TIGAR Manual

Xiaoran Meng
mengxiaoran0629@gmail.com

Emory University

Contents

1 Introduction

"**TIGAR**" stands for **T**ranscriptome-**I**ntegrated **G**enetic **A**ssociation **R**esource, which is developed using Python and BASH scripts. TIGAR can fit both Elastic-Net and nonparametric Bayesian model (Dirichlet Process Regression, i.e. DPR), impute genetically regulated gene expression (GReX) from genotype data, and conduct transcriptome-wide association studies (TWAS) using both individual-level and summary-level GWAS data for univariate and multivariate phenotypes.

2 Installation

- DPR
  - DPR module is saved under folder ./Model_Train_Pred/Functions
  - Please run following command before running DPR in TIGAR

```
Command Line
$ chmod a+x ./Model_Train_Pred/Functions/DPR
```

- Python 3.5
  - dfply : Work similar as dplyr package in R.
  - io : Decode genotype data from TABIX result.
  - subprocess : Read in TABIX result.
  - multiprocessing
- BGZIP : http://www.htslib.org/doc/bgzip.html
- TABIX : http://www.htslib.org/doc/tabix.html

3 TIGAR

3.1 cis-eQTL Effect-Sizes Calculation

Generally, SNPs within 1Mb of the gene boundary will be included in regression model and genetically regulated gene expression (GReX) can be imputed through $\widehat{\boldsymbol{GReX}} = \mathbf{X}_{new}\hat{\mathbf{w}}$ with new genotype data $\mathbf{X}_{new}$.

3.1.1 Elastic-Net Regression (PrediXcan)

Elastic-Net regression method assumes linear regression model as follow:

$$\mathbf{E}_g = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{\sigma^2}) \tag{1}$$

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\mathbf{argmin}}(\|\mathbf{E}_g - \mathbf{X}\mathbf{w}\|_2^2 + \boldsymbol{\lambda}(\boldsymbol{\alpha}\|\mathbf{w}\|_1 + \frac{\mathbf{1}}{\mathbf{2}}(\mathbf{1}-\boldsymbol{\alpha})\|\mathbf{w}\|_2^2)), \boldsymbol{\alpha} \in [\mathbf{0}, \mathbf{1}] \tag{2}$$

$\mathbf{E}_g$ represent gene expression level for specific gene g, usually corrected for confounding covariates like age, gender and genotype principle components. $\mathbf{X}$ is the genotype matrix, $\mathbf{w}$ denotes effect-size vector of corresponding SNPs and $\boldsymbol{\epsilon}$ is the error term. In this model, cis-eQTL effect-size $\mathbf{w}$ is estimated by adding a mixture of LASSO ($\boldsymbol{L_1}$) and Ridge ($\boldsymbol{L_2}$) penalties, where $\boldsymbol{\alpha}$ denotes proportion of $\boldsymbol{L_1}$ and $\boldsymbol{L_2}$ penalty and $\boldsymbol{\lambda}$ is the penalty parameter. Specifically, PrediXcan assumes $\boldsymbol{\alpha} = \mathbf{0.5}$ and picks $\boldsymbol{\lambda}$ by 5-folds cross-validation.

3.1.2 Dirichlet Process Regression (DPR)

The linear regression model is quite similar as (1). According to latent Dirichlet process regression, the model assumes

$$\mathbf{E}_g = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{\sigma^2}), \boldsymbol{\sigma^2} \sim \boldsymbol{IG}(\boldsymbol{a_\epsilon}, \boldsymbol{b_\epsilon}) \tag{3}$$

$$\boldsymbol{w_i} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{\sigma_w^2}), \boldsymbol{\sigma_w^2} \sim \boldsymbol{D}, \boldsymbol{D} \sim \boldsymbol{DP}(\boldsymbol{ID}(\boldsymbol{a}, \boldsymbol{b}), \boldsymbol{\xi}) \tag{4}$$

Where $\boldsymbol{w_i}$ denotes effect-size for each SNP within in gene g, which follows a normal distribution with mean 0 and variance $\boldsymbol{\sigma^2}$ with Dirichlet process prior D that has base distribution inverse gamma IG(a,b) and concentration parameter $\boldsymbol{\xi}$. After integrating out latent variable $\boldsymbol{\sigma^2}$, an equivalent non-parametric prior distribution of $\boldsymbol{w_i}$ can be driven as follow:

$$\boldsymbol{w_i} \sim \sum_{\boldsymbol{k}=\mathbf{1}}^{+\infty} \boldsymbol{\pi_k} \boldsymbol{N}(\mathbf{0}, \boldsymbol{\sigma_k^2}), \boldsymbol{\sigma_k^2} \sim \boldsymbol{IG}(\boldsymbol{a_k}, \boldsymbol{b_k}), \boldsymbol{\pi_k} = \boldsymbol{v_k} \prod_{\boldsymbol{l}=\mathbf{1}}^{\boldsymbol{k}-\mathbf{1}} (\mathbf{1} - \boldsymbol{v_l}), \boldsymbol{v_k} \sim \boldsymbol{Beta}(\mathbf{1}, \boldsymbol{\xi}) \tag{5}$$

Here, $\boldsymbol{\xi}$ means the same concentration parameter in (3) with a hyper prior $\boldsymbol{\xi} \sim \boldsymbol{Gamma}(\boldsymbol{a_\xi}, \boldsymbol{b_\xi})$. DPR model is more robust in detect gene structure due to non-informative prior for $\boldsymbol{\sigma_k^2}$,$\boldsymbol{\sigma^2}$ and $\boldsymbol{\xi}$, which usually assumes $\boldsymbol{a_k}$, $\boldsymbol{b_k}$,$\boldsymbol{a_\epsilon}$ and $\boldsymbol{b_\epsilon}$ as 0.1 and $(\boldsymbol{a_\xi}, \boldsymbol{b_\xi})$ as (1,0.1), then $\boldsymbol{\sigma_k^2}$, $\boldsymbol{\sigma^2}$ and $\boldsymbol{\xi}$ can be estimated through data and make $\boldsymbol{w_i}$ data-driven.

3.2 TWAS

3.2.1 Type One Association Study

With given weight (SNP effect-sizes) $\mathbf{w}$, individual genotype $\mathbf{X}_{new}$, phenotype Y and covariance matrix $\mathbf{C}$, the association test of $\widehat{\mathbf{GReX}}$ and $\mathbf{Y}$ is conducted through linear regression model as follow

- Single Phenotype
  - Association test for single phenotype and imputed GReX is conducted through model as follow
$$\boldsymbol{f}(\boldsymbol{E}[\boldsymbol{Y}|\mathbf{X},\mathbf{C}]) = \boldsymbol{\eta}\mathbf{C} + \boldsymbol{\beta}\widehat{\mathbf{GReX}} \tag{6}$$
  - $\boldsymbol{f}(\cdot)$ is a pre-specified function and $\boldsymbol{H_0}: \boldsymbol{\beta} = \mathbf{0}$ is the same with gene-based association test.
- Multivariate Phenotype
  - Association test for multivariate phenotype (number of Y $> \mathbf{1}$)and imputed GReX is conducted through model as follow
$$\begin{aligned} &\boldsymbol{Y_j} = \boldsymbol{\eta}\mathbf{C} + \boldsymbol{\epsilon}, \boldsymbol{j} = \mathbf{1}, \mathbf{2}, ..., \boldsymbol{n} \\ &\widetilde{\boldsymbol{Y}}_{\boldsymbol{j}} = \boldsymbol{Y_j} - \hat{\boldsymbol{Y}}_{\boldsymbol{j}} \end{aligned} \tag{7}$$
$$\widehat{\boldsymbol{GReX}}_{\boldsymbol{g}} = \sum_{\boldsymbol{j}=\mathbf{1}}^{\boldsymbol{n}} \boldsymbol{\beta_j}\widetilde{\boldsymbol{Y}}_{\boldsymbol{j}} + \boldsymbol{\epsilon} \tag{8}$$
  - Here $\boldsymbol{Y_j}, \boldsymbol{j} = \mathbf{1}, \mathbf{2}, ..., \boldsymbol{n}$ represent n different phenotypes and C is a covariance matrix. In (8), TIGAR first adjust for covariates by calculating residual $\widetilde{\boldsymbol{Y}}_{\boldsymbol{j}}, \boldsymbol{j} = \mathbf{1}, \mathbf{2}, ..., \boldsymbol{n}$ for each phenotype. Association study is conducted base on $\boldsymbol{R^2}$ from (9), which is the same as $\boldsymbol{H_0}: \boldsymbol{R^2} \neq \mathbf{0}$.

3.2.2 Type Two Association Study

- TIGAR can run association test through summary-level data when new genotype data is not provided. Let $\mathbf{Z}$ represent single-variance test for all cis-SNPs. Burden Z-score of association test is defined as
$$\widetilde{\boldsymbol{Z}} = \frac{\mathbf{Z}\hat{\mathbf{w}}}{\sqrt{\hat{\mathbf{w}}^T\mathbf{V}\hat{\mathbf{w}}}} \tag{9}$$
- Here, $\mathbf{V}$ denotes covariance matrix across SNPs, which TIGAR can calculated through original genotype data.

4 Input

- Example input files provided here are generated artificially.
  (Except ./example_data/block_annotation_EUR.txt)
- All input files are **tab delimited** text files.
- Example of input files are provided in
  https://github.com/xmeng34/TIGAR/tree/master/example_data.

4.1 cis-eQTL Effect-Sizes Calculation

Script : ***TIGAR_Model_Train.sh***
(Output Files : $*$_training_weight.txt[6.1.1], $*$_training_info.txt[6.1.2])

- model : **elastic_net** or **DPR**. Training imputation model for transcriptomic data.
  - elastic_net : Elastic-Net Regression Model
  - DPR : Dirichlet Process Regression Model
- Gene_Exp : Combination of gene annotation and expression file.
  - Example File : ***./example_data/Gene_Exp.txt***
  - First 5 columns specify
    * Chromosome Number (CHROM)
    * Gene Starting Position (GeneStart)
    * Gene Ending Position (GeneEnd)
    * Target Gene ID (TargetID/GeneID)
    * Gene Name (GeneName, optional, could be the same as Gene ID.)
  - Sample gene expression data start from the 6^{th} column.

CHROM	GeneStart	GeneEnd	TargetID	GeneName	GTEX-111FC	GTEX-1128S
1	196621008	196716634	ENSG00000000971.11	CFH	-2.643948	-2.330538

- sampleID : A column of sampleIDs use for training, which is based on genotype file provided here.
  - Example File : ***./example_data/sampleID.txt***
- chr : Chromosome Number
- genofile_type : **vcf** or **dosages**.

- genofile : Genotype Data (Tabixed)
  - Sorted by chromosome and base pair position, zipped by **bgzip**, and **tabix**.
  - Example tabix command

```
Command Line
$ tabix -f -p vcf *.vcf.gz
```

  - vcf
    * Example File : ***./example_data/Genotype/example.vcf.gz***
    * Vcf genotype data start from the 10^{th} column.
    * More information about vcf file format : vcf Format Link

CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT
1	22346009	1_22346609_G_A_b37	G	A	.	PASS	AF=0.1617;MAF=0.1617;R2=0.61796	GT:DS

  - dosages
    * The first 5 columns are of the same format as vcf file.
    * Dosage genotype data start from the 6^{th} column.

CHROM	POS	ID	REF	ALT	GTEX-1117F	GTEX-111CU
1	22346009	1_22346009_G_A_b37	G	A	0.93	0.31

- Format : **GT** or **DS**. Format using for genotype data. For example, $0|0$ or $0/0$ stands for GT format, 0.31 stands for DS format.
- maf : Threshold for Minor Allele Frequency (range from 0 to 1) with 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 with default **0.001**. TIGAR will select snps with test 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, say thread=10, TIGAR will run data for 10 genes simultaneously, which will accelerate training procedure.
- out : Path of TIGAR to save output files.

4.1.1 Elastic-Net Regression Only

  - cv : Number of folds used in cross-validation to select parameter for Elastic-Net regression. TIGAR uses **5**-folds as default.
  - alpha : Ratio for L_1 and L_2 penalty for Elastic-Net regression. Default is **0.5**.

4.1.2 DPR Only

- dpr : **1** or **2** or **3**. Model used in DPR with default **1**.
  - $--$ ***dpr*** **1** fits DPR using variation Bayesian algorithm.
  - $--$ ***dpr*** **2** fits DPR using MCMC sampling with fixed number of normal components in mixture prior.
  - $--$ ***dpr*** **3** fits DPR using MCMC sampling with adaptively selected number of the normal components in mixture prior.
- ES : **fixed** or **additive**. Effect-Size with default **fixed**.
  - b : Prior for effect-size of corresponding snp
  - beta : Posterior mean estimate for effect-size
  - For $--$ ***ES fixed***, ES=beta.
  - For $--$ ***ES additive***, ES=b+beta.

4.2 GReX Prediction

Script : **TIGAR_Model_Pred.sh** (Output File : $*$_GReX_prediction.txt[6.2])

- model : **elastic_net** or **DPR**. Training imputation model used for transcriptomic data.
- chr : Chromosome Number
- train_weight_path : Contains training parameters of each snps. See exact format in **Output** ($*$_training_weight.txt[6.1.1]).
- train_info_path : Contains information of each gene. See exact format in **Output** ($*$_training_info.txt[6.1.2]).
- genofile_type : **vcf** or **dosages**.
- genofile : Genotype data for prediction with vcf or dosages format.
- sampleID : A column of sampleIDs use for prediction, which is based on genotype file provided here.
- Format : **GT** or **DS**. Format using for Genotype Data.
- window : Window size around gene boundary. Default is 10^6BP.
- maf_diff : Threshold of difference between maf calculated in **cis-eQTL Effect-Sizes Calculation** [4.1] and **GReX Prediction** [4.2]. TIGAR will select snps with difference less than this threshold for prediction. Default is **0.2**.
- thread : Number of thread for multiprocessing with default **1**.
- out : Path of TIGAR to save output files.

4.3 TWAS

Script : **TIGAR_TWAS.sh**

- asso : **1** or **2**. Method of association study.
- thread : Number of thread for multiprocessing with default **1**.
- out : Path for TIGAR to save output files.

4.3.1 Type One Association Study

Running TWAS with individual-level GWAS data.
(Output Folder : TIGAR_TWAS_Type_One[6.3.1])

- Gene_Exp : Predicted gene expression data. See exact format in **Output** (∗_GReX_prediction.txt[6.2]).
- PED : PED file
  - Example File : ***./example_data/example_PED.ped***
  - More information bout PED file format: http://zzz.bwh.harvard.edu/plink/data.shtml#ped
- Asso_Info : Instruction for association study.
  - Example File : ***./example_data/Asso_Info/Asso_Info_∗.txt***
  - The variables specified in this file will be used in TWAS.
  - Two columns with the first column specifying the Phenotype (P) and Covariate variables (C) from the PED file, and the second column specifying the corresponding variable names in the PED file.
- method : **OLS** or **Logit**. Link Function with default **OLS**.
  - OLS stands for ordinary least square regression.
  - Logit stands for logistic regression.
  - TIGAR only uses **OLS** for Multivariate Phenotype (Number of phenotype >1).

4.3.2 Type Two Association Study

- Running TWAS with summary-level GWAS data
  (Output Folder : TIGAR_TWAS_Type_Two[6.3.2])
  - Gene_Exp : Combination of gene annotation and expression level file. The same format in **cis-eQTL Effect-Sizes Calculation** [4.1].
  - Weight : File contains snp effect-sizes.
    - File specifying chromosome number, base pair position, reference allele, alternative allele, and target gene ID.
    - See example file in **Output** (∗_training_weight.txt[6.1.1]).

* Zscore : Z-score file from previous GWAS study (Tabixed).
  · Example File : ***./example_data/example_Zscore/∗_GWAS_Zscore.txt.gz***
  · Sorted by chromosome and base pair position, zipped by **bgzip**, and tabixed.
  · The first 4 columns are of the same format as **Weight** input.

CHROM	POS	REF	ALT	Zscore
1	22346158	G	A	0.195928

* Covar : Reference covariance matrix (Tabixed File. Scripts are provided under folder ***./TWAS/Covar***).
* chr : Chromosome Number.
* window : Window size around gene boundary. Default is 10^6BP.

– Reference Covariance Matrix Calculation
* block : Genome block annotation file.
  · Example File : ***./example_data/block_annotation_EUR.txt***
  · CHROM : Chromosome Number.
  · Start : Block Starting Position.
  · End : Block Ending Position.
  · Boundaries (CHROM, Start, End) in this example file are based on the LD structure of European samples, which can be used directly when calculating user's own covariance matrix.
  · Reference genotype files shall be of one per chromosome, or one for the whole genome-wide variants.
  · Block annotation files of other ethnicities can be adopted from the genome segmentation generated by **LDetect** : https://bitbucket.org/nygcresearch/ldetect-data/src/master/.
* genofile_type : **vcf** or **dosages**.
* genofile : Genotype data for calculating reference covariance matrix (Tabixed).
* chr : Chromosome Number.
* Format : **GT** or **DS**. Format using for Genotype Data.
* maf : Threshold for Minor Allele Frequency (range from 0-1). Default is **0.05**.

5 Example Usage

5.1 cis-eQTL Effect-Sizes Calculation

- Gene_Exp_path=./example_data/Gene_Exp.txt
- sampleID=./example_data/sampleID.txt
- genofile=./example_data/Genotype/example.vcf.gz
- out_prefix=./Result

5.1.1 Elastic-Net Regression

Command Line

```
$ cd TIGAR
$ ./TIGAR_Model_Train.sh --model elastic_net \
$ --Gene_Exp ${Gene_Exp_path} --sampleID ${sampleID} \
$ --chr 1 --genofile_type vcf \
$ --genofile ${genofile} --Format GT \
$ --out ${out_prefix}
```

5.1.2 DPR

Command Line

```
$ cd TIGAR
$ ./TIGAR_Model_Train.sh --model DPR \
$ --Gene_Exp ${Gene_Exp_path} --sampleID ${sampleID} \
$ --chr 1 --genofile_type vcf \
$ --genofile ${genofile} --Format GT \
$ --out ${out_prefix}
```

5.2 GReX Prediction

- genofile=./example_data/Genotype/example.vcf.gz
- sampleID=./example_data/sampleID.txt
- out_prefix=./Result

5.2.1 Elastic-Net Regression Based

- train_weight_path=./Result/elastic_net_CHR1/CHR1_elastic_net_training_weight.txt
- train_info_path=./Result/elastic_net_CHR1/CHR1_elastic_net_training_info.txt

Command Line

```
$ cd TIGAR
$ ./TIGAR_Model_Pred.sh --model elastic_net \
$ --chr 1 \
$ --train_weight_path ${train_weight_path} \
$ --train_info_path ${train_info_path} \
$ --genofile_type vcf \
$ --genofile ${genofile} \
$ --sampleID ${sampleID} \
$ --Format GT \
$ --out ${out_prefix}
```

5.2.2 DPR Based

- train_weight_path=./Result/DPR_CHR1/CHR1_DPR_training_weight.txt
- train_info_path=./Result/DPR_CHR1/CHR1_DPR_training_info.txt

Command Line

```
$ cd TIGAR
$ ./TIGAR_Model_Pred.sh --model DPR \
$ --chr 1 \
$ --train_weight_path ${train_weight_path} \
$ --train_info_path ${train_info_path} \
$ --genofile_type vcf \
$ --genofile ${genofile} \
$ --sampleID ${sampleID} \
$ --Format GT \
$ --out ${out_prefix}
```

5.3 TWAS

5.3.1 Type One Association Study

- Gene_Exp_path=./Result/DPR_CHR1/CHR1_DPR_GReX_prediction.txt
- PED=./example_data/example_PED.ped
- Asso_Info=./example_data/Asso_Info/Asso_Info_SinglePheno_OLS.txt
- out_prefix=./Result/DPR_CHR1

Command Line

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

5.3.2 Type Two Association Study

- Gene_Exp_path=./example_data/Gene_Exp.txt
- Zscore=./example_data/example_Zscore/CHR1_GWAS_Zscore.txt.gz
- Weight=./Result/DPR_CHR1/CHR1_DPR_training_weight.txt
- Covar=./Result/reference_cov/CHR1_reference_cov.txt.gz
- out_prefix=./Result/DPR_CHR1

Command Line

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

5.3.3 Reference Covariance Matrix Calculation

- block=./example_data/block_annotation_EUR.txt
- genofile=./example_data/Genotype/example.vcf.gz
- out_prefix=./Result

Command Line

```
$ cd TIGAR
$ ./TWAS/Covar/TIGAR_Covar.sh --block ${block} \
$ --genofile_type vcf --genofile ${genofile} \
$ --chr 1 \
$ --Format GT \
$ --out ${out_prefix}
```

5.4 Change Default Values

- cis-eQTL Effect-Size Calculation
  - Change values of alpha and cv for Elastic-Net model.

Command Line

```
$ cd TIGAR
$ ./TIGAR_Model_Train.sh --model elastic_net \
$ --Gene_Exp ${Gene_Exp_path} --sampleID ${sampleID} \
$ --chr 1 --genofile_type vcf \
$ --genofile ${genofile} --Format GT \
$ --alpha 0.8 --cv 10 \
$ --out ${out_prefix}
```

- GReX Prediction
  - Change value of maf_diff in prediction.

Command Line

```
$ cd TIGAR
$ ./TIGAR_Model_Pred.sh --model elastic_net \
$ --chr 1 \
$ --train_weight_path ${train_weight_path} \
$ --train_info_path ${train_info_path} \
$ --genofile_type vcf \
$ --genofile ${genofile} --Format GT \
$ --sampleID ${sampleID} \
$ --maf_diff 0.1 \
$ --out ${out_prefix}
```

- TWAS
  - Change value of model from OLS to Logit
  - Asso_Info=./example_data/Asso_Info/Asso_Info_SinglePheno_Logit.txt

Command Line

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

6 Output

- All output files are **tab delimited** text files.
- Example of output files are shown in https://github.com/xmeng34/TIGAR/tree/master/Result

6.1 cis-eQTL Effect-Size Calculation

6.1.1 Training Weight Files

File : ∗**_training_weight.txt**

- CHROM : Chromosome Number
- POS : Snp Position
- REF : Reference Allele
- ALT : Alternative Allele
- TargetID : GeneID
- ES : Estimated effect-size. Only keep $\boldsymbol{ES \neq 0}$.
- MAF : Minor Allele Frequency (range from 0-1).
- p_HWE: P-value of Hardy Weinberg Equilibrium Exact Test for this snp.
- Elastic-Net Training Weight File
  - ID : rsID

CHROM	POS	REF	ALT	TargetID	ID	ES	MAF	p_HWE
1	195731020	T	C	ENSG00000000971.11	1_195731020_T_C_b37	-0.05166	0.299302	0.485607

- DPR Training Weight File
  - n_miss: Number of samples that have missing genotypes.
  - b: Prior for effect-size of corresponding snp.
  - beta: Posterior mean estimate for effect-size.
  - gamma: Indicator variable for whether DPR has beta estimation.
    - If gamma=0, then beta=0.
    - If gamma=1, then beta $\neq$ 0.

CHROM	POS	REF	ALT	TargetID	n_miss	b	beta	gamma	ES	MAF	p_HWE
1	195621022	G	A	ENSG00000000971.11	0	-0.000859	-0.000043	1	-0.000043	0.641167	0.343892

6.1.2 Training Information

File : ∗**_training_info.txt**

- CHROM : Chromosome Number
- GeneStart : Gene Starting Position
- GeneEnd : Gene Ending Position
- TargetID : Target Gene ID
- GeneName : Gene Name
- snp_size : Number of snps used for regression.
- effect_snp_size : Number of snps that have nonzero ($\boldsymbol{ES} \neq \mathbf{0}$) effect-size.
- sample_size : Number of sampleIDs used.
- 5-fold-CV-R2 : Average cross-validation $\boldsymbol{R^2}$.
  - TIGAR will run 5-folds cross-validation before training model with whole samples.
  - 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.
- Elastic-Net Training Information File
  - k-fold : Folds we use for cross-validation (ex.5-folds)
  - alpha : $\boldsymbol{L_1}$ and $\boldsymbol{L_2}$ 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	TargetID	GeneName
1	196621008	196716634	ENSG00000000971.11	CFH

snp_size	effect_snp_size	sample_size	5-fold-CV-R2	TrainPVALUE	Train-R2	k-fold	alpha	Lambda	cvm
5784	38	128	0.02819	0.000007	0.084624	5	0.5	1.0	-0.110839

- DPR Training Information File

CHROM	GeneStart	GeneEnd	TargetID	GeneName	snp_size	effect_snp_size	sample_size	5-fold-CV-R2	TrainPVALUE	Train-R2
1	196621008	196716634	ENSG00000000971.11	CFH	4515	4515	128	0.019943	0.000011	0.142995

6.2 GReX Prediction

- File : ∗**_GReX_prediction.txt**
- CHROM, GeneStart, GeneEnd, TargetID and GeneName share the same explanation in **Training Information** [6.1.2] file.
- Predicted sample gene expression data start from the 6^{th} column.

CHROM	GeneStart	GeneEnd	TargetID	GeneName	GTEX-111FC	GTEX-1128S
1	196621008	196716634	ENSG00000000971.11	CFH	-0.044587	0.010128

6.3 TWAS

CHROM, GeneStart, GeneEnd, TargetID and GeneName share the same explanation in **Training Information** [6.1.2] file.

6.3.1 Type One Association Study

Results are saved under folder **TIGAR_TWAS_Type_One**

- Single Phenotype (association_study_Single_∗.txt)
  - R2 : Regression $\boldsymbol{R^2}$.
  - BETA : Regression coefficient of Gene.
  - BETA_SE : Standard error of BETA.
  - T_STAT : T-test Statistics for corresponding Gene.
  - PVALUE : T-test P-value for corresponding Gene.
  - N : Sample Size.

CHROM	GeneStart	GeneEnd	TargetID	GeneName	R2	BETA	BETA_SE	T_STAT	PVALUE	N
1	196621008	196716634	ENSG00000000971.11	CFH	0.103382	0.95573	5.561181	0.171857	0.863839	128

- Multivariate Phenotype (association_study_Multi_∗.txt)
  - R2 : Regression $\boldsymbol{R^2}$.
  - F_STAT : Value of F statistics for Regression Model.
  - F_PVALUE : P-value of F-test.
  - N : Sample Size.

CHROM	GeneStart	GeneEnd	TargetID	GeneName	R2	F_STAT	F_PVALUE	N
1	196621008	196716634	ENSG00000000971.11	CFH	0.00671	0.42559	0.65432	128

6.3.2 Type Two Association Study

- Summary Statistics
  - Results are saved under folder **TIGAR_TWAS_Type_Two**
  - ZSCORE **:** Value of Burden Z-score.
  - PVALUE **:** P-value for chi-square test for burden Z-score.

CHROM	GeneStart	GeneEnd	TargetID	GeneName	ZSCORE	PVALUE
1	196621008	196716634	ENSG00000000971.11	CFH	-0.09699	0.922735

- Reference Covariance Matrix
  - Results are saved under folder **reference_cov** (∗_reference_cov.txt.gz).
  - CHROM, POS, ID, REF and ALT share the same explanations in **Training Weight** [6.1.1] file.
  - COV **:** A string of covariance for corresponding snp with other snps within the same block. TIGAR only records upper triangle of the covariance matrix.

CHROM	POS	ID	REF	ALT	COV
1	12041	5_12041_A_T_b37	A	T	0.37715874458643,0.3699702365039476,0.35080354...

7 Source Code

- cis-eQTL Effect-Sizes Calculation
  - Elastic-Net Regression : Elastic_Net_Train.py
  - DPR : DPR_Train.py, call_DPR.sh
  - TIGAR_Model_Train.sh
- GReX Prediction
  - Predict GReX from a given genotype file : Prediction.py
  - TIGAR_Model_Pred.sh
- TWAS
  - Association Study : TWAS.py
    - Reference covariance matrix calculation : covar_calculation.py, TIGAR_Covar.sh
  - TIGAR_TWAS.sh

8 Reference

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