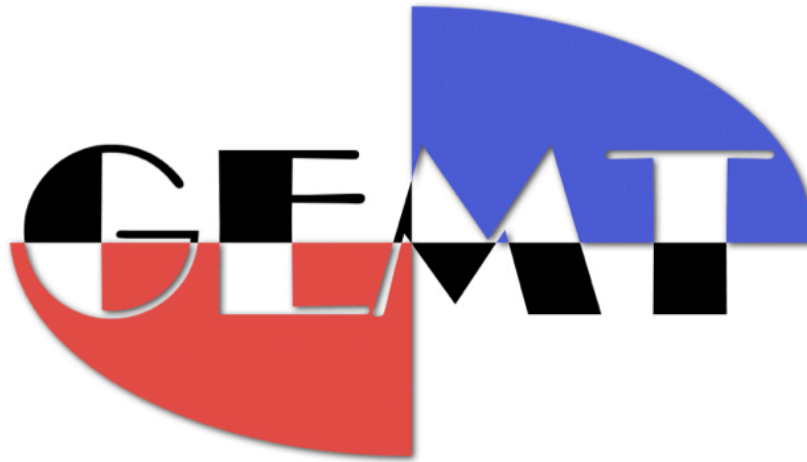


User Manual for



Genomic Environments and Multiple Traits association analysis tool

(Version 1.01)

Last updated on November 4, 2017

Zhiwu Zhang Laboratory
 *For Statistical Genomics*
ZZLab.Net

Disclaimer: While extensive testing has been performed by Zhiwu Zhang Lab at Washington State University, results are, in general, reliable, correct or appropriate. However, results are not guaranteed for any specific set of data. We strongly recommend that users validate GEMT results with other software packages, such as GAPIT, and FarmCPU,BLINK.

Support documents: Extensive support documents, including this user manual, source code, demonstration scripts, data, and results, are available at GEMT website at Zhiwu Zhang Laboratory: <http://zzlab.net/GEMT>

Questions and comments: Users and developers are recommended to post questions and comments at GAPIT forum: <https://groups.google.com/forum/#!forum/GEMT>. Answers from other users and developers are appreciated. The GEMT team members will periodically go through these questions and comments and address them accordingly.

The GEMT project is partially supported by USDA, DOE, NSF, the Agricultural Research Center at Washington State University, and Washington Grain Commission



Contents

<u>1</u>	<u>INTRODUCTION</u>	<u>4</u>
<u>2</u>	<u>GETTING STARTED</u>	<u>4</u>
2.1	OPEN COMMAND LINE WINDOW	4
2.2	DOWNLOAD GEMT	5
2.3	DOWNLOAD INPUT FILES	5
2.4	RUN GEMT	5
<u>3</u>	<u>PHENOTYPE FILE</u>	<u>6</u>
3.1	FORMAT	6
3.2	MISSING VALUES	6
3.3	COVARIANCE FORMAT	7
<u>4</u>	<u>GWAS</u>	<u>7</u>
4.1	WORKING WITH DIFFERENT GENOTYPE FORMATS	7
4.2	CHANGING OUTPUT FILE NAME	8
<u>5</u>	<u>ADVANCED OPERATIONS</u>	<u>8</u>
5.1	MEMORY SAVING	8
5.2	PARALLEL COMPUTATION	8
5.3	SPECIFYING A TRAIT	8
5.4	OPTIMIZATION	8
5.5	RUN GEMT FROM R	9
<u>6</u>	<u>Q&A</u>	<u>10</u>
1.	WHY DO I USE GEMT?	10
2.	HOW DO I CITE GEMT?	10
<u>7</u>	<u>REFERENCES</u>	<u>10</u>

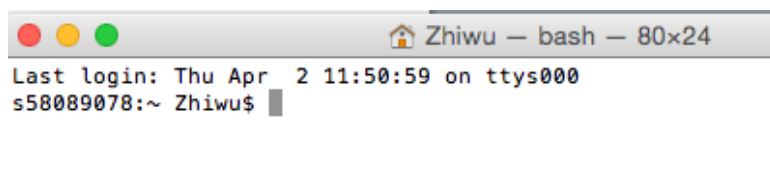
1 Introduction

The performance of computing tools for genome-wide association studies (GWAS) are measured by their computing speed, memory requirements, and statistical power¹. These three factors are determined by the statistical methods a tool implemented and how these methods are engineered to make full use of computer hardware resources. We developed a computing tool named BLINK (www.zzlab.net/blink) that implements a new statistical method. GEMT (Genomic Environments and Multiple Traits association analysis tool) was used to detect intereaction SNPs between G&E or Multiple Traits. GEMT was written in C computer language to maximize the capability of direct electronic circuit operations, including binary formatting of genotype input files and bit operations for matrix manipulations. To further increase computing speed, GEMT was developed with parallel computational capacity, so that computing times decrease linearly with the number of central processing units. Furthermore, the parallel components are dissected small enough so that graphic processing units are also able to perform parallel computations. To solve the memory footprint bottleneck, GEMT allows users to directly control memory usage when big data are analyzed on computers with limited memory. That is, users have the option to trade computing time for less memory usage. Based on these features above, GEMT makes analyses of large and complex datasets feasible without supercomputers.

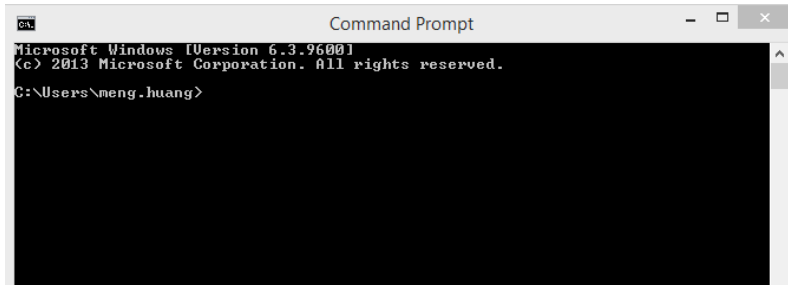
2 Getting started

2.1 Open command line window

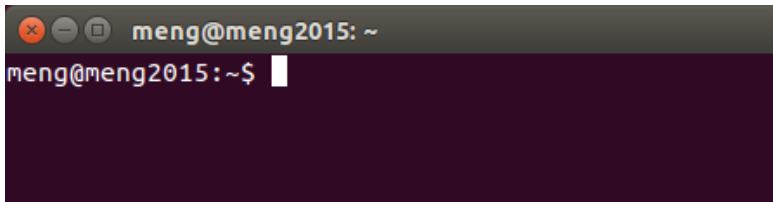
GEMT use Command-Line Interface (CLI). In Mac, the application is called Terminal. From Applications window, click Utilities and then Terminal.



In Windows, the application is called Command Prompt. From search window, input “cmd” and then choose it from results.



In Linux (Ubuntu), the application is also called Terminal. From search window, input “terminal” and then choose it from results.



2.2 Download GEMT

The GEMT executable program (GEMT) can be download at <http://ZZLab.net/GEMT>. Create a folder on your hard disk, for example, myGEMT and save the GEMT executable program in the folder.

2.3 Download input files

Go to <http://ZZLab.net/GEMT> and download the demo data, then copy all the files including data and GEMT executable program to the same folder (e.g. myGEMT).

NOTE: Although most of the file have the same format as GAPIT² and TASSEL³, differences do exist.

2.4 Run GEMT

Users need to specify the pathway and names of GEMT executable file and input files. A convenient way is to change current pathway to the one containing these files. This can be done with this command in the Terminal:

```
cd /users/Zhiwu/myGEMT
```

To perform GWAS between phenotype and one of the genotype formats, for example the compress format, type the following command:

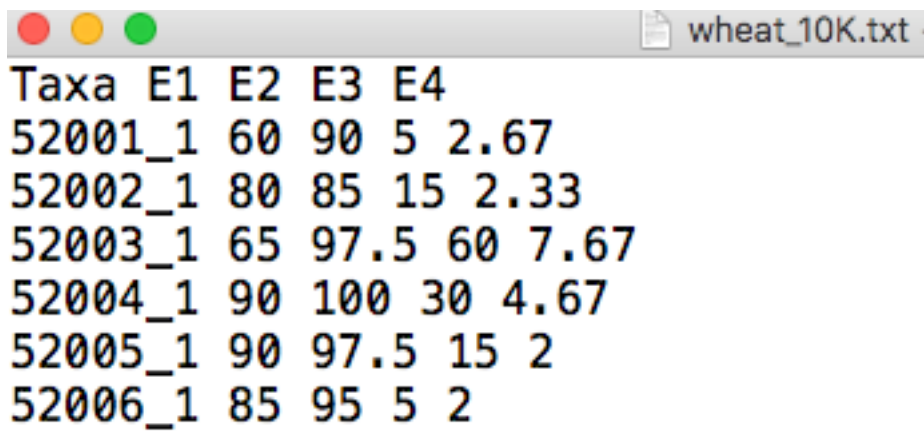
```
GEMT --gwas --file myData --numeric --interaction 3
```

There are five input files involved in this analyses: myData.pre, myData.pos, myData.val, myData.map, and myData.txt. UNIX operating system (e.g. Mac and Ubuntu) may require adding “./” in front of these command lines to specify the current directory.

3 Phenotype file

3.1 Format

Phenotype file is coded as text file with extension of “txt”. The file name must be the same as genotype file(s) so they can be analyzed together. GEMT supports multiple traits. The first column is reserved for individual name. Each trait occupies one column. The first row is reserved as the header of each column. The following figure demonstrates the first eight rows of the phenotype file from the demonstration dataset.



Taxa	E1	E2	E3	E4
52001_1	60	90	5	2.67
52002_1	80	85	15	2.33
52003_1	65	97.5	60	7.67
52004_1	90	100	30	4.67
52005_1	90	97.5	15	2
52006_1	85	95	5	2

The individuals have to be in the same order as the genotype data.

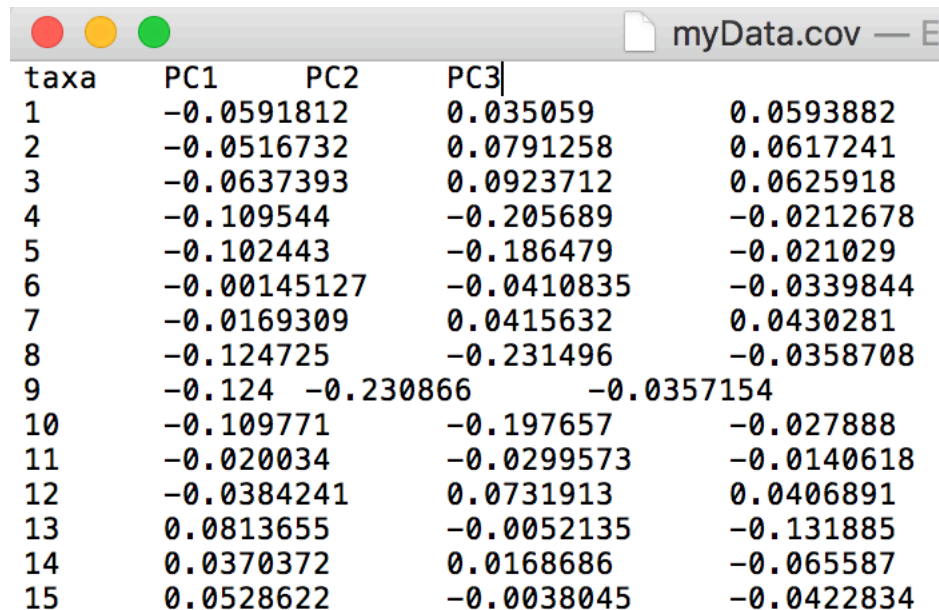
3.2 Missing values

Missing data are allowed in phenotype data. Missing data in genotype can be any character (such as N, NA, or NaN) except numerical number and decimal. But missing data in phenotype should only be “NaN”. Traits are analyzed independently. None missing values of each trait are matched with genotype for each trait.

NOTE: When the trait has missing value and do GWAS

3.3 Covariance format

The covariance will be saved column by column with title and ID into the text file. Different column means different covariance and different row means different individuals. When you want to add covariance into model, just keep its file name same as genotype files and with the extension “.cov” (e.g. myData.cov), then put them into same folder and do GWAS analysis.



taxa	PC1	PC2	PC3
1	-0.0591812		0.035059
2	-0.0516732		0.0791258
3	-0.0637393		0.0923712
4	-0.109544		-0.205689
5	-0.102443		-0.186479
6	-0.00145127		-0.0410835
7	-0.0169309		0.0415632
8	-0.124725		-0.231496
9	-0.124	-0.230866	-0.0357154
10	-0.109771		-0.197657
11	-0.020034		-0.0299573
12	-0.0384241		0.0731913
13	0.0813655		-0.0052135
14	0.0370372		0.0168686
15	0.0528622		-0.0038045

4 GWAS

Both phenotype and genotype files are required to perform GWAS. These files must share a common name with different extensions specified by phenotype and different genotype formats. Analyses of GWAS is specified with “--gwas” option.

4.1 Working with different genotype formats

To perform GWAS with GEMT on one of the genotype formats, type the one of the corresponding file format:

```
GEMT --gwas --file myData --binary
```

```
GEMT --gwas --file myData --numeric
```

```
GEMT --gwas --file myData --hapmap
```

```
GEMT --gwas --file myData --vcf
```

```
GEMT --gwas --file myData --plink
```

The GWAS result contains map information of the marker and corresponding p values. The output file is named by the trait name followed by “_GWAS_result.txt” in format of ‘TraitName_GWAS_result.txt’. The file can be directly used by third-party software (e.g. GAPIT in R) for visualizations, such as Manhattan and QQ plots.

4.2 Changing output file name

Users have the need to change output file name in some cases. GEMT provides an option to fit the need. The default output file name can be changed by using “--out” option as following:

```
GEMT --file myData --out newData
```

5 Advanced operations

GEMT provide more options for analyses with special needs, such as analyses on particular trait, memory saving and customized optimization.

5.1 Memory saving

Define the memory usage by control the number of markers in one cycle (default value is 1000). `--cycle_size 2000`

5.2 Parallel computation

Choose parallel or not. `GEMT --file myData --gwas --parallel 1`

This option will let GEMT switch to parallel computing in CPU device and the number of threads is specified by `--cycle_size`.

5.3 Specifying a trait

GEMT only analyze the first trait by default. A specific trait, for example the third trait, can be analyzed by option “--trait 3” as following:

```
GEMT --file myData --numeric --gwas --trait 3
```

When “--trait 0” is specified, GEMT will automatically analyses on all the available traits.

5.4 Optimization

1. Define the size of bin divided in whole genome, and the unit is 1+e6 bp. The first number is the length of bin_size array, the numbers start from second one are the length of bin size.


```
--bin_size 3 50 5 0.5
```

2. Define the chosen number of top SNPs coming from each bin. The first number is the length of bin_selection array, the numbers start from second one are the value of bin selection. `--bin_selection 3 10 20 30`

3. Define the max number of iteration. `--max_loop 5`

4. Add prior QTN. The first number is the total number of prior QTN, the numbers start from second one are the order of prior QTN in all the SNPs in .map file.

```
--prior 3 12345 54321 43215
```

5.5 Run GEMT from R

As a command, GEMT can be run from R by using system function. The following R code demonstrates the usage of GAPIT demonstration data, simulation of phenotype, analyses with GEMT and visualization.

```
#Import data
setwd("/Users/Zhiwu/myGAPIT")
myGD <- read.table("mdp_numeric.txt", head = TRUE)
myGM <- read.table("mdp_SNP_information.txt", head = TRUE)

#Import library
#source("http://www.bioconductor.org/biocLite.R")
#biocLite("multtest")
#install.packages("gplots")
#install.packages("scatterplot3d")

library('MASS') # required for ginv
library(multtest)
library(gplots)
library(compiler) #required for cmpfun
library("scatterplot3d")
source("http://www.zzlab.net/GAPIT/emma.txt")
source("http://www.zzlab.net/GAPIT/gapit_functions.txt")

#Simulating phenotype
set.seed(99163)
myPheno=GAPIT.Phenotype.Simulation(GD=myGD,h2=.75,NQTN=20,QTNDist="geometry",
  effectunit=.92)
myY=myPheno$Y
QTN.position=myPheno$QTN.position
```

```
#Create GEMT input data
setwd("/Users/Zhiwu/myGEMT/Simulation")
GD=t(myGD[,-1])
write.table(GD,file="myData.dat",quote=F,sep="\t",col.name=F,row.name=F)
write.table(myGM,file="myData.map",quote=F,sep="\t",col.name=T,row.name=F)
write.table(myY,file="myData.txt",quote=F,sep="\t",col.name=T,row.name=F)

#Run GEMT
system("/Users/Zhiwu/myGEMT/GEMT --file /Users/Zhiwu/myGEMT/Simulation/myData
--out /Users/Zhiwu/temp/myData -interaction 3 ")

#Manhattan and QQ plots
GMP <- read.delim("myData_GWAS_result.txt", head = T)
GMP=GMP[,c(2,3,5)]
GAPIT.Manhattan(GI.MP = GMP, name.of.trait = "Trait",plot.type =
"Genomewise",DPP=50000,cutOff=0.01,band=2,seqQTN=QTN.position)
GAPIT.QQ(P.values=GMP[,3], plot.type = "log_P_values", name.of.trait =
"Trait",DPP=50000)
```

6 Q&A

1. Why do I use GEMT?

A: GEMT is designed to make you more successful for finding genes of your interest, such as the ones lead to cure of cancers, or reduction of using pesticides. It also aims to reduce computing time and memory usage so that big can be analyzed.

2. How do I cite GEMT?

A: We are in the process for your convenience of citation. Please cite: “Huang, M and Zhiwu Zhang, GEMT, <http://zzlab.net>, access data”.

7 References

1. Zhang, Z., Buckler, E. S., Casstevens, T. M. & Bradbury, P. J. Software engineering the mixed model for genome-wide association studies on large samples. *Br. Bioinform* **10**, 664–675 (2009).

2. Lipka, A. E. *et al.* GAPIT: genome association and prediction integrated tool. *Bioinformatics* **28**, 2397–2399 (2012).
3. Bradbury, P. J. *et al.* TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).