TisCoMM User Manual

Xingjie Shi
shixingjie0615@gmail.com

2020-08-13

Contents

Introduction

TisCoMM package provides a unified probabilistic model for TWAS, leveraging the co-regulation of genetic variations across different tissues explicitly. **TisCoMM** not only performs hypothesis testing to prioritize gene-trait associations, but also detects the tissue-specific role of candidate target genes in complex traits. To make use of widely available GWAS summary statistics, TisCoMM is extended to use summary-level data, namely, TisCoMM-S^2 .

Models

TisCoMM tests for gene-trait associations one gene at a time. Assume $\mathcal{D}_1 = \{\mathbf{Y}_g, \mathbf{X}_{1g}\}$ denote the reference transcriptome data set of gene g for n_1 samples over T tissues, e.g. $\mathbf{Y}_g \in \mathbb{R}^{n_1 \times T}$ is the expression matrix for

this gene over T tissues, $\mathbf{X}_{1g} \in \mathbb{R}^{n_1 \times M_g}$ is the genotype matrix for cis-SNPs within this gene. Denote the GWAS data $\mathcal{D}_2 = \{\mathbf{z}, \mathbf{X}_{2g}\}$, where $\mathbf{z}$ is an $n_2 \times 1$ vector of phenotypic values, $\mathbf{X}_{2g}$ is the genotype matrix for M_g cis-SNPs.

To simplify notation we will omit the subscript g in all the expression that has dependence on gene g. Our model is

$$\mathbf{Y} = \mathbf{X}_1\mathbf{B} + \mathbf{E},$$
$$\mathbf{z} = \mathbf{X}_2\mathbf{B}\alpha + \mathbf{e}_z,$$

where $\alpha \in \mathbb{R}^T$, $\mathbf{E} \sim \mathcal{MN}(0, \mathbf{I}_m, \mathbf{V}_e)$, and $\mathbf{e}_z \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$. Note that we assume $\mathcal{D}_1$ and $\mathcal{D}_2$ are centered and thus intercepts can be omitted.

Statistical Inference in TisCoMM

Problem:

1. Parameter Estimation: $\mathbf{V}_e, \sigma_b^2, \sigma^2, \alpha$;
2. (testing gene-trait association) do joint testing: $\alpha = 0$ to detect significant candidate genes;
3. (testing tissue-specific effects) for significant genes only, do tissue-specific tesing: $\alpha_t = 0$, $t = 1, \ldots T$.

Methods:

- the EM algorithm with parameter expansion (Liu et al. 1998, Biometrika)
- the log-likelihood ratio test (LRT)
- multiple testing (p-values $\leq 5 \times 10^{-6}$ for joint test)

Installation

To install the development version of **TisCoMM**, it's easiest to use the 'devtools' package. Note that **TisCoMM** depends on the 'Rcpp' package, which also requires appropriate setting of Rtools and Xcode for Windows and Mac OS/X, respectively.

```
library(devtools)
install_github("XingjieShi/TisCoMM")
```

Real Data Analysis with GWAS Individual Data

1. Preparing eQTL data

The eQTL data consists of matched genotype data and gene expression data in multiple target tissues.

Genotype data ($\mathbf{X}_1$)

Genotype data of the eQTL samples in the PLINK binary format (.**bed**), and must be accompanied by .**bim** and .**fam** files with the same prefix. For example:

- GTEx.bed,
- GTEx.bim,
- GTEx.fam.

Gene expression across multiple tissues ($\mathbf{Y}_1$)

- Please note that gene expression in each tissue should be previously normalized for covariates which may confound the eQTL associations. We can achieve this by regressing the phenotypes (gene expressions) on the covariates and use the residuals as new phenotypes. After peforming this procedures for all the tissues, they can be specified as input of TisCoMM.
- Each tissue file should be a **tab-delimited** text file and include following information for each gene:
  - Start location
  - End location
  - Gene type
  - Hugo name
  - Ensembl ID
  - Chromosome number
  - normalized expression levels cross samples

  If some gene annotation is not included in the original gene expression file, one will have to extract these information by performing gene ID mapping with other annotation files. GENCODE (https://www.gencodegenes.org/human/) provides comprehensive gene annotation files.
- *TisCoMM* will use the headers in the expression files to extract information. It is required to have specific columns in all the formatted expression file. See Table 1 for a demonstration. Note that the first six column names should be exactly the same as those in Table 1. Expression levels across individuals should be appended after the first six columns.

Table 1: The first three rows and eight columns in an example gene expression file (rows for genes, and columns after the first six columns for samples).

lower	up	genetype1	genetype2	TargetID	Chr	ID1	ID2	⋯
59783540	59843484	lincRNA	PART1	ENSG00000152931.6	5	0.51	0.71	⋯
48128225	48148330	protein_coding	UPP1	ENSG00000183696.9	7	1.41	-0.01	⋯
57846106	57853063	protein_coding	INHBE	ENSG00000139269.2	12	0.58	-1.02	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

After this step, we will have the eQTL gene expression files:

- Skin_Sun_Exposed_Lower_leg_gene_expression.txt
- Whole_Blood_gene_expression.txt

2. Preparing GWAS data

TisCoMM can handle two different types of GWAS input dataset, the individual level data and summary statistics. There are some difference between these two types of input, here we discuss them separately.

Formatting GWAS individual data

The GWAS individual data files consist of genotype ($\mathbf{X}_2$) and phenotype data $\mathbf{z}$ for all GWAS samples. They should be in the PLINK binary format. For example:

- NFBC1966.bed,

- NFBC1966.bim,
- NFBC1966.fam.

You could optionally add the covariate file, which contains all confounding covariates used to adjust population stratification in the GWAS data. The covariate file should be formatted in a similar manner to the plink phenotype file, which should also be a **tab-delimited** file.

3. Testing gene-trait associations with GWAS individual data

```
# eQTL genotype file
file1 <- "GTEx_qc"

# GWAS individual level data
file2 <- "NFBC1966_qc"

# eQTL gene expression files
file3 <- c("Skin_Sun_Exposed_Lower_leg_gene_expression.txt",
           "Whole_Blood_gene_expression.txt")

# eQTL covariates file. Since normalized GE is provided, we do not need this file.
file4 <- ""

# GWAS covariates file
file5 <- ""

wihchPheno <- 1
bw         <- 5e5
coreNum    <- 24

fit <-mammot_paral(file1, file2, file3, file4, file5,
                   wihchPheno, bw, coreNum)
```

There are other three arguments.

- whichPheno specifies which phenotype in the phenotype file (GTEX_qc.fam) is used for association tests.
- bw defines the cisSNPs within a gene: either up to bw proximal to the start of gene, or up to bw distal to the end of the gene.
- corNum sets the number of threads the program will use.

4. Testing tissue-specific effects with GWAS individual data

```
# eQTL genotype file
file1 <- "GTEx_qc"

# GWAS individual level data
file2 <- "NFBC1966_qc"

# eQTL gene expression files
file3 <- c("Skin_Sun_Exposed_Lower_leg_gene_expression.txt",
           "Whole_Blood_gene_expression.txt")
```

```
# eQTL covariates file. If normalized GE is provided, we do not need this file.
file4 <- ""

# GWAS covariates file
file5 <- ""

# genes (TargetID) on which we want to perform tissue-specific test.
targetList  <- c("ENSG00000196666.3"
                 "ENSG00000213619.5"
                 "ENSG00000149187.13")

wihchPheno <- 1
bw         <- 5e5
coreNum    <- 24

fit <-mammot_part_paral(file1, file2, file3, file4, file5,
                   targetList, wihchPheno, bw, coreNum)
```

Compared with the input for gene-trait association test, there is a new arguments "targetList". It is a character vecter containing genes' names. Note that it should be the same identifiers as "targetID" in the eQTL data. In general, You can perform the tissue-specific test on any genes. Usually, we would like to focus on genes which are significantly associated with the trait.

Real Data Analysis with GWAS Summary Statisitc Data

1. Preparing eQTL data

This step is the same as analysis with GWAS individual data.

2. Formatting GWAS summary statisitc data

Reference panel ($\mathbf{X}_r$)

Reference panel in the PLINK binary format (**.bed**, **.bim**, **.fam**). For example,

- 1000G.bed,
- 1000G.bim,
- 1000G.fam.

GWAS summary statistic data

GWAS summary statistic data is required to have specific columns. See Table 2 for a demonstration. Note that all the column names should be exactly the same as those in Table 2.If GWAS summary statistic in the original downloaded file do not come with all the information *TisCoMM* needs, one will have to compute them manually. For example, if odds ratio is included, then beta can be computed as log(Odds Ratio). Assume our interested trait is the late-onset Alzheimer's disease (LOAD), and we download the summary statistic file from http://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php. After this step, the summary statistic is formatted correctly in following file:

- LOAD.txt

Table 2: An example for the GWAS summary statistics.

SNP	chr	BP	A1	A2	beta	se
rs3094315	1	752566	G	A	-0.0122	0.0294
rs3128117	1	944564	C	T	-0.0208	0.0278
rs1891906	1	950243	C	A	-0.0264	0.0260
rs2710888	1	959842	T	C	-0.0439	0.0297
rs4970393	1	962606	G	A	-0.0252	0.0233
rs7526076	1	998395	A	G	-0.0512	0.0229
⋮	⋮	⋮	⋮	⋮	⋮	⋮

3. Testing gene-trait associations with GWAS sumary statistics

```
# eQTL genotype file
file1 <- "GTEx_qc"

# GWAS summary statistic file
file2 <- "LOAD.txt"

# reference panel file
file3 <- "1000G"

# eQTL gene expression files
file4 <- c("Skin_Sun_Exposed_Lower_leg_gene_expression.txt",
           "Whole_Blood_gene_expression.txt")

# eQTL covariates file. Since normalized GE is provided, we do not need this file.
file5 <- ""

lam     <- 0.95
bw      <- 5e5
coreNum <- 24



fit <- mammotSS_paral(file1, file2, file3, file4, file5,
                      lam, bw, coreNum)
```

There are other three arguments.

- lam is the shrinkage intensify for the reference panel.
- bw defines the cisSNPs within a gene: either up to bw proximal to the start of gene, or up to bw distal to the end of the gene.
- corNum sets the number of threads the program will use.

4. Testing tissue-specific effects with GWAS individual data

```
# eQTL genotype file
file1 <- "GTEx_qc"

# GWAS summary statistic file
```

```
file2 <- "LOAD.txt"

# reference panel file
file3 <- "1000G"

# eQTL gene expression files
file4 <- c("Skin_Sun_Exposed_Lower_leg_gene_expression.txt",
           "Whole_Blood_gene_expression.txt")

# eQTL covariates file. Since normalized GE is provided, we do not need this file.
file5 <- ""

# genes (TargetID) on which we want to perform tissue-specific test.
targetList  <- c("ENSG00000196666.3"
                 "ENSG00000213619.5"
                 "ENSG00000149187.13")

lam     <- 0.95
bw      <- 5e5
coreNum <- 24

fit <- mammotSS_part_paral(file1, file2, file3, file4, file5,
                      targetList, lam, bw, coreNum)
```

Compared with the input for gene-trait association test, there is a new arguments "targetList". It is a character vecter containing genes' names. Note that it should be the same identifiers as "targetID" in the eQTL data. In general, You can perform the tissue-specific test on any genes. Usually, we would like to focus on genes which are significantly associated with the trait.

Replicate simulation results in Shi et al. (2020)

All the simulation results can be reproduced by using the code at simulation. Before running simulation to reproduce the results, please familiarize yourself with **TisCoMM** using 'TisCoMM User Manual'.

1. Simulation results for multi-tissue joint can be reproduced by following steps:
    - ExampleOne.R: This function can be run in a HPC cluster (with minor revisions, it could be run on a PC), it will output files, named pvalue_hz0.1_hc0.25_rhoX5_s5_batch-6.txt, which contain inference results of each replicate, for all multi-tissue TWAS methods: TisCoMM, TisCoMM-S^2, MultiXcan, S-MultiXcan and UTMOST.
    - ExampleOnePlot.R: This function produces simulation figures of joint test in Shi et al. (2019).
2. Simulation results for tissue-specific test can be reproduced by following steps:
    - PartCoMMCOR.R: This function can be run in a HPC cluster (with minor revisions, it could be run on a PC), it will output files, named part_hc4_rhoX8_rhoW8nz_ti2_batch-2.rds, which contain inference results of each replicate, for all single-tissue TWAS methods: CoMM, PrediXcan, and TWAS.
    - SummaryCOR.R: This function produces simulation figures of tissue-specific test in Shi et al. (2019).

Reference

Shi et al (2020). A tissue-specific collaborative mixed model for jointly analyzing multiple tissues in transcriptome-wide association studies