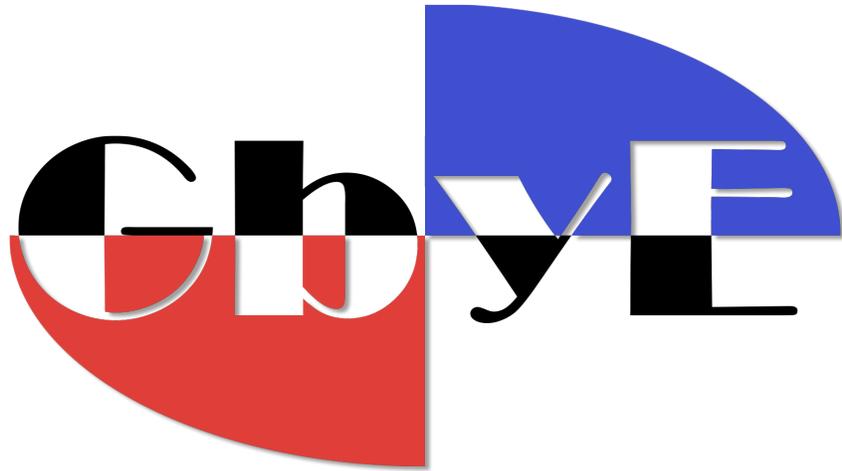


User Manual of



Tool for Genetic and Environmental interaction analysis

(Version 1.0)

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 GbyE results with other software packages, such as GAPIT, and FarmCPU, or BLINK.

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

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

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



**WASHINGTON
GRAIN
COMMISSION**

Contents

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

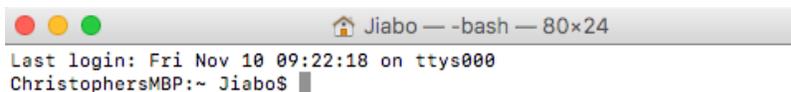
1 Introduction

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. GbyE (Genomic by Environments association analysis tool) was used to detect intereaction SNPs between G&E or Multiple Traits. GbyE 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, GbyE 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, GbyE 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, GbyE makes analyses of large and complex datasets feasible without supercomputers.

2 Getting started

2.1 Open command line window

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



```
Jiabo - -bash - 80x24
Last login: Fri Nov 10 09:22:18 on ttys000
ChristophersMBP:~ Jiabo$
```

2.2 Download GbyE

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

2.3 Download input files

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

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

2.4 Run GbyE

Users need to specify the pathway and names of GbyE 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/Username/myGbyE
```

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

```
GbyE --gwas --file myData --numeric --interaction 4
```

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 commmand 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. GbyE 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	X13_FC_Irr	X13_FC_Dry	X12_Gr_Irr	X12_Gr_Dry
2174-05	117.76	78.75	180.12	85.75
2180	131.75	69.87	187.31	94.15
ABOVE	146.25	72.04	181.04	77.58
AGATE	124.6	75.7	118.3	77
ALICE	122.52	80.94	171.83	89.01
ALLIANCE	125.06	72.65	185.69	80.08
ANTELOPE	117.54	72	160.52	90.07
ANTON	126.34	71.27	172.27	95.55
ARAPAHOE	147.52	72.95	145.66	76.42
ARLIN	144.48	71.47	211.14	79.14
AVALANCHE	111.33	80.46	175.02	93.83
BENNETT	115.1	78.93	165.52	93.28
BIG_SKY	127.91	73.49	75.25	102.14
BILL_BROWN	133.57	78.94	191.88	85.11
BOND_CL	137.08	74.6	170.21	83.96
BRONZE	111.76	64.41	86.83	84.57

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.

```
myData.cov
taxa_myY   PCA_1   PCA_2   PCA_3
2174-05 7.04916623224607 -19.8945255188523 13.435309665714
2180 3.7953640451235 3.52302293624172 -17.9210176633054
ABOVE 14.4197055770282 -38.1749461088588 46.8914765326278
AGATE 22.5675014823421 8.83787997427119 -1.94363849353979
ALICE 6.55238852857415 4.10791336842176 -12.4390433743214
ALLIANCE 20.3734846021749 10.391907016351 3.7175910932417
ANTELOPE 11.3712322299268 4.24713173211229 -10.6821288473065
ANTON 16.0029192308066 -0.495956350783792 -10.1256580494246
ARAPAHOE 17.4061145562358 8.44394033216448 -8.31048937869399
ARLIN 7.53361290656697 -2.50029744993498 -18.7071516701156
AVALANCHE 12.5582866176084 -8.84090229123803 10.7160725829823
BENNETT 23.2970133202436 12.7284893976113 -0.349025352307705
BIG_SKY 20.7124397262195 17.9888184672319 -2.45964291776306
BILL_BROWN 14.9335511312985 11.8126675657642 -8.62839609241744
BOND_CL 15.9734944391446 -4.18825231173657 22.0235664841814
BRONZE 13.3558687327898 5.26336506612441 -9.63969294989979
BUCKSKIN 23.5773820024464 15.0284938323578 4.67763834307303
C003W043 11.2571229981201 -12.6565124081549 7.13779608511083
THUNDER CL 8.36527346691113 -22.8260303196807 11.8607198010806
```

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 Changing output file name

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

```
GbyE --file myData --out newData
```

5 Advanced operations

GbyE 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. `GbyE --file myData --gwas --parallel 1`

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

5.3 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`

5.4 Run GbyE from R

As a command, GbyE 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 GbyE and visualization.

```
rm(list=ls())
library('MASS') # required for ginv
library(multttest)
library(gplots)
library(compiler) #required for cmpfun
library("scatterplot3d")
library("EMMREML")
library(ape)
source("http://www.zzlab.net/GAPIT/emma.txt")
```

```

source("http://www.zzlab.net/GAPIT/gapit_functions.txt")
  ha2=c(0.8,0.8)
  rg=diag(1,2)
  rg[1,2]=rg[2,1]=0.9
  re=diag(1,2)
  re[1,2]=re[2,1]=.1
  NQTN=20
  NE=2
rep=50
n_e=NE
set.seed(99163)

system(paste("./GbyE --gwas --file ","myData --numeric ","--interaction
",n_e,sep=""))
result=read.table(paste("GbyE","_GWAS_result.txt",sep=""),head=T)
result=result[,c(1,2,3,5,4)]
chrom=result[1:(nrow(result)/n_e),2]
for(i in 1:n_e)
{
  result[((1+(i-1)*(nrow(result)/n_e)):((nrow(result)/n_e+(i-1)*(nrow(resu
lt)/n_e))),2]=chrom
}

source("http://www.zzlab.net/GbyE/Manhattan_g&e.R")
#result=cbind(newmap,result[,4])
GAPIT.Manhattan(GI.MP=result[,-1],name.of.trait="GW",Name_environ=c("13_FC_Irr",
"13_FC_Dry","12_Gr_Irr","12_Gr_Dry"),Nenviron=n_e)

```

6 Q&A

1. Why do I use GbyE?

A: GbyE is designed to make you more successful for finding interaction genes between genome and multiple environments, 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 GbyE?

A: We are in the process for your convenience of citation. Please cite: “Jiabo Wang and Zhiwu Zhang, GbyE, <http://zzlab.net/GbyE>, 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).