GIFT User Manual

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

2024-10-24

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

Liu, L., Yan, R., Guo, P. et al. Conditional transcriptome-wide association study for fine-mapping candidate causal genes. Nat Genet 56, 348–356 (2024).

1.3 The GIFT method

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

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

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

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 $MN_{n_1,k}(0, \boldsymbol{I}_{\boldsymbol{n}_1}, \boldsymbol{\Omega})$, where $\boldsymbol{I}_{\boldsymbol{n}_1}$ 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

3.1 GIFT: Using individual-level data as input

3.1.1 Data pre-process

The function pre_process_individual can convert common genotype data formats from different softwares to GIFT inputs.

```
pre_process_individual(filelocation, plinkexe="plink")
```

The input from individual level data:

- **filelocation**: a directory representing a location only contains the genotype files to be processed.
- **plinkexe**: a directory indicating a path for plink, or using the default "plink" when you have added the executable file path of plink to the system's environment variable.

The output is a list containing:

- **gene**: the vector representing the gene names.
- **Z**: standardized cis-genotypes matrix for all the genes in a region.
- **pindex**: the vector representing the number of cis-SNPs for each gene.

Specifically, this function is flexible to handle plink binary format (.bim/.fam./.bed), vcf, ped/map format, csv, and tsv file. Note that, cis-genotype matrix has been standardized to have a mean of zero and standard derivation of one.

3.1.2 Running GIFT

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

```
GIFT_individual(X, Y, Zx, Zy, gene, pindex, maxiter =100, tol=1e-3, pleio=0, ncores=1, filter=T)
```

The input from individual level data:

- **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 for each individual. For example, a phenotype vector with three individuals can be represented as follows:
  ```
  -0.3657089
  1.1313758
  -0.7656669
  ```
- **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.
- **gene**: the vector representing the gene names in a region, the order of the gene name should be consistent with that in **X**.
- **pindex**: the vector representing the number of cis-SNPs for each gene, e.g. pindex = c(2, 2) for the above example.
- **maxiter**: the user-defined maximum iteration, with the default to be 100 (maxiter = 100).
- **tol**: 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-3 (tol = 1e-3).

- **pleio**: the user-defined option of controlling the pleiotropy, with the default to be 0 (pleio = 0). If pleio is set to 0, analysis will perform without controlling any SNP; If pleio is set to 1, analysis will perform controlling the top SNP; If pleio is set to 2, analysis will perform controlling the top two SNPs.
- **ncores**: the user-defined number of cores used in this algorithm, with the default to be 1 (ncores = 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.
- **filter**: The user-defined logical value, with the default to be T. If filter is set to T, the analysis will be performed using the SNPs with a GWAS p-value < 0.05 when the GWAS sample size over 100,000. This step will improve the computational speed.

The output from individual level data is a data frame containing:

- **gene**: the vector representing the gene names.
- **causal_effect**: the vector representing the causal effect estimate 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
- **p**: the vector representing p values 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 GIFT: Using summary statistics as input

3.2.1 Data pre-process

The function pre_process_summary can convert common summary statistics and LD matrix format from different softwares to GIFT inputs.

```
pre_process_summary(eQTLfilelocation, eQTLLDfile, GWASfile, GWASLDfile, snplist, pindex)
```

The input from summary statistics:

- **eQTLfilelocation**: a directory representing a file location only contains the summary statistics files from eQTL data.
- **eQTLLDfile**: a directory representing the LD matrix from eQTL data.
- **GWASfile**: a directory representing the summary statistics from GWAS data.
- **GWASLDfile**: a directory representing the LD matrix from GWAS data.
- **snplist**: the vector representing the cis-SNP list for all the genes in a region.
- **pindex**: the vector representing the number of cis-SNPs for each gene.

The output is a list containing:

- **gene**: the vector representing the gene names.
- **pindex**: the vector representing the number of cis-SNPs for each gene.
- **Zscore1**: the matrix representing the z-scores of each cis-SNPs for all the genes from eQTL

data.

- **Zscore2**: the vector representing the z-scores of each cis-SNPs for all the genes from GWAS data.
- **LDmatrix1**: the LD matrix from eQTL data.
- **LDmatrix2**: the LD matrix from GWAS data.
- **n1**: the sample size from eQTL data.
- **n2**: the sample size from GWAS data.

Specifically, this function is flexible to handle association test output from plink (.qassoc), GEMMA (.assoc.txt) and SAIGE (.txt). Meanwhile, this function is also flexible to handle LD matrix either from matrix or a long format such as h5 format. Note that, the summary statistics version of GIFT often requires the in-sample LD matrix. If the in-sample LD matrix is unavailable, it can be also calculated from the reference panel data (e.g., 1,000 Genomes project). It would be better to ensure the ethnicity of the reference panel is consistent with that of the analyzed data.

3.2.2 Running GIFT

The main function for GIFT method with summary statistics is

```
GIFT_summary(Zscore1, Zscore2, LDmatrix1, LDmatrix2, R, n1, n2, gene, pindex, maxiter =100,
tol=1e-3, pleio = 0, ncores=1, in_sample_LD=F, filter=T, split=5)
```

The input from summary statistics:

- **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
  -3.8318312   0.0164746
  ```
- **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 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.
- **gene**: the vector representing the gene names in a region, the order of the gene name should be consistent with that in **Zscore_1**.
- **pindex**: the vector representing the number of cis-SNPs for each gene, e.g. pindex = c(2, 2) for the above example.
- **maxiter**: the user-defined maximum iteration, with the default to be 100 (maxiter = 100).
- **tol**: 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-3 (tol = 1e-3).
- **pleio**: the user-defined option of controlling the pleiotropy, with the default to be 0 (pleio = 0). If pleio is set to 0, analysis will perform without controlling any SNP; If pleio is set to 1, analysis will perform controlling the top SNP; If pleio is set to 2, analysis will perform controlling the top two SNPs.
- **ncores**: the user-defined number of cores used in this algorithm, with the default to be 1 (ncores = 1). If the number of cores is greater than 1, analysis will perform with fast parallel computing. If you use the Windows system, foreach() depends on another R package “doParallel” would be used. Otherwise, mclapply() depends on another R package “parallel” would be used.
- **in_sample_LD**: the logical value representing whether in-sample LD was used, with the default to be F. If in-sample LD was not used, the LD matrix is regularized to be (1-s1)*Sigma1+s1*E and (1-s2)*Sigma2+s2*E, where both s1 and s2 are estimated by the function estimate_s_rss() in susieR, E is an identity matrix. A grid search algorithm is performed over the range from 0.1 to 1 once the estimation from susieR does not work well. The function estimate_s_rss() depends on another R package "susieR".
- **filter**: The user-defined logical value, with the default to be T. If filter is set to T, the analysis will be performed using the SNPs with a GWAS p-value < 0.05 when the GWAS sample size over 100,000. This step will improve the computational speed.
- **split**: If you use the Windows system, foreach() depends on another R package "doParallel" would be used. Otherwise, mclapply() depends on another R package "parallel" would be used.

The output from summary statistics data is a data frame containing:

- **gene**: the vector representing the gene names.
- **causal_effect**: the vector representing the causal effect estimate 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
- **p**: the vector representing p values for testing the causal effect. For example, the result of an analysis involving two genes can be represented as follows:
  0.5785106
  0.7038768

3.3 Two-stage version of GIFT: Using pre-trained weights and summary statistics as input

3.3.1 Data pre-process

The function weightconvert can convert the weights from all the genes into a required input of two-stage version of GIFT. Gene expression prediction is the key for two-stage TWAS methods. The commonly used prediction models include lasso and elastic net (enet) as implemented in prediXcan, Best Linear Unbiased Prediction (BLUP), the top SNPs (top1) and Bayesian sparse linear mixed model (BSLMM) as implemented in TWAS/FUSION, latent Dirichlet process regression (DPR) as implemented in both DPR and TIGAR.

```
weightconvert(weightlist)
```

The input from eQTL data is:

- **weightlist**: a list containing the weights for all the genes in a region.

The output is:

- **weightinput**: a block diagonal matrix.

The function pre_process_twostage can convert common convert common summary statistics and LD matrix format from GWAS data to match GIFT input.

```
pre_process_twostage(GWASfile, GWASLDfile, snplist)
```

The input from summary statistics:

- **GWASfile**: a directory representing the summary statistics from GWAS data.
- **GWASLDfile**: a directory representing the LD matrix from GWAS data.
- **snplist**: the vector representing the cis-SNP list for all the genes in a region.
- **pindex**: the vector representing the number of cis-SNPs for each gene.

The output is a list containing:

- **beta**: beta vector of each cis-SNPs for all genes from GWAS data.
- **se**: corresponding se vector of each cis-SNPs for all genes from GWAS data.
- **LDmatrix**: the LD matrix from GWAS data.

Specifically, this function is flexible to handle association test output from plink (.qassoc), GEMMA (.assoc.txt) and SAIGE (.txt). Meanwhile, this function is also flexible to handle LD matrix either from matrix or a long format such as h5 format. Note that, the summary statistics version of GIFT often requires the in-sample LD matrix. If the in-sample LD matrix is unavailable, it can be also calculated from the reference panel data (e.g., 1,000 Genomes project). It would be better to ensure the ethnicity of the reference panel is consistent with that of the analyzed data.

3.3.2 Running two-stage version of GIFT

The main function for the two-stage version of GIFT method using pre-trained weights and summary statistics as input is

```
GIFT_two_stage_summ(betax, betay, se_betay, Sigma, n, gene, in_sample_LD=F)
```

The input to run the two-stage version of GIFT:

- **betax**: the weight matrix of the cis-SNP effect size for all the genes within the target region in eQTL data. For example, the weight matrix with two genes containing 2, 2 cis-SNPs respectively can be represented as follows:

0.099600571	0
0.007784031	0
0	0.013852086
0	0.098520337

- **betay**: the beta vector of the cis-SNP effect size for the phenotype in GWAS data. For example, the beta vector for four cis-SNPs on the GWAS phenotype can be represented as follows:

  -0.0032658667

  -0.0288285441

  -0.0092631461

  -0.0273599280
- **se_betay**: the se vector of the cis-SNP effect size for the phenotype in GWAS data. For example, a se vector with four cis-SNPs can be represented as follows:

  0.01414214

  0.01414214

  0.01414214

  0.01414214
- **Sigma**: the LD matrix in GWAS 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

- **n**: the sample size of GWAS data.
- **gene**: the vector of gene names.
- **in_sample_LD**: the logical value representing whether in-sample LD was used, with the default to be F. If in-sample LD was not used, the LD matrix is regularized to be (1-s)*Sigma+s*E, where s is estimated by the function estimate_s_rss() in susieR, E is an identity matrix. The function estimate_s_rss() depends on another R package "susieR".

The output from two-stage version of GIFT is a data frame containing:

- **gene**: the vector representing the gene names.
- **z**: the vector representing Z value for testing the causal effect. For example, the result of an analysis involving two genes can be represented as follows:

  0.8849015

  0.6931964
- **p**: 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:

  0.3762097

  0.4881863