GIFT User Manual

Zhongshang Yuan and Xiang Zhou
E-mail: yuanzhongshang@sdu.edu.cn and xzhousph@umich.edu

2023-04-22

Contents

1 Introduction

1.1 What is GIFT

GIFT (Gene-based Integrative Fine-mapping through conditional TWAS), is an R package for efficient statistical inference of conditional TWAS fine-mapping. GIFT examines one genomic region at a time, jointly models the genetically regulated expression (GReX) of all genes residing in the focal region, and carries out TWAS conditional analysis in a maximum likelihood framework. GIFT explicitly models the gene expression correlation and cis-SNP LD across different genes in the region, accounts for the uncertainty in the constructed GReX through joint inference and provides calibrated *p* values.

1.2 How to Cite GIFT

Lu Liu, Ran Yan, Ping Guo, Jiadong Ji, Weiming Gong, Fuzhong Xue, Zhongshang Yuan, and Xiang

Zhou (2023). Conditional transcriptome-wide association study for fine-mapping causal genes.

1.3 The GIFT method

GIFT tests for gene-trait associations one region at a time. The model is

$$\boldsymbol{x}_i = \boldsymbol{Z}_i \boldsymbol{\beta}_i + \boldsymbol{e}_i, i = 1, \cdots, k, \tag{1}$$

$$\boldsymbol{y} = \sum_{i=1}^{k} \alpha_i \widetilde{\boldsymbol{Z}}_i \boldsymbol{\beta}_i + \tilde{\boldsymbol{e}}, \tag{2}$$

where equation (1) is for the gene expression data and equation (2) is for the GWAS data. Here, $\boldsymbol{x}_i$ is an n_1-vector as the i-th gene expression levels measured on n_1 individuals in the gene expression study; $\boldsymbol{Z}_i$ is an n_1 by p_i matrix of genotypes for p_i *cis*-SNPs of the i-th gene; $\boldsymbol{\beta}_i$ is a p_i-vector of *cis*-SNP effect sizes on the i-th gene expression; $\boldsymbol{e}_i$ is an n_1-vector of residual error and $(\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_k)$ following a matrix normal distribution $MN_{n_1,k}(0, \boldsymbol{I}_{n_1}, \boldsymbol{\Omega})$, where $\boldsymbol{I}_{n_1}$ is an n_1 by n_1 identity matrix and $\boldsymbol{\Omega}$ is a k by k symmetric variance-covariance matrix among the k different gene expression levels with $\boldsymbol{\Omega} = \boldsymbol{DRD}$, where $\boldsymbol{R}$ is the estimated gene expression correlation based on the gene expression mapping study and $\boldsymbol{D}$ is a diagonal matrix of standard deviations; $\boldsymbol{y}$ is an n_2-vector of outcome trait measured on n_2 individuals in the GWAS; $\widetilde{\boldsymbol{Z}}_i$ is an n_2 by p_i matrix of genotypes for the same p_i *cis*-SNPs of the i-th gene; α_i represents the causal effect of i-th gene expression on the outcome; $\tilde{\boldsymbol{e}}$ is an n_2-vector of residual error with each element independently and identically distributed from the same normal distribution $N(0, \sigma_y^2)$. Regarding the SNP effect sizes $\boldsymbol{\beta}$ as missing data, GIFT develops a parameter-expanded version of expectation-maximization (PX-EM) algorithm for inference. PX-EM is able to improve the convergence rate while enjoys the stability of traditional EM algorithm.

2 Installation

To install the development version of GIFT, it's easiest to use the 'devtools' package. Appropriate setting of Rtools is required, given that GIFT relies on the 'Rcpp' package.

```
# install.packages("devtools")
library(devtools)
install_github("yuanzhongshang/GIFT")
```

3 Application analysis

GIFT can handle both the individual level data and summary statistics, with details provided below.

3.1 Individual level Data

The main function for GIFT method with individual-level data is

```
GIFT_individual(X, Y, Zx, Zy, pindex, max_iterin =1000, epsin=1e-4, Cores=1)
```

3.1.1 Input format

- **X**: standardized gene expression matrix in eQTL data. Each row contains all the gene expressions residing the target region for each individual. For example, an analysis involving three individuals and two genes in analyzed region can be represented as follows:

  ```
  -0.2878446    -0.5181887
  1.1123536     -0.6345557
  -0.8245090    1.1527444
  ```
- **Y**: vector of standardized phenotypes. Each row is the phenotype e for each individual. For example, a phenotype vector with four individuals can be represented as follows:

  ```
  -0.6570064
  0.8205077
  -1.0517360
  0.8882348
  ```
- **Zx**: standardized cis-genotype matrix in eQTL data. For example, a standardized cis-genotype matrix with three individuals and two genes containing 2, 2 cis-SNPs respectively, can be represented as follows:

  ```
  -1.1547005    1     -0.5773503    0.5773503
  0.5773503     -1    -0.5773503    0.5773503
  0.5773503     0     1.1547005     -1.1547005
  ```
- **Zy**: standardized cis-genotype matrix in GWAS data.
- **pindex**: the vector representing the number of cis-SNPs for each gene, e.g. pindex=c(2,2) for the above example.
- **max_iterin**: the user-defined maximum iteration with the default to be 1000 (max_iterin =1000).
- **epsin**: the user-defined convergence tolerance of the absolute value of the difference between the nth and (n+1)th log likelihood, with the default to be 1e-4 (epsin=1e-4).
- **Cores**: the user-defined number of cores used in this algorithm, with the default to be 1 (Cores=1). If the number of cores is greater than 1, analysis will perform with fast parallel computing. The function mclapply() depends on another R package "parallel" in Linux.

3.1.2 Running GIFT

Users are suggested to follow the procedure of the simple example within the "example" folder to run GIFT.

3.1.3 Output

- **causal_effect**: the vector representing the estimate of causal effect for each gene in a region. For example, the result of an analysis involving two genes can be represented as follows:

  ```
  0.04599541
  0.04628380
  ```
- **gene_based_pvalue**: the vector representing p value for testing the causal effect. For example, the result of an analysis involving two genes can be represented as follows:

  ```
  0.5785106
  0.7038768
  ```

3.2 Summary Statistics Data

The main function for GIFT method with summary statistics is

```
GIFT_summary(Zscore1, Zscore2, LDmatrix1, LDmatrix2, R, n1, n2, pindex, max_iterin =1000,
epsin=1e-4, Cores=1)
```

3.2.1 Input format

- **Zscore_1**: the Zscore matrix of the cis-SNP effect size for all the genes within the target region in eQTL data. For example, the Zscore matrix with two genes containing 2, 2 cis-SNPs respectively can be represented as follows:
  ```
  2.1454620     -0.0354723
  0.1676732     -2.1554618
  -1.7967432    1.6588060
  ```
- **Zscore_2**: the Zscore vector of the cis-SNP effect size for the phenotype in GWAS data. For example, the Zscore vector for four cis-SNPs on the GWAS phenotype in the above example can be represented as follows:
  ```
  -0.2309086
  -2.0382820
  -0.6549378
  1.4678436
  ```
- **LDmatrix1**: the LD matrix in eQTL data. For example, a standardized LD matrix with four cis-SNPs can be represented as follows:
  ```
  1        -0.8660254     0.5      -0.5
  -0.8660254 1            0         0
  0.5           0           1 -0.5773503
  -0.5          0.5     -0.8660254 1
  ```
- **LDmatrix2**: the LD matrix in GWAS data.
- **n1**: the sample size of eQTL data.
- **n2**: the sample size of GWAS data.
- **R**: the estimated correlated matrix of gene expressions.
- **pindex**: the vector representing the number of cis-SNPs for each gene, e.g. pindex=c(2,2) for the above example.
- **max_iterin**: the user-defined maximum iteration with the default to be 1000 (max_iterin =1000).
- **epsin**: the user-defined convergence tolerance of the absolute value of the difference between the nth and (n+1)th log likelihood, with the default to be 1e-4 (epsin=1e-4).
- **Cores**: the user-defined number of cores used in this algorithm, with the default to be 1 (Cores=1). If the number of cores is greater than 1, analysis will perform with fast parallel computing. The function mclapply() depends on another R package "parallel" in Linux.

3.2.2 Running GIFT

Users are suggested to follow the procedure of the simple example for the summary statistics within the "example" folder to run GIFT. Of note,

3.2.3 Output

- **causal_effect**: the vector representing the estimate of causal effect for each gene in a region. For example, the result of an analysis involving two genes can be represented as follows:
  0.04599541
  0.04628380
- **gene_based_pvalue**: the vector representing p value for testing the causal effect. For example, the result of an analysis involving two genes can be represented as follows:
  0.5785106
  0.7038768

- **Notes:** The summary statistics version of GIFT often requires the LD matrix calculated from the reference panel data, it would be better to ensure the ethnicity of the reference panel is consistent with that of the analyzed data.