# 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.
- genofile_type **:** vcf or dosages.
- genofile_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 |

- Format **:** 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 is 0.001. TIGAR will select snps p-value greater than this threshold for training.
- window : Window size around gene boundary, default is $\mathbf{10^6}$BP.
- thread : Number of thread for multiprocessing with default 1. If thread>1, say thread=10, it will run 10 genes simultaneously. In other words, it will accelerate training procedure.
- out : Path of TIGAR 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 uses 5-fold as default.
  - alpha : Ratio for $\boldsymbol{L_1}$ & $\boldsymbol{L_2}$ penalty for elastic net regression, default is 0.5.
- Input only for DPR
  - dpr : $-\boldsymbol{dpr1}$ fits DPR using variation Bayesian algorithm, $-\boldsymbol{dpr2}$ fits DPR using MCMC sampling with fixed number of normal components in mixture prior and $-\boldsymbol{dpr3}$ 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, $\boldsymbol{ES = beta}$. For additive effect size, $\boldsymbol{ES = b + beta}$. Default is fixed.

## 3.2 Prediction

- model : Training imputation model for transcriptomic data (elastic_net or DPR).
- chr : Chromosome number.
- train_result_path : Contains training parameters of each snps. See exact format in **Output** part.
- train_info_path : Contains information of each gene. See exact format in **Output** part.
- genofile_type : vcf or dosages.
- genofile_dir : Genotype data for prediction with vcf or dosages format.This file should be tabixed (contains .gz and .gz.tbi).
- test_sample : A column of sampleIDs use for training.
- Format : Format using for training data(GT or DS).
- window : Window size around gene boundary, default is $\mathbf{10^6}$BP.
- 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.
- thread : Number of thread for multiprocessing with default 1.
- out : Path of TIGAR to save output files.

## 3.3 Association Study

- asso :
  - If $\boldsymbol{asso = 1}$, 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 $\boldsymbol{asso = 2}$, 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 $\mathbf{10^6}$BP.
- 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.
  - genofile_path **:** Folder of Genotype files (tabixed)
  - genofile_type **:** 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 of TIGAR to save output files.

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/sampleID.txt
  - genofile_dir=./example_data/Genotype.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 --genofile_dir ${genofile_dir} \
$ --genofile_type vcf --Format 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 --genofile_dir ${genofile_dir} \
$ --genofile_type vcf --Format DS \
$ --out ${out_prefix}
```

## 4.2 Prediction

- Predicion Inputs
  - genofile_dir=./example_data/Genotype.vcf.gz
  - test_sample_path=./example_data/sampleID.txt
- Based on Elastic Net Regression
  - train_result_path=./Result/elastic_net_CHR1/CHR1_elastic_net_training_param.txt
  - train_info_path=./Result/elastic_net_CHR1/CHR1_elastic_net_training_info.txt

Command Line

```
$ ./TIGAR_Model_Pred.sh --model elastic_net \
$ --chr 1 \
$ --train_result_path ${train_result_path} \
$ --train_info_path ${train_info_path} \
$ --genofile_type vcf \
$ --genofile_dir ${genofile_dir} \
$ --test_sample ${test_sample_path} \
$ --Format DS \
$ --out ${out_prefix}
```

- Based on DPR
  - train_result_path=./Result/DPR_CHR1/CHR1_DPR_training_param.txt
  - train_info_path=./Result/DPR_CHR1/CHR1_DPR_training_info.txt

Command Line

```
$ ./TIGAR_Model_Pred.sh --model DPR \
$ --chr 1 \
$ --train_result_path ${train_result_path} \
$ --train_info_path ${train_info_path} \
$ --genofile_type vcf \
$ --genofile_dir ${genofile_dir} \
$ --test_sample ${test_sample_path} \
$ --Format GT \
$ --out ${out_prefix}
```

## 4.3 Association Study

- Association Study Input ($\boldsymbol{asso = 1}$)
  - Gene_Exp_path=./Result/DPR_CHR1/CHR1_DPR_prediction.txt
  - PED=./example_data/example_PED.ped
  - Asso_Info=./example_data/Asso_Info.txt
  - out_prefix=./Result/DPR_CHR1

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=./Result/DPR_CHR1/CHR1_DPR_prediction.txt
  - Zscore=./example_data/example_Zscore/CHR1_GWAS_Zscore.txt.gz
  - Weight=./Result/DPR_CHR1/CHR1_DPR_trainining_param.txt
  - Covar=./example_data/CHR1_reference_cov.txt.gz
  - out_prefix=./Result/DPR_CHR1

Command Line

```
$ ./TIGAR_TWAS.sh --asso 2 \
$ --Gene_Exp ${Gene_Exp_path} \
$ --Zscore ${Zscore} --Weight ${Weight} --Covar ${Covar} \
$ --chr 1 \
$ --out ${out_prefix}
```

- Reference Covariance Matrix Calculation
  - block=./example_data/block_annoation.txt
  - genofile_path=./example_data
  - out_prefix=./Result/reference_cov

Command Line

```
$ ./covar_calculation.sh --block ${block} \
$ --genofile_path ${genofile_path} --genofile_type vcf \
$ --chr 1 \
$ --Format GT \
$ --out ${out_prefix}
```

## 4.4 Change Default Values

- Model Training
  - To change default value, like alpha and cv for elastic-net model.

Command Line

```
$ ./TIGAR_Model_Train.sh --model elastic_net \
$ --Gene_Exp ${Gene_Exp_path} --train_sample ${train_sample_path} \
$ --chr 1 --genotype_dir ${genotype_dir} \
$ --genofile_type vcf --Format GT \
$ --alpha 0.8 --cv 10 \
$ --out ${out_prefix}
```

- Model Prediction
  - To change default value, say maf_diff in prediction part.

Command Line

```
$ ./TIGAR_Model_Pred.sh --model elastic_net \
$ --chr 1 \
$ --train_result_path ${train_result_path} \
$ --train_info_path ${train_info_path} \
$ --genofile_type vcf \
$ --genofile_dir ${genofile_dir} \
$ --test_sample ${test_sample_path} \
$ --Format DS \
$ --maf_diff 0.1 \
$ --out ${out_prefix}
```

- TWAS
  - Change model from OLS to Logit

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
  - 5-fold-CV-R2 **:** Average cross-validation $\boldsymbol{R^2}$. TIGAR will run 5-fold cross validation before training model with whole training sample. If 5-fold-CV-R2$< \mathbf{0.01}$, TIGAR will assume Elastic-Net or DPR model is not suitable for this gene and skip this gene.
  - TrainPVALUE **:** P-value of F-test for final training model with whole samples.
  - Train-R2 **:** Regression $\boldsymbol{R^2}$ 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
  - ES**:** Effect size estimation based on elastic net regression.
    We only keep snps that have beta$\neq$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 **:** $\boldsymbol{L_1}$ & $\boldsymbol{L_2}$ 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: If $ES = fixed$, $ES = beta$. If $ES = addictive$, $ES = b + beta$. 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 : Regression $R^2$.
  - BETA : Regression coefficient of gene
  - BETA_SE : Standard error of BETA.
  - PVALUE : P-value of F-test for regression model.
  - N : Sample size.
- Multiple Phenotype
  - R2 : Regression $R^2$.
  - F_STAT : Value of F statistics for regression model.
  - F_PVALUE : P-value of F-test.
  - N : Sample size.
- Using summary statistics
  - Zscore : Value of burden Z-score.
  - Pvalue : p-value for chi-square test for Zscore.

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

# 6 Source Code

- Model Training
  - Elastic-Net Model
    - Model Training : Elastic_Net_Train.py
    - Elastic_Net.sh
  - DPR Model
    - Model Training : DPR_Train.py, call_DPR.sh
    - DPR.sh
  - TIGAR_Model_Train.sh
- Prediction
  - Predict transcriptome from a given genotype file : Prediction.py
  - TIGAR_Model_Pred.sh
- TWAS
  - Association Study with Individual-level GWAS data ($\boldsymbol{asso = 1}$) : Asso_Study_01.py
  - Association Study with Summary-level GWAS data ($\boldsymbol{asso = 2}$) : Asso_Study_02.sh, summary_stat.py
    - Reference covariance matrix calculation : covar_calculation.py, TIGAR_Covar.sh
  - TIGAR_TWAS.sh

# 7 Reference

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