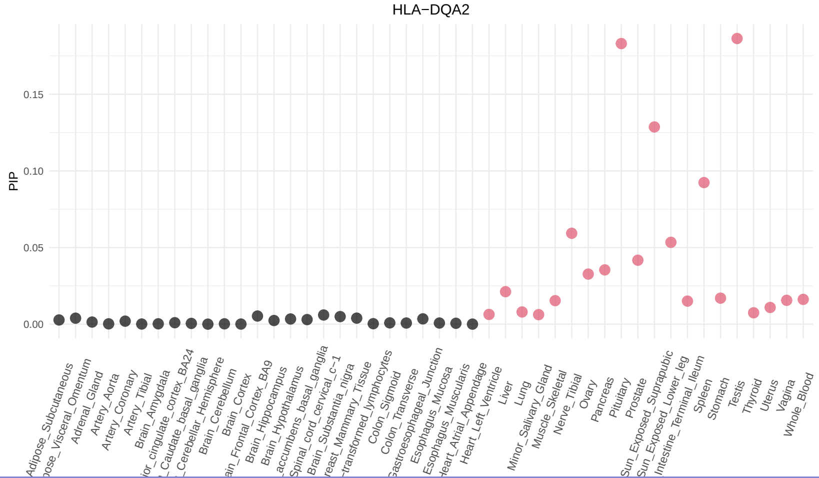
Result

```
plot_pip('HLA-DRB1',dat_ad,cov_matrix,all_res)
```



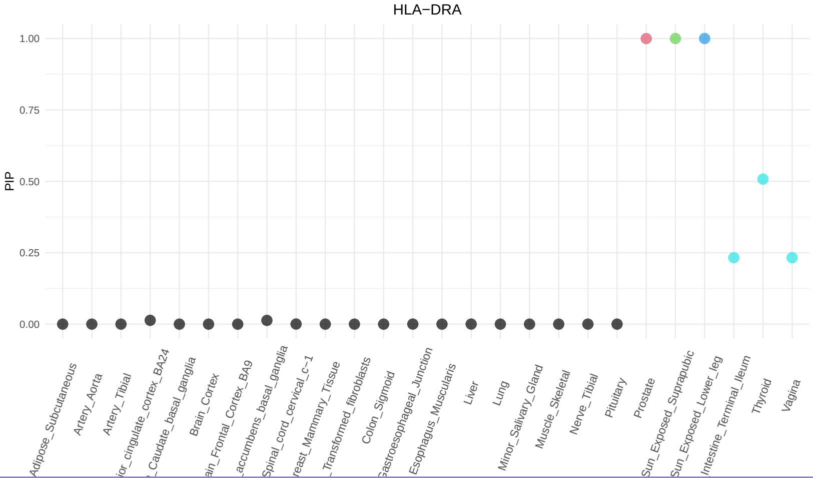
Result

```
plot_pip('HLA-DQA2',dat_ad,cov_matrix,all_res)
```



Result

```
plot_pip('HLA-DRA',dat_ad,cov_matrix,all_res)
```



Section 4

Reference

Reference

- Gusev et al. "Integrative approaches for large-scale transcriptome-wide association studies" 2016 Nature Genetics
- Wang, G., Sarkar, A., Carbonetto, P., & Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. Journal of the Royal Statistical Society: Series B (Statistical Methodology). https://doi.org/10.1111/rssb.12388
- Schaid, D.J., Chen, W. & Larson, N.B. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat Rev Genet 19, 491–504 (2018). https://doi.org/10.1038/s41576-018-0016-z
- ▸ Hilary Finucane, Broad Institute