**Reproducibility Form for**

“**Retrospective Association Analysis of Longitudinal Binary Traits Identifies Important Loci and Pathways in Cocaine Use”**

## Datasets

**Abstract**

This dataset is cocaine use genome-wide association study data from the Veterans Aging Cohort Study (VACS). Longitudinal cocaine-use in patient surveys was collected at six clinic visits on 2,470 participants who provided a blood sample. Among them, 69.8% are African Americans (AAs), 19.3% are European Americans (EAs), and 10.9% are of other races.

**Availability**

The genome-wide genotype data and longitudinal cocaine use phenotypes will be deposited to the database of Genotypes and Phenotypes (dbGap) if this manuscript is accepted.

**Description**

Our use of the VACS data was approved by the Yale Human Research Protection Program and the Institutional Review Board of the Veterans Affairs Connecticut Healthcare System. This dataset will be deposited to dbGap if the manuscript is accepted. The genotype data is in plink binary format and the phenotype data are stored in .csv files.

**Optional Information**

Unique identifier / DOI will be available once we deposit the dataset.

## Software Package

**Abstract (Mandatory)**

R package for Retrospective Association Test for Longitudinal Binary Traits

**Description (Mandatory)**

The R package for L-BRAT and RGMMAT will be available at <https://github.com/ZWang-Lab/LBRAT>. The description of the R package is as follows:

Package: LBRAT

Type: Package

Title: Retrospective Association Test for Longitudinal Binary Traits

Version: 0.0.0.9000

Author: Weimiao Wu and Zuoheng Wang

Maintainer: Weimiao Wu<weimiao.wu @yale.edu>

License: GPL (>= 2)

**Optional Information (complete as necessary)**

Requires access to cluster that uses slurm job scheduler.

Depends: R (>= 2.12.1), R libraries: SKAT, mvtnorm, snpStats, snowfall

## Instructions for Use

**Reproducibility**

The following instructions can be used to reproduce Table 1 (Type I error table) and Figure 1 (Empirical power plot) of the simulation studies.

**R package Description**

The LBRAT\_0.0.0.9000.tar.gz file is the R package for L-BRAT and RGMMAT. A more detailed description of each function can be found in the manual of the R package on the Github (<https://github.com/ZWang-Lab/LBRAT>). To install the R package, run the following command in terminal.

1. git clone https://github.com/ZWang-Lab/LBRAT.git
2. R CMD INSTALL LBRAT

Alternatively, one can install the package in R using following R scripts

1. library(devtools)
2. install\_github("ZWang-Lab/LBRAT")

**Folder Setup**

Please set up the folders as follows to replicate the simulation results.

* **type\_I/** folder contains scripts to replicate the type I error simulations and to summarize results into a type I error table.
  + **type1\_lbrat.R** – This is an R script that generates one set of phenotypes and 1,000 SNPs and performs the L-BRAT and RGMMAT tests.
  + **slurm\_summary.sh** – This is a shell script for submitting a job to summarize results of all models and tests.
  + **summaryjobs.sh** – This is a shell script that goes into each directory of the trait model and ascertainment scheme and calls summaryglm.R and summarygee.R to summarize results.
  + **summaryglm.R/summarygee.R** – This is needed to be included in the type\_I/ directory.
* **power/** folder contains scripts to replicate the power simulations and to summarize results into a power plot figure.
  + **power\_lbrat.R** -- This is an R script that generates one set of phenotypes and 2 causal SNPs at the gamma level {0.3, 0.32, 0.34, 0.36 and 0.38} and performs the L-BRAT and RGMMAT tests.
  + **slurm\_summary.sh** – This is a shell script for submitting a job to summarize results of all models and tests.
  + **summaryjobs.sh** – This is a shell script that goes into each directory of the trait model and ascertainment scheme and calls summarypower.R and summarypowerglm.R to summarize results into an output file that can be used as an input for power\_plot.R**.**
  + **summarypower.R/** **summarypowerglmm.R** – This is needed to include in the power/ directory.
  + **power\_plot.R** – Read lbrat\_power.txt and generate Figure 1.

**Workflow**

1. **Table 1 reproduction:**

To reproduce Table 1, follow the three steps listed below:

*Step 1: Generate 106 simulation replicates and calculate test statistics and P-values*

1. cd type\_I/
2. sbatch ../slurm\_1000jobs.sh type1\_lbrat.R

This submits an array of slurm jobs to run the type I error simulations. It creates six subdirectories with the names logistic\_sum, logistic\_random, logistic\_baseline, liability\_sum, liability\_random, and liability\_baseline. Each folder contains 1000 .RData files and each file contains 1000 type I tests results, which gives a total of 106 test results for each of the phenotype and ascertainment model combination.

*Step 2: Summarize results of the 106 simulation replicates*

1. sbatch slurm\_summary.sh #jobID

This summarizes type I error results under each of the six directories and creates summary level .RData files with prefix “summary-“ followed by the test name and the jobID. For example, “summary-glmm6565959.RData”.

*Step 3: Create Type I error table*

1. Rscript type\_1\_table.R

This pulls information over .RData from all six directories and creates a ‘type1\_table.csv’ file which is the Type I error table.

1. **Figure 1 reproduction:**

To reproduce Figure 1, follow the three steps listed below:

*Step 1: Generate 1,000 simulation replicates and calculate test statistics and P-values*

1. cd power/
2. sbatch ../slurm\_1000jobs.sh power\_lbrat.R

This submits an array of slurm jobs to run the power simulations. It creates six subdirectories with the names logistic\_sum, logistic\_random, logistic\_baseline, liability\_sum, liability\_random, and liability\_baseline. Each folder contains 5,000 RData objects with the name indicating it’s corresponding gamma level, test name, and jobID. For example, mixture\_gamma\_0.3\_gee6566477-989.RData.

*Step 2: Summarize results of the 1,000 simulation replicates at* *each gamma value*

1. sbatch slurm\_summary.sh #jobID

This summarizes power results under each of the six directories and creates summary level .RData files with prefix “summary-“ followed by the gamma level and test name and the jobID. For example, ‘gamma\_0.3\_glmm6566477-987.RData’ and also a ‘power.txt’ file that summarizes empirical power for each test at each gamma level.

*Step 3: Create Figure 1*

1. Rscript power\_plot.R

This reads the lbrat\_power.txt file (which is formatted from power.txt) and generates ‘fig1.tiff’.

**Replication**

The following scripts is an example of using the L-BRAT in real data analysis. Given longitudinally measured phenotype data stored in file ‘pheno.csv’, time-dependent or independent covariate data stored in file “cov.csv”, and genotype plink binary files: ‘plink.bim’, ‘plink.fam’ and ‘plink.bed’, the following code can be used to perform genome-wide association analysis using L-BRAT and RGMMAT

1. library(LBRAT)
2. y.file  <-'pheno.csv'
3. cov.file  <- 'cov.csv'
4. plink.file <- 'plink'
5. p0 <- lbrat\_read\_phe(y.file, cov.file,plink.file)
6. File.Bim <- ‘plink.bim’; File.Fam <- ‘plink.fam’; File.Bed <-‘plink.bed’; File.SetID<-‘plink.SetID’; File.SSD = ‘plink.SSD’; File.Info <-‘plink.SSD.Info’
7. gen\_batch\_SetID(File.Bim, File.SetID)
8. Generate\_SSD\_SetID(File.Bed, File.Bim, File.Fam, File.SetID, File.SSD, File.Info)
9. gee0 = lbrat\_est.gee(p0$phe.long, p0$phe.time, p0$phe.cov, timecov = T, corstr = 'ar1')
10. lbrat\_result =lbrat.SSD.All(SSD.INFO, gee0)
11. glmm0 = lbrat\_est.glmm(p0$phe.long, p0$phe.time, p0$phe.cov, timecov = T)
12. rgmmat\_result =.SSD.All(SSD.INFO, glmm0)

## Notes

The dataset will be deposited to dbGap (<https://www.ncbi.nlm.nih.gov/gap>) if the manuscript is accepted.