TIGAR Manual

Xiaoran Meng
mengxiaoran0629@gmail.com

Emory University

Contents

1 Introduction

"**TIGAR**" standing for **T**ranscriptome-**I**ntergrated **G**enetic **A**ssociation **R**esource, which is developed using Python and BASH scripts.TIGAR can fit both Elastic-Net and nonparametric Beyesian model(Dirichlet Process Regression, i.e. DPR), impute transcriptomic data, and conduct genetic association studies using both individual-level and summary-level GWAS data for univariate and multivariate phenotypes.

2 Installation

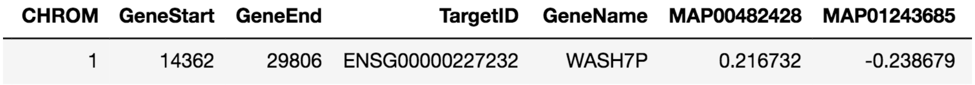
To run TIGAR, we need following software

- Python 3.5
  - dfply **:** Work similar to R's dplyr package.
  - io **:** Decode genotype data input from TABIX result.
  - subprocess **:** Read in TABIX result.
  - multiprocessing
- TABIX

3 Input

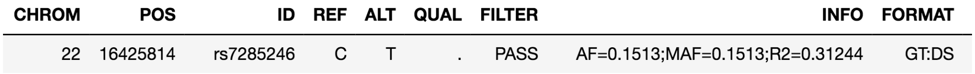
3.1 Training

- model **:** Training imputation model for transcriptomic data (elastic_net or DPR).
- Gene_Exp **:** Combination of gene annotation and expression level file, with first five columns CHROM, GeneStart, GeneEnd, TargetID/GeneID and GeneName.



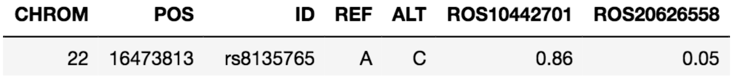
CHROM	GeneStart	GeneEnd	TargetID	GeneName	MAP00482428	MAP01243685
1	14362	29806	ENSG00000227232	WASH7P	0.216732	-0.238679

- train_sample **:** A column of sampleIDs use for training.
- chr **:** Chromosome number.
- geno_train **:** vcf or dosages.
- train_dir **:** Training genotype data with vcf or dosages format.This file should be tabixed (contains .gz and .gz.tbi).
  - vcf **:** First nine columns fixed.



CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT
22	16425814	rs7285246	C	T	.	PASS	AF=0.1513;MAF=0.1513;R2=0.31244	GT:DS

  - dosages **:** First five columns fixed.



CHROM	POS	ID	REF	ALT	ROS10442701	ROS20626558
22	16473813	rs8135765	A	C	0.86	0.05

- FT **:** Format using for training data(GT or DS).
- maf **:** Threshold for Minor Allele Frequency (range from 0-1),default 0.01. TIGAR will select snps with maf greater than this threshold for training.

- hwe : Threshold of p-value for Hardy Weinberg Equilibrium exact test,default 0.001. TIGAR will select snps p-value greater than this threshold for training.
- window : Window size around gene boundary, default is 10^6BP.
- thread : number of thread for multiprocessing with default 1. If thread>1, for example thread=10, it will run 10 genes simultaneously. In other words, it will accelerate training procedure.
- out : Path you want to save output files.
- Input only for elastic net
  - cv : Number of folds used in cross-validation to select prameter for elastic net regression. TIGAR use 5-fold as default.
  - alpha : Ratio for L_1 & L_2 penalty for elastic net regression.
- Input only for DPR
  - dpr : $-dpr1$ fits DPR using variation Bayesian algorithm, $-dpr2$ fits DPR using MCMC sampling with fixed number of normal components in mixture prior and $-dpr3$ fits DPR using MCMC sampling with adaptively selected number of the normal components in mixture prior. Default is 1.
  - ES : Effect size (fixed or additive). For fixed effect size, $ES = beta$. For additive effect size, $ES = b + beta$. Default is fixed.

3.2 Prediction

- pred : y or n. If pred is y, TIGAR will run predict the transcriptome from a given genotype file based on parameters from model training part.
- geno_test : vcf or dosages
- test_dir : Genotype data used for prediction with vcf or dosages format. Requirement is the same as training genotype data.
- FP : Format using for prediction data(GT or DS).
- maf_diff : Threshold of difference between training maf and testing maf. If difference correspond to a snp is larger than this threshold, TIGAR will drop this snp. Default is 0.2.

3.3 Association Study

- asso :
  - If $asso = 1$, run TWAS with predicted gene expression data provide by model training.
    - PED : PED file.
    - Asso_Info : Instruction for association study.
      - P stands for column names for corresponding phenotype in PED file.
      - C stands for column names for covariates in PED file.
    - method : Link function. OLS for ordinary least square regression. Logit for logistic regression.
  - If $asso = 2$, run TWAS with additional Z-score from GWAS.
    - Zscore : Zscore file from previous GWAS study (tabixed).
    - Weight : File contains snps effect size (Same format as prediction output file).
    - Covar : Reference covariance matrix (Scripts is provided, see in covar_calculation.py, covar_calculation.sh, tabixed)
    - chr : Chromosome number.
    - window : Window size around gene boundary, default is 10^6BP.

- thread : Number of thread for multiprocessing with default 1.
- out : Path you want to save output files.
- Reference covariance matrix calculation (covar_calulation.py, covar_calculation.sh, additional part)
  - block : Provided in the example_data (./example_data/block_annotation.txt). Block annotation is based on the LD structure of European samples.
  - geno_path : Genotype file (tabixed)
  - geno : vcf or dosages. (Same format as in model training)
  - chr : Chromosome number.
  - Format : GT or DS.
  - maf : Threshold for Minor Allele Frequency (range from 0-1). Default is 0.05.
  - thread : Number of thread. Default is 1.
  - out : Path you want to save output file.

Example of input files are shown in https://github.com/xmeng34/GEtools/tree/master/TIGAR

4 Example Usage

4.1 Model Training

- Training Inputs
  - Gene_Exp_path=./example_data/Gene_Exp_combination.txt
  - train_sample_path=./example_data/training_sampleID.txt
  - train_dir=./example_data/Train_VCF.vcf.gz
  - out_prefix=./result
- Elastic Net Regression

Command Line

```
$ ./TIGAR_Model_Train.sh --model elastic_net \
$ --Gene_Exp ${Gene_Exp_path} --train_sample ${train_sample_path} \
$ --chr 1 --train_dir ${train_dir} \
$ --geno_train vcf --FT DS \
$ --out ${out_prefix}
```

- DPR

Command Line

```
$ ./TIGAR_Model_Train.sh --model DPR \
$ --Gene_Exp ${Gene_Exp_path} --train_sample ${train_sample_path} \
$ --chr 1 --train_dir ${train_dir} \
$ --geno_train vcf --FT DS \
$ --out ${out_prefix}
```

4.2 Prediction

- Predicion Inputs
  - test_dir=./example_input/Test_dosages.txt.gz
- Based on Elastic Net Regression

Command Line

```
$ ./TIGAR_Model_Train.sh --model elastic_net \
$ --Gene_Exp ${Gene_Exp_path} --train_sample ${train_sample_path} \
$ --chr 1 --train_dir ${train_dir} \
$ --geno_train vcf --FT DS \
$ --pred y \
$ --geno_test dosages --FP DS \
$ --out ${out_prefix}
```

- Based on DPR

Command Line

```
$ ./TIGAR_Model_Train.sh --model DPR \
$ --Gene_Exp ${Gene_Exp_path} --train_sample ${train_sample_path} \
$ --chr 1 --train_dir ${train_dir} \
$ --geno_train vcf --FT DS \
$ --pred y \
$ --geno_test dosages --FP DS \
$ --out ${out_prefix}
```

4.3 Association Study

- Association Study Input ($\boldsymbol{asso = 1}$)
  - Gene_Exp_path=./example_data/Gene_Exp_combination.txt
  - PED=./example_data/PED_file.ped
  - Asso_Info=./example_data/Asso_Info.txt
  - out_prefix=./result

Command Line

```
$ ./TIGAR_TWAS.sh --asso 1 \
$ --Gene_Exp ${Gene_Exp_path} \
$ --PED ${PED} \
$ --Asso_Info ${Asso_Info} \
$ --out ${out_prefix}
```

- Association Study Input ($\boldsymbol{asso = 2}$)
  - Gene_Exp_path=./example_data/Gene_Exp_combination.txt
  - Zscore=./example_data/Zscore.txt.gz
  - Weight=./example_data/Weight.txt
  - Covar=./example_data/CHR22_reference_cov.txt.gz
  - out_prefix=./result

Command Line

```
$ ./TIGAR_TWAS.sh --asso 2 \
$ --Gene_Exp ${Gene_Exp_path} \
$ --Zscore ${Zscore} --Weight ${Weight} --Covar ${Covar} \
$ --chr 22 \
$ --out ${out_prefix}
```

- Reference Covariance Matrix Calculation
  - block=./example_data/block_annoation.txt
  - geno_path=./example_data/Train_VCF.vcf.gz
  - out_prefix=./result

Command Line

```
$ ./covar_calculation.sh --block ${block} \
$ --geno_path ${geno_path} --geno vcf \
$ --chr 22 \
$ --Format GT \
$ --out ${out_prefix}
```

4.4 Change Default Values

- Model Training
  - If you want to change default value within scripts, say $\boldsymbol{alpha}$ and $\boldsymbol{cv}$ in elastic net model in model training and $\boldsymbol{maf\_diff}$ in prediction

Command Line

```
$ ./TIGAR_Model_Train.sh --model elastic_net \
$ --Gene_Exp ${Gene_Exp_path} --train_sample ${train_sample_path} \
$ --chr 1 --train_dir ${train_dir} \
$ --geno_train vcf --FT GT \
$ --alpha 0.8 --cv 10 \
$ --pred y \
$ --geno_test dosages --FP DS \
$ --maf_diff 0.3 \
$ --out ${out_prefix}
```

- TWAS

Command Line

```
$ ./TIGAR_TWAS.sh --asso 1 \
$ --Gene_Exp ${Gene_Exp_path} \
$ --PED ${PED} \
$ --Asso_Info ${Asso_Info} \
$ --method Logit \
$ --out ${out_prefix}
```

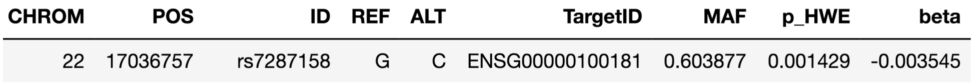
5 Output

For model training and prediction, some share output variables are listed as follow

- Training Parameter Files
  - CHROM: Chromosome number
  - POS: Snp position
  - TargetID: Gene correspond to this snp(GeneID)
  - MAF: Minor Allele Frequency(range from 0-1)
  - p_HWE: P-value for Hardy Weinberg Equilibrium exact test for this snp
- Training Information & Prediction Files
  - CHROM: Chromosome number
  - GeneStart: Position of this gene start
  - GeneEnd: Posistion of this gene end
  - GeneName: Name of this gene
  - GeneFunction: Function of this gene
  - TargetID: GeneID
  - sample_size: Number of snps used for regression
  - effect_sample_size: Number of snps that have regression coefficient not equal to 0
  - R2: Regression R^2 for model training

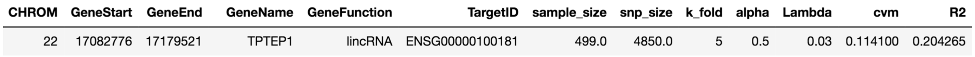
Some unique variable for specific output files are listed as follow

- Elastic Net Training Parameter File
  - ID: rsID
  - REF: Reference allel
  - ALT: Alternative allel
  - beta: Effect size estimation based on elastic net regression.
    We only keep snps that have beta≠0.



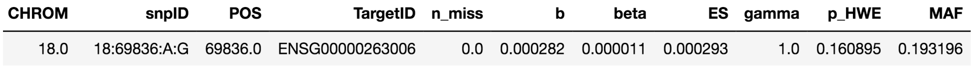
CHROM	POS	ID	REF	ALT	TargetID	MAF	p_HWE	beta
22	17036757	rs7287158	G	C	ENSG00000100181	0.603877	0.001429	-0.003545

- Elastic Net Training Information File
  - k_fold: folds we use for crossvalidation(ex.5-folds)
  - alpha: L_1 & L_2 ratio for elastic net regression
  - Lambda: Constant that multiplies the penalty terms. Selected by cross-validation.
  - cvm: Mean cross-validated score corresponding to selected lambda.



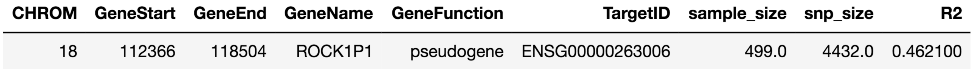
CHROM	GeneStart	GeneEnd	GeneName	GeneFunction	TargetID	sample_size	snp_size	k_fold	alpha	Lambda	cvm	R2
22	17082776	17179521	TPTEP1	lincRNA	ENSG00000100181	499.0	4850.0	5	0.5	0.03	0.114100	0.204265

- DPR Training Parameter File
  - snpID: chromosom: snp position:reference allel:alternative allel
  - n_miss: Number of samples that have missing genotypes.
  - b: Prior for effect size of corresponding snp
  - beta: Posterior mean estimate for effect size
  - ES: b+beta, totol effect size estimation, which is used in model predition.We only keep snps that have ES$\neq$0.
  - gamma: Indicator variable of whether we have beta estimation.If gamma=0, beta=0.If gamma=1, beta$\neq$0.



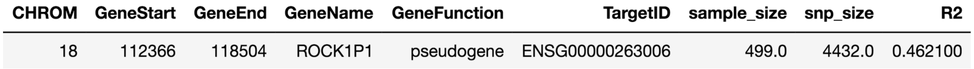
CHROM	snpID	POS	TargetID	n_miss	b	beta	ES	gamma	p_HWE	MAF
18.0	18:69836:A:G	69836.0	ENSG00000263006	0.0	0.000282	0.000011	0.000293	1.0	0.160895	0.193196

- DPR Training Information File



CHROM	GeneStart	GeneEnd	GeneName	GeneFunction	TargetID	sample_size	snp_size	R2
18	112366	118504	ROCK1P1	pseudogene	ENSG00000263006	499.0	4432.0	0.462100

- Prediction File



CHROM	GeneStart	GeneEnd	GeneName	GeneFunction	TargetID	sample_size	snp_size	R2
18	112366	118504	ROCK1P1	pseudogene	ENSG00000263006	499.0	4432.0	0.462100

For association study, explaination of output variables CHROM, GeneStart, GeneEnd, GeneName, GeneFunction and TargetID keep the same as model training and prediction part. Unique output variables for association study are listed as follow.

- Single Phenotype
  - R2 :
  - BETA:
  - BETA_SE :
  - PVALUE :
  - N :
- Multiple Phenotype
  - R2 :
  - F_STAT :
  - F_PVALUE :
  - N :

Example of output files are shown in https://github.com/xmeng34/GEtools/tree/master/src

6 Source Code

- Model Training
  - Elastic Net Model
    - Model Training : Elastic_Net_Train.py
    - Predict transcriptome from a given genotype file : Prediction.py
    - Elastic_Net.sh
  - DPR Model
    - Model Training : DPR_Train.py, call_DPR.sh
    - Predict transcriptome from a given genotype file : Prediction.py
    - DPR.sh
  - TIGAR_Model_Train.sh
- TWAS
  - For $\boldsymbol{asso} = \mathbf{1}$: Asso_Study_01.py
  - For $\boldsymbol{asss} = \mathbf{2}$: Asso_Study_02.sh, summary_stat.py
    - Reference covariance matrix calculation : covar_calculation.py, covar_calculation.sh
  - TIGAR_TWAS.sh

7 Reference

- PrediXcan : `https://github.com/hakyimlab/PrediXcan`
- DPR : `https://github.com/biostatpzeng/DPR`