User Manual for

Genomic Association and Prediction Integrated Tool

(Version 3)

Last updated on May 4, 2022

Zhiwu Zhang Laboratory

WASHINGTON STATE UNIVERSITY

Disclaimer: While extensive testing has been performed by the Zhiwu Zhang Lab at (2014 to present) at Washington State University and Edward Buckler Lab (2012-2014) at Cornell University, respectively. Results are, in general, reliable, correct, and appropriate. However, results are not guaranteed for any specific set of data. We strongly recommend that users validate GAPIT results with other software packages, such as SAS and TASSEL.

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

Questions and comments: To benefit GAPIT community, questions and comments should be addressed to GAPIT forum: https://groups.google.com/forum/#!forum/gapit-forum. The GAPIT team members will periodically go through these questions and comments and address them accordingly. For countries with restriction on Google, questions and comments are welcome to Jiabo Wang by email: wangjiaboyifeng@163.com.

Citation: Multiple statistical methods are implemented in GAPIT version 1, 2 and 3. Citations of GAPIT vary depending on methods and versions used in the analysis:

Method	Method paper	Version 1[1]	Version 2[2]	Version 3[3]
General Linear Model (GLM)	Price et al, 2006, *Nature Genetics*[4]	✓	✓	✓
Mixed Linear Model (MLM)	Yu et al, 2005, *Nature Genetics*[5]	✓	✓	✓
Compression MLM (CMLM)	Zhang et al, 2010, *Nature Genetics*[6]	✓	✓	✓
gBLUP	Zhang et al, 2007, *J. Anim. Science*[7]	✓	✓	✓
Enriched CMLM	Li et al, 2014, *BMC Biology*[8]		✓	✓
SUPER	Wang et al, 2014, *PLoS One*[9]		✓	✓
MLMM	Segura et al, 2012, *Nature Genetics*[10]			✓
FarmCPU	Liu et al, 2016, *PloS Genetics*[11]			✓
cBLUP and sBLUP	Wang et al, 2019, *Heredity*[12]			✓
BLINK	Huang et al, 2019, *GigaScience*[13]			✓

Note: These references are listed in section of Reference.

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

Table of Contents

1 INTRODUCTION

1.1 Why GAPIT?

GAPIT implemented a series of methods for Genome Wide Association (GWAS) and Genomic Selection (GS). The GWAS models include General Linear Model (GLM), Mixed Linear Model (MLM or Q+K), Compressed MLM (CMLM), Enriched CMLM, SUPPER, Multiple Loci Mixed Model (MLMM), FarmCPU and BLINK. The GS models include gBLUP, Compressed BLUP, and SUPER BLUP.

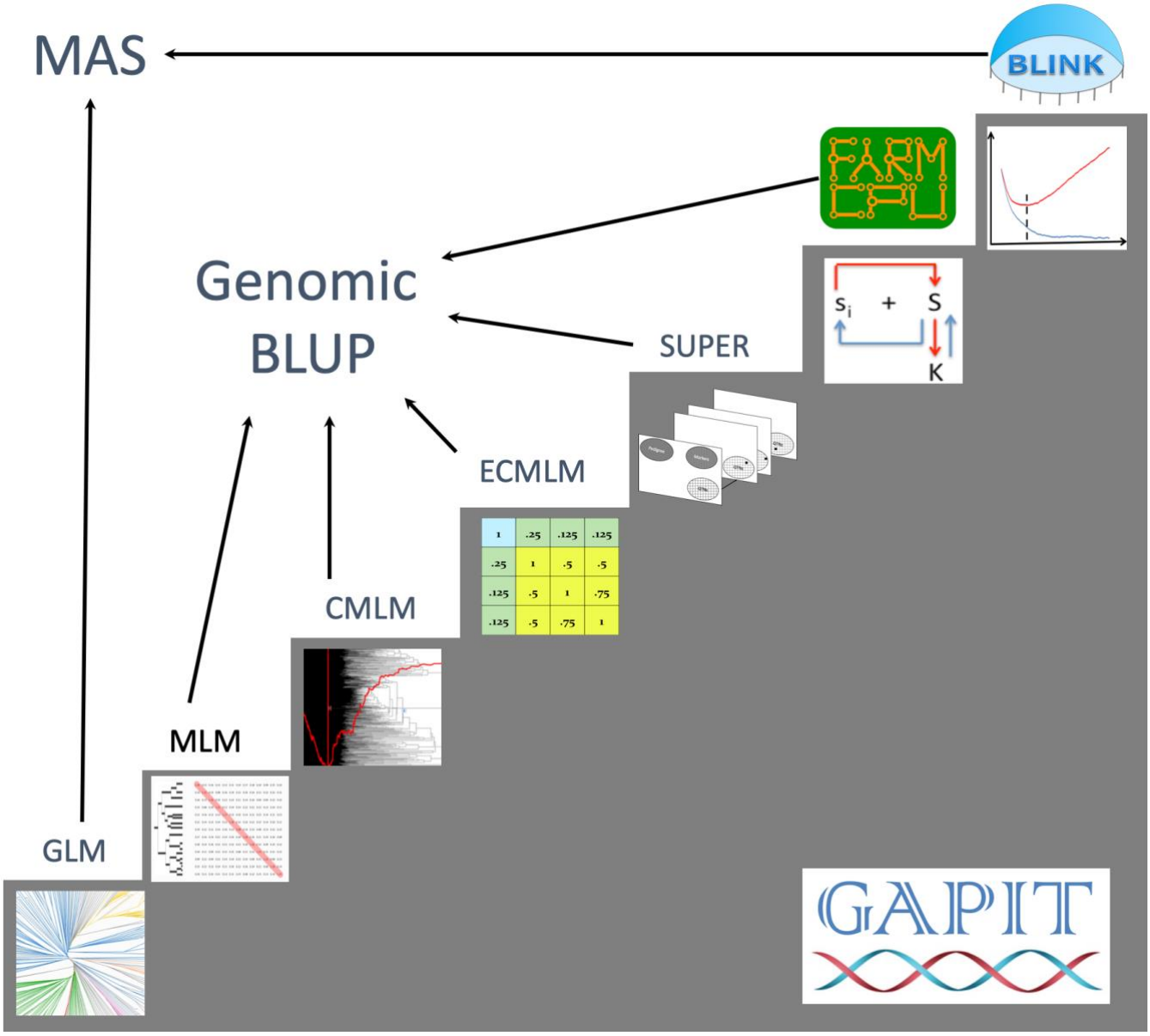


Figure 1.1. Methods implemented in GAPIT for GWAS and genomic selection. All the methods support GWAS, including General Linear Model (**GLM**), Mixed Linear Model (**MLM**), Compressed MLM (**CMLM**), Enriched CMLM (**ECMLM**), Settlement of MLM Under Progressively Exclusive Relationship (**SUPER**), Fixed and random model Circulating Probability Unification (**FarmCPU**), and Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (**BLINK**). Some of these methods support genomic selection, including MLM, CMLM, ECMLM, SUPER, and FarmCPU. The remaining (GLM and BLINK) can be used for breeding through marker assisted selection (MAS).

1.2 Getting Started

GAPIT is a package that is run in the R software environment, which can be freely downloaded from http://www.r-project.org or http://www.rstudio.com. There are two sources to install GAPIT package.

Zhiwu Zhang Lab website:

```
source("http://zzlab.net/GAPIT/GAPIT.library.R")
source("http://zzlab.net/GAPIT/gapit_functions.txt")
```

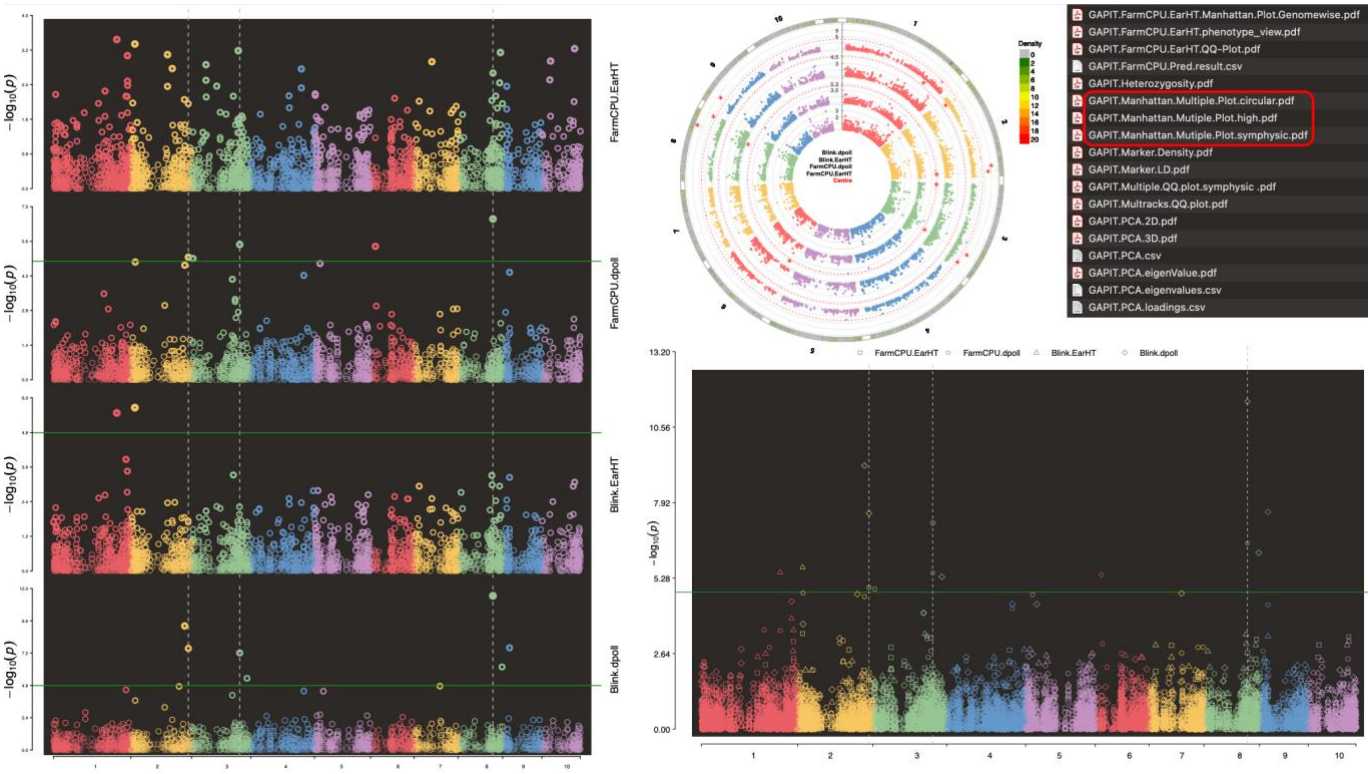
Or GitHub:

```
install.packages("devtools")
devtools::install_github("jiabowang/GAPIT3",force=TRUE)
library(GAPIT3)
```

The easiest way is to COPY/PASTE GAPIT tutorial script. Here are example code and outputs:

```
#Import data from Zhiwu Zhang Lab
myY  <- read.table("http://zzlab.net/GAPIT/data/mdp_traits.txt", head = TRUE)
myGD=read.table(file="http://zzlab.net/GAPIT/data/mdp_numeric.txt",head=T)
myGM=read.table(file="http://zzlab.net/GAPIT/data/mdp_SNP_information.txt",head=T)

#GWAS
myGAPIT=GAPIT(
Y=myY[,c(1,2,3)], #fist column is ID
GD=myGD,
GM=myGM,
PCA.total=3,
model=c("FarmCPU", "Blink"),
Multiple_analysis=TRUE)
```



As demonstrated above, users can specify any one or multiple models. GAPIT accepts multiple input data formats, including both numeric, hapmap, and PLINK genotype formats. GAPIT produces comprehensive reports to interpret data and results in publication ready formats. For examples, the distribution of marker density and decay of linkage equilibrium inform user if the markers are dense enough. When GWAS were conducted with multiple traits, environments, or multiple models, GAPIT produces the integrated Manhattan plots with overlapped associated markers highlighted. The above analysis should be completed within couple minutes. In your current R working directory, you should find multiples files with three types of extensions: pdf, csv, and txt. The three types of the Manhattan plots are displayed above.

1.3 How to use the GAPIT user manual?

The next three chapters (2-5) describe details on the input data, GWAS, GS, and output results. Chapter 6 presents scenarios to demonstrate the applications. Chapter 7 is for users to use GAPIT for prototyping. The last chapter (8) lists frequently questions and answers. Before reading the next three chapters, we recommend you go directly to the tutorial chapter and run other tutorials.

1.4 How to cite GAPIT?

Although historical version of GAPIT (1 and 2) are available, the newest version (3) is recommended for full support from GAPIT team. Citations should specify the version and models used. For an example, a GWAS run by GAPIT version 3 using BLINK can cite as "The GWAS was conducted by GAPIT (version 3)[3] using BLINK model[13]" . A GS with run by GAPIT version 3 using gBLUP/cBLUP can cite as "GS was conducted by GAPIT (version 3)[3] using gBLUP model[7] and cBLUP model[12]".

2 Input Data

There are six types of input data: phenotype (Y), genotype in hapmap format (G), genotype data in numerical format (GD), genotype map (GM), kinship (K), and covariate variables (CV), see **Table 2.1**. Phenotypic data must be provided, and the rest are optional, including genotype data, map, kinship, and covariate. Kinship can be provided by users or be generated from genotype data, or even omitted by suing BLINK method. Genotypic data may not be needed for genomic prediction if the kinship matrix is provided. Covariate variables (fixed effects), such as population structure represented by the Q matrix (subpopulation proportion) or principal components (PCs), are optional. GAPIT provides the option to calculate PCs from the genotypic data. All input files should be saved as a "Tab" delimited text file.

Notice: *It is important that each taxa name is spelled, punctuated, and capitalized (R is case sensitive) the same way in each of the input data sets. If this is not done, they will be excluded from the analysis. Additionally, the taxa names must not be numeric.*

Table 2.1 Gallery of GAPIT input data

Parameter	Default	Options	Description	Tutorial files*
Y	NULL	User	Phenotype	mdp_traits.txt
KI	NULL	User	Kinship Matrix	KSN.txt
CV	NULL	User	Covariate Variables	mdp_PC.txt
G	NULL	User	Genotype Data in Hapmap Format	mdp_genotype_test.hmp.txt
GD	NULL	User	Genotype Data in Numeric Format	mdp_numeric.txt
GM	NULL	User	Genotype Map for Numeric Format	mdp_SNP_information.txt

The tutorial file can be downloaded at: http://zzlab.net/GAPIT/GAPIT_Tutorial_Data.zip. These files can read into R with following commands:

```
#Phenotypic Data
#myY  <- read.table("mdp_traits.txt", head = TRUE)

#HapMap genotype format
myG <- read.delim("mdp_genotype_test.hmp.txt", head = FALSE)

#Numerical genotype format
#-------------------A pair of Genotypic Data and map files-----------------------------
myGD <- read.table("mdp_numeric.txt", head = TRUE)
myGM <- read.table("mdp_SNP_information.txt", head = TRUE)

#Kinship matrix
myKI <- read.table("KSN.txt", head = FALSE)

#covaraite variables (such as population structure represented by Q matrix or PC)
myCV <- read.table("mdp_PC", head = TRUE)
```

2.1 Phenotypic Data

The user has the option of performing GWAS on multiple phenotypes in GAPIT. This is achieved by including multiple phenotype columns in phenotypic file. Taxa names should be in the first column of the phenotypic data file and the remaining columns should contain the observed phenotype from each individual. Missing data should be indicated by either "NaN" or "NA". The first ten observations in the tutorial data (mdp_traits.txt) are displayed as follows:

Taxa	EarHT	dpoll	EarDia
811	59.5	NaN	NaN
4226	65.5	59.5	32.21933
4722	81.13	71.5	32.421
33-16	64.75	64.5	NaN
38-11	92.25	68.5	37.897
A188	27.5	62	31.419
A214N	65	69	32.006
A239	47.88	61	36.064
A272	35.63	70	NaN
A441-5	53.5	67.5	35.008

The file is "Tab" delimited. The first row consists of column labels (i.e., headers). The column labels indicate the phenotype name, which is used for the remainder of the analysis.

The phenotype file can be input to R by typing command line:

```
myY <- read.table("mdp_traits.txt", head = TRUE)
```

2.2 Genotypic Data

Genotypic data are required for GWAS, but are optional for GS. In the later case, genomic prediction is performed using a kinship matrix provided by the user. GAPIT accepts genotypic data in either HapMap format or in numeric format.

2.2.1 Hapmap Format

Hapmap is a commonly used format for storing sequence data where SNP information is stored in the rows and taxa information is stored in the columns. This format allows the SNP information (chromosome and position) and genotypes of each taxa to be stored in a single file.

The first 11 columns display attributes of the SNPs and the remaining columns show the nucleotides observed at each SNP for each taxa. The first row contains the header labels and each remaining row contains all the information for a single SNP. The first five individuals on the first seven SNPs from the tutorial data (mdp_genotype.hmp.txt) are presented below.

rs	alleles	chrom	pos	strand	assembly	center	protLSID	assayLSID	panel	QCcode	33-16	38-11	4226	4722	A188
PZB00859.1	A/C	1	157104	+	AGPv1	Panzea	NA	NA	maize282	NA	CC	CC	CC	CC	AA
PZA01271.1	C/G	1	1947984	+	AGPv1	Panzea	NA	NA	maize282	NA	CC	GG	CC	GG	CC
PZA03613.2	G/T	1	2914066	+	AGPv1	Panzea	NA	NA	maize282	NA	GG	GG	GG	GG	GG
PZA03613.1	A/T	1	2914171	+	AGPv1	Panzea	NA	NA	maize282	NA	TT	TT	TT	TT	TT
PZA03614.2	A/G	1	2915078	+	AGPv1	Panzea	NA	NA	maize282	NA	GG	GG	GG	GG	GG
PZA03614.1	A/T	1	2915242	+	AGPv1	Panzea	NA	NA	maize282	NA	TT	TT	TT	TT	TT
PZA00258.3	C/G	1	2973508	+	AGPv1	Panzea	NA	NA	maize282	NA	GG	CC	CC	CG	CC
PZA02962.13	A/T	1	3205252	+	AGPv1	Panzea	NA	NA	maize282	NA	TT	TT	TT	TT	TT
PZA02962.14	C/G	1	3205262	+	AGPv1	Panzea	NA	NA	maize282	NA	CC	CC	CC	CC	CC
PZA00599.25	C/T	1	3206090	+	AGPv1	Panzea	NA	NA	maize282	NA	CC	TT	CC	TT	TT

This file can be read into R by typing the following command line:

```
myG <- read.table("mdp_genotype_test.hmp.txt", head = FALSE)
```

Although all of the first 11 columns are required, GAPIT uses only 3 of these: the "rs" column, which is the SNP name (e.g. "PZB00859.1"); the "chrom" column, which is the SNP's chromosome; and the "pos", which is the SNP's base pair (bp) position. It is sufficient to fill in the requested information in the remaining eight columns with "NA"s. To be consistent with HapMap naming conventions, missing genotypic data are indicated by either "NN" (double bit) or "N" (single bit).

For genotypic data in HapMap format, GAPIT accepts genotypes in either double bit or in the standard IUPAC code (single bit) as following:

Genotype	AA	CC	GG	TT	AG	CT	CG	AT	GT	AC
Code	A	C	G	T	R	Y	S	W	K	M

By default, the HapMap numericalization is performed so that the sign of the allelic effect estimate (in the GAPIT output) is with respect to the nucleotide that is second in alphabetical order. For example, if the nucleotides at a SNP are "A" and "T", then a positive allelic effect indicates that "T" is favorable. Selecting "Major.allele.zero = TRUE" in the GAPIT() function will result in the sign of the allelic effect estimate being with respect to the minor allele. In this scenario, a positive allelic effect estimate will indicate that the minor allele is favorable.

2.2.2 Numeric format

GAPIT also accepts the numeric format. The order of taxa and SNPs is reversed from the HapMap format. Columns are used for SNPs and rows are used for taxa in the numeric format. This format is problematic in Excel because the number of SNPs used in a typical analysis exceeds the Excel column limit. Additionally, this format does not contain the chromosome and position of the SNPs. Therefore, two separate files must be provided to GAPIT. One file contains the numeric genotypic data (called the "GD" file), and the other contains the position of each SNP along the genome (called the "GM" file).
Note: The SNPs in the "GD" and "GM" files NEED to be in the same order.

Homozygotes are denoted by "0" and "2" and heterozygotes are denoted by "1" in the "GD" file. Any numeric value between "0" and "2" can represent imputed SNP genotypes. The first row is a header file with SNP names, and the first column is the taxa name. The example file (mdp_numeric.txt from tutorial data set) is as following:

taxa	PZB00859.1	PZA01271.1	PZA03613.2	PZA03613.1
33-16	2	0	0	2
38-11	2	2	0	2
4226	2	0	0	2
4722	2	2	0	2
A188	0	0	0	2
…				

This file can be read into R by typing the following command line:

```
myGD <- read.table("mdp_numeric.txt", head = TRUE)
```

The genetic map ("GM") file contains the name and location of each SNP. The first column is the SNP id, the second column is the chromosome, and the third column is the base pair position. As seen in the example, the first row is a header file. The example file (mdp_SNP_information.txt from tutorial data set) is as following:

Name	Chromosome	Position
PZB00859.1	1	157104
PZA01271.1	1	1947984
PZA03613.2	1	2914066
PZA03613.1	1	2914171
PZA03614.2	1	2915078
…		

This file is read into R by typing the following command line:

```
myGM <- read.table("mdp_SNP_information.txt", head = TRUE)
```

2.3 Kinship

The kinship matrix file (called "KI" in GAPIT) is formatted as an n by n+1 matrix where the first column is the taxa name, and the rest is a square symmetric matrix. Unlike the other input data files, the first row of the kinship matrix file does not consist of headers. The example (KSN.txt from tutorial data set) is as following:

33-16	2	0.228837	0.229322	0.268842	0.237145	0.0781	0.347107
38-11	0.228837	2	0.244965	0.293708	0.175211	0.079276	0.295606
4226	0.229322	0.244965	2	0.214859	0.236153	0.082693	0.283713
4722	0.268842	0.293708	0.214859	2	0.25935	0.061573	0.160104
A188	0.237145	0.175211	0.236153	0.25935	2	0.061469	0.232799
A214N	0.0781	0.079276	0.082693	0.061573	0.061469	2	0.110364
A239	0.347107	0.295606	0.283713	0.160104	0.232799	0.110364	2

This file is read into R by typing the following command line:

```
myKI <- read.table("KSN.txt", head = FALSE)
```

2.4 Covariate variables

A file containing covariates (called "CV" in GAPIT) can include information such as population structure (commonly called the "Q matrix"), which are fitted into the GWAS and GS models as fixed effects. These files are formatted similarly to the phenotypic files. Specifically, the first column consists of taxa names, and the remaining columns contain covariate values. The first row consists of column labels. The first column can be labeled "Taxa", and the remaining columns should be covariate names. The example file (mdp_population_structure.txt from tutorial data set) is as following

Taxa	Q1	Q2	Q3
33-16	0.014	0.972	0.014
38-11	0.003	0.993	0.004
4226	0.071	0.917	0.012
4722	0.035	0.854	0.111
A188	0.013	0.982	0.005
A214N	0.762	0.017	0.221
A239	0.035	0.963	0.002
A272	0.019	0.122	0.859
A441-5	0.005	0.531	0.464

This file is read into R by typing the following command line:

```
myCV <- read.table("mdp_population_structure.txt", head = TRUE)
```

3 GWAS

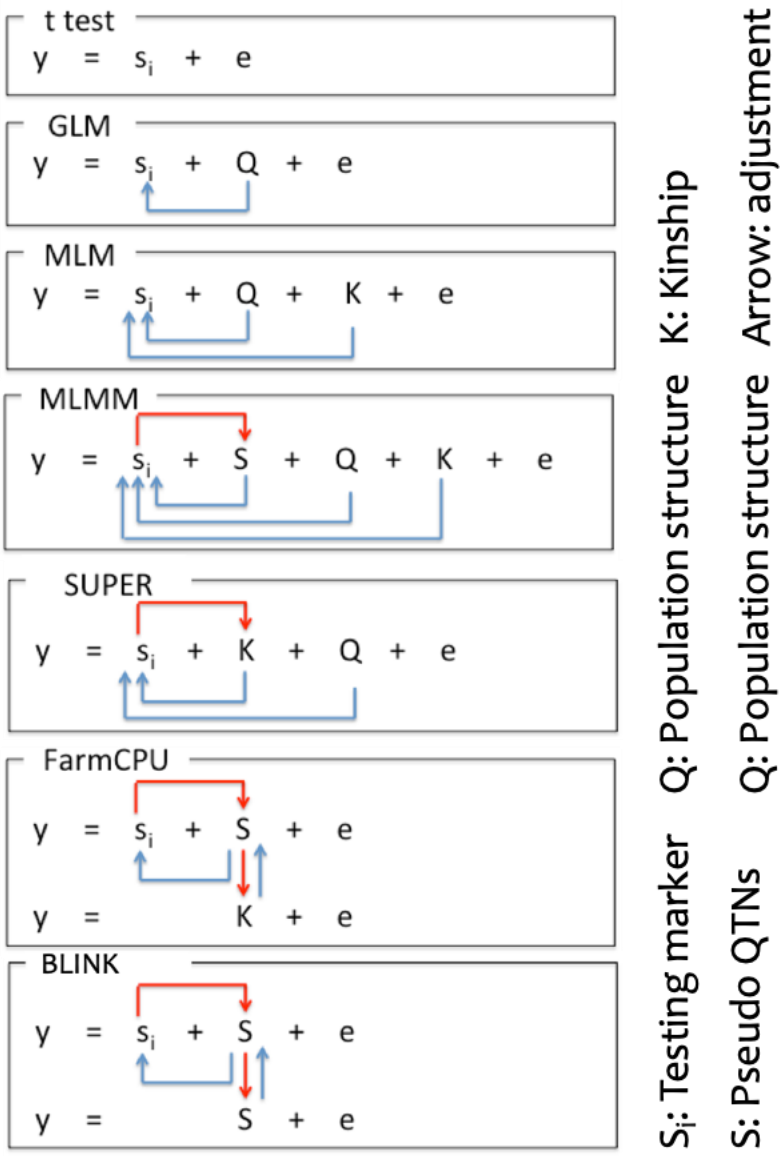
3.1 GWAS model overview

Currently, GAPIT has implemented more than ten models. The similarity and difference among seven milestone models are summarized in the figure below. The simplest model (t test) is to directly detect the association between a phenotype (y) and markers (Si) one at a time, where i=1 to m, and m is number of markers. When a cofactor, such as population structure (Q) is introduced through a general linear model (GLM), the cofactor may not only account residuals (e) partially, but also adjust some effect that does not belong to the testing markers and consequently reduce false positives. The mixed linear model (MLM) applies the same principle by adding individuals' genetic effects as random cofactor effects with variance structure defined by the kinship (K) among individuals. In both Q or Q+K models, Q and K stay the same. There are no cofactors that are adjusted by the marker tests.

Inclusion of cofactors benefits the reduction of false positives for testing markers in GLM and MLM. The disadvantage is these cofactors are also confounded with the testing markers. In MLM particularly, the kinship defines the genetic effect of individuals which equal the sum of causal genes. Many know genes identified by GLM had signals below threshold using MLM[14].

The compressed MLM (CMLM) was proposed to reduce the confounding problem of MLM[6]. Individuals are compressed into groups. The individual genetic effects are replaced with the group genetic effects. Correspondingly, kinship among individuals is replaced with kinship among groups with grouping maximized using maximum likelihood method. The optimization of kinship among groups further improve statistical power[8] in the enriched CMLM (ECMLM).

GLM and MLM are the special cases of CMLM which is a general format. When number of groups is forced to be one in CMLM, CMLM becomes GLM. Similarly, when number of groups is forced to be the number of individuals in CMLM, CMLM becomes MLM. The optimization of groping improves statistical power[15].



The optimization of groping in CMLM and optimization of kinship among groups in ECMLM are thoroughly based on phenotypes. There is no impact from marker tests. The situation was changed by the multiple loci mixed model (MLMM). Through marker association tests, the associated markers are fitted as the cofactors for marker test. The cofactors are adjusted through forward and backward stepwise regression of mixed model. However, both Q and K remain unchanged.

In the SUPER method, K is derived from the associated markers and is adjected accordingly by the marker tests. As the K is derived from a smaller number of markers than the K in MLM and MLMM that are derived from all the markers, the confounding between K and some of the markers becomes more severe. SUPER eliminate the confounding by using the complimentary kinship derived from associated markers except for the ones that are in strong linkage disequilibrium (LD) with the testing markers under a user-defined threshold.

To eliminate the ambiguity of determining associated markers are in LD with a testing marker, FarmCPU completely removes the confounding from kinship by using a fixed-effect model without a kinship derived either from all markers, or associated markers. Instead, the kinship derived from the associated markers is used to select the associated markers using the maximum likelihood method. This process overcomes the model overfitting problems of stepwise regression. FarmCPU uses both the fixed effect model and the random effect model iteratively.

In both SUPER and FarmCPU models, the bin approach is used to avoid selecting markers from the same locations with bin size and the number of bins optimized using the maximum likelihood method. The underlying assumption is that causal genes are distributed equally across the genome. BLINK eliminates the assumption to improve statistical power by using the linkage disequilibrium (LD) method. Markers are sorted with the most significantly associated maker on the top as reference. The remaining markers are removed if they are in LD with the most associated marker. Among the remaining makers, the most significantly associated maker is selected as the reference. The process is repeated until no markers can be removed. The random effect model in FarmCPU to select associated markers using the maximum likelihood method remains a high computing cost for a large number of individuals. BLINK approximate the maximum likelihood using Bayesian Information Content (BIC) in a fixed-effect model to eliminate the computational burden.

3.2 Model selection

With the multiple models implemented in GAPIT, a common question is which to choose. Many people make the selection based on their trust gained over experience. For example, some researchers must choose GLM implemented in PLINK[16] because it is the only software accepted by the reviewers and editors in their fields. In general, computing efficiency and statistical power should be the criteria for the selection.

Two models use the fixed-effect model only which is the most computing efficient, including GLM and BLINK. FarmCPU is a hybrid that uses both the fixed-effect model and the random effect model. The rest use a fixed and random effects mixed model which is computationally expensive, including MLM, CMLM, ECMLM, SUPER, and MLMM. CMLM uses groups and is cubic time faster than MLM. Due to additional optimizations, ECMLM and SUPER are slower than CMLM. For a trial analysis, GLM and BLINK are good to start with.

Regarding statistical power, multiple loci models (e.g. MLMM, FarmCPU, and BLINK) are superior to the rest. Within multiple loci model category, FarmCPU is superior to MLMM[11] and BLINK is superior to FarmCPU[13]. Within the single locus model category, MLM is superior to GLM[5], CMLM is superior to MLM[6], ECMLM is superior to ECMLM[8], SUPER and MLM are superior to MLM[9,10]. These relationships

are summarized by the model stairs in the first chapter. The method on a higher stair has higher statistical power than the one on a lower stair. The magnitude of the differences among models may change from case to case, however, their order stays the same. The inversion of the order has not been found. Therefore, BLINK is selected as the default GAPIT model because of its high computing efficiency and statistical power. Users are welcome to use the following statement to justify the usage of BLINK.

"In addition to the capability to incorporate principal components as covariates to reduce false positives due to population stratification, BLINK iteratively incorporates associated markers as covariates for testing markers to eliminate their connection to the cryptic relationship among individuals. The associated markers are selected according to linkage disequilibrium, optimized for Bayesian information content, and reexamined across multiple tests to reduce false negatives".

3.3 Model description

The detailed model description is critical for readers to understand exactly how the analyses were performed or to replicate the analyses. As all the implemented models are well described elsewhere, the model description should focus on the covariates that are specific to the analyses. All covariates should be described in detail, including the levels for the category covariates. Here is an example:

"GAWAS was conducted by GAPIT (version 3)[3] using BLINK model[17]. The covariate variables include the first three principal components derived from all the markers and the origin-group. The origin group was coded as indicators (0/1) for each of the origin groups except the last one to avoid linear dependency".

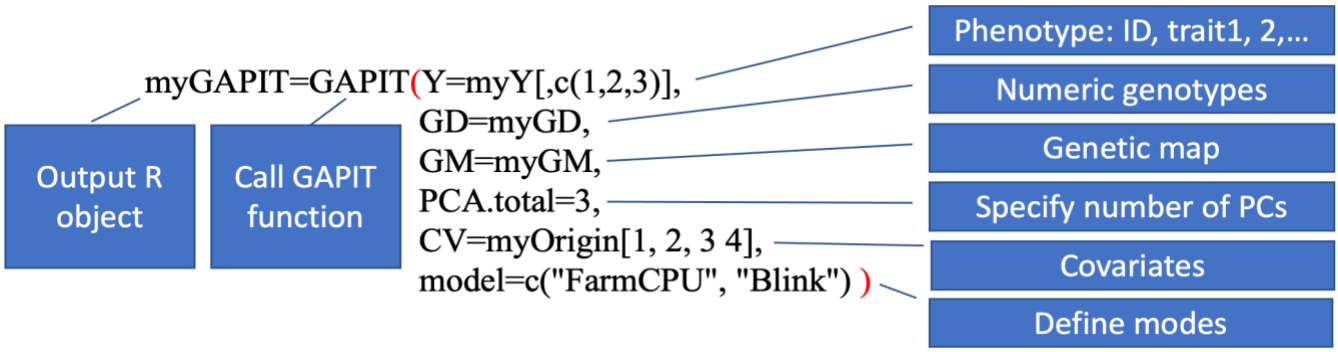
3.4 Model justification

It was found during the development of FarmCPU that causal genes can be detected even when they are confounded with population structure and population structure such as the first three principal components were fitted as covariates for testing markers. As an anonymous FarmCPU reviewer suggested, fitting several PCs does not hurt the degree of freedom very much, however, it helps in situations there are non-genetic effects associated with population structure during phenotyping. Otherwise, a false positive marker would appear to capture the non-genetic effect. Therefore, fitting several PCs is recommended for all analyses. The related justification is as follows.

"Principal component analysis was performed with GAPIT (version 3)[3] using all available SNPs. GAWAS was conducted by GAPIT (version 3)[3] using BLINK model[17]. The first principal components were fitted as covariate variables to reduce the false positives due to population stratification".

3.5 GAPIT Syntax

GAPIT can be executed by calling "GAPIT()" with inputs and parameters included in "()". The inputs include phenotypes, genotype data, genetic map, covariate variables. The general parameters include number of PCs as covariates and models. More general parameters can be found in Table 3.5.1.



There are also parameters specific to models. For example, the CMLM model involves number of groups. These model specific parameters will be described within the sections of specific models.

Table 3.5.1. GAPIT input parameters.

Parameter	Default	Options	Description
model	Blink	GLM, MLM, CMLM, SUPER, MLMM, FarmCPU, and Blink	Choose one or multiple models to conduct GWAS
kinship.algorithm	VanRaden	Zhang, Loiselle and EMMA	Algorithm to Derive Kinship from Genotype
kinship.cluster	average	complete, ward, single, mcquitty, median, and centroid	Clustering algorithm to group individuals based on their kinship
kinship.group	Mean	Max, Min, and Median	Method to derive kinship among groups
LD.chromosome	NULL	User	Chromosome for LD analysis
LD.location	NULL	User	Location (center) of SNPs for LD analysis
LD.range	NULL	User	Range around the Central Location of SNPs for LD Analysis
PCA.total	0	>0	Total Number of PCs as Covariates
PCA.scaling	None	Scaled, Centered.and.scaled	Scale And/Or Center And Scale The SNPs Before Conducting PCA
SNP.FDR	1	>0 and <1	Threshold to Filter SNP on FDR
SNP.MAF	0	>0 and <1	Minor Allele Frequency to Filter SNPs in GWAS Reports
SNP.effect	Add	Dom	Genetic Model
SNP.P3D	TRUE	FALSE	Logic Variable to Use P3D or Not for Testing SNPs
SNP.fraction	1	>0 and <1	Fraction of SNPs Sampled to Estimate Kinship and PCs
SNP.test	TRUE	FALSE	Logic Variable to Test SNPs or Not

3.6 Mixed Linear Model (MLM)

MLM includes both fixed and random effects. Including individuals as random effects gives an MLM the ability to incorporate information about relationships among individuals. This information about relationships is conveyed through the kinship (K) matrix, which is used in an MLM as the variance-covariance matrix between the individuals. When a genetic marker-based kinship matrix (K) is used jointly with population structure (commonly called the "Q" matrix, and can be obtained through STRUCTURE[18] or conducting a principal component analysis[19]), the "Q+K" approach improves statistical power compared to "Q" only[20]. An MLM can be described using Henderson's matrix notation as follows:

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e}, \quad (1)$$

where **Y** is the vector of observed phenotypes; **β** is an unknown vector containing fixed effects, including the genetic marker, population structure (Q), and the intercept; **u** is an unknown vector of random additive genetic effects from multiple background QTL for individuals/lines; **X** and **Z** are the known design matrices; and **e** is the unobserved vector of residuals. The **u** and **e** vectors are assumed to be normally distributed with a null mean and a variance of:

$$\mathrm{Var}\begin{pmatrix}\mathbf{u}\\ \mathbf{e}\end{pmatrix} = \begin{pmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{pmatrix} \quad (2)$$

where $\mathbf{G} = \sigma^2_a\mathbf{K}$ with σ^2_a as the additive genetic variance and **K** as the kinship matrix. Homogeneous variance is assumed for the residual effect; i.e., $\mathbf{R} = \sigma^2_e\mathbf{I}$, where σ^2_e is the residual variance. The proportion of the total variance explained by the genetic variance is defined as heritability (h^2).

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}, \quad (3)$$

3.7 Compressed MLM (CMLM)

As kinship is derived from all the markers, incorporating with the kinship for testing markers in a MLM causes the confounding between the testing markers and the individuals' genetic effects with variance structure defined by the kinship. To reduce the confounding, individuals are replaced by their corresponding groups in the compressed MLM developed by Zhang et al in 2010[21]. Cluster analysis is used to assign similar individuals into groups. The elements of the kinship matrix are used as similarity measures in the clustering analysis. Various linkage criteria (e.g., unweighted pair group method with arithmetic mean, UPGMA) can be used to group the lines together. The number of groups is specified by the user. Once the lines are assigned into groups, summary statistics of the kinship between and within groups are used as the elements of a reduced kinship matrix. This procedure is used to create a reduced kinship matrix for each compression level.

A series of mixed models are fitted to determine the optimal compression level. The value of the log likelihood function is obtained for each model, and the optimal compression level is defined as the one whose fitted mixed model yields the largest log likelihood function value. There are three parameters to determine the range and interval of groups for examination: group.from, group.to and group.by. Their defaults are 0, *n* and 10, where *n* is the total number of individuals.

3.8 General Linear Model (GLM)

Regular MLM[22] is an extreme case of CMLM where each individual is considered as a group. It can be simply performed by setting the number of groups equal to the total number of individuals, e.g. group.from = *n* and group.to = *n*, where *n* is total number of individuals shared in both the genotype and phenotype files. Similarly, general linear model (GLM) is another extreme case of CMLM where all individuals are considered as one group. It can be simply performed by setting the number of groups equal to one, i.e. group.from = 1 and group.to = 1. GLM is the working model in PLINK[23], a primary software for studies in human genetics.

3.9 P3D/EMMAx

In addition to implementing compression, GAPIT uses EMMAx/P3D[6,24] to reduce computing time for MLM, CMLM, ECMLM, and SUPER. If specified, the additive genetic (σ^2_a) and residual (σ^2_e) variance components will be estimated prior to conducting GWAS. These estimates are then used for each SNP where a mixed model is fitted.

3.10 SUPER

SUPER is an advanced version of FaST-Select, developed Wang et al. in 2016. The major difference between SUPER and FaST-Select is that SUPER uses bin approach to select associated markers. The entire genome is divided into equal sized bins and each bin is represented by the most significant marker on the bin. The bin size and number of bins selected are optimized using maximum likelihood method in a random model with the kinship derived from the selected bins. Consequently, the confounding between the kinship and some of markers become more severe than the kinship derived from all markers. SUPER eliminate the confounding by using the complementary kinship derived from associated markers except the ones that are in strong linkage disequilibrium (LD) with the testing markers under a user defined threshold. Both simulation and real data demonstrated that SUPER had higher statistical power than regular MLM.

To run SUPER in GAPIT, simply specify model= "SUPER".

3.11 Multiple Locus Mixed Linear Model (MLMM)

GAPIT implemented a Multiple Loci Mixed Linear Model (MLMM) which use forward-backward stepwise linear mixed-model regression to include associated markers as covariates.

To run MLMM in GAPIT, simply specify model= "MLMM".

3.12 FarmCPU

To solve the problem of false positive control and confounding between testing markers and cofactors simultaneously, an iterative method, named Fixed and random model Circulating Probability Unification (**FarmCPU)**, was developed in 2016[11]. The associated markers detected from the iterations are fitted as the cofactors to control false positives for testing the rest markers in a fixed effect model. To avoid the over model fitting problem in stepwise regression, a random effect model is used to select the associated markers using maximum likelihood method[11].

In the cycle of fixed effect model of iterations, markers are tested against the associated markers, not the confounded kinship used by MLM, CMLM, ECMLM, SUPER, and MLMM. In the cycle of random effect model of iterations, markers are selected among a small number of associated markers using maximum likelihood method to avoid the over model fitting problem in stepwise regression used by MLMM, which select marker among all available markers. Consequently, FarmCPU exhibits higher statistical power than MLMM[11]. As FarmCPU tests markers in a fixed effect model, it is computational efficient than the methods that test markers in random effect model, such as MLM, CMLM, ECMLM, SUPER, and MLMM[11].

To run FarmCPU in GAPIT, simply specify model= "FarmCPU".

3.13 BLINK

BLINK method was designed to have both high statistical power and computational efficiency[13]. It was inspired by FarmCPU method with two major changes to achieve the objectives. One is to eliminate the assumption that causal genes are evenly distributed across genome that required by FarmCPU. As the assumption cause either inclusion of non causal genes, or missing the causal genes that are in the same bin with another causal genes with stronger signal. BLINK works directly on markers instead of bins. Markers that are in linkage disequilibrium (LD) with the most significant marker are excluded. For the second remaining marker, the exclusion is conducted in the same way as the most significant marker, so on and so forth until no marker can be excluded.

The other change is to use Bayesian Information Content (BIC) of a fixed effect model to approximate the maximum likelihood of a random effect model to select the associated markers among the markers remained the exclusion based on LD. As both the models of testing markers and selecting associated markers as cofactors are fixed effect model, the computation complexity reach the maximum. A dataset with one million individuals and one million markers can be solved in hours by using BLINK C version. The BLINK R version can be run as standard alone, or through GAPIT. To run BLINK in GAPIT, simply specify model= "Blink". The performances of the two versions were documented by the BLINK article on GigaScience[13].

4 Genomic Selection

Genomic selection, or genomic prediction termed in human genetics, is to use genetic markers across the whole genome to predict individual performances if formats of phenotypes or predicted genetic merit. In contrast to GWAS, there is a strong interaction between prediction methods and the traits measured in a particular condition. The reversion of method superiorities has been found in many cases. The genomic selection based on SUPER, named SUPER BLUP, has higher prediction accuracy than the genomic selection based on MLM known as genomic BLUP (gBLUP) for traits controlled with a smaller number of genes. The prediction accuracies are reversed for traits controlled by a large number of genes. The genomic selection based on CMLM, named Compressed BLUP, has higher accuracy for traits with low heritability than gBLUP. GAPIT implemented a series of methods for GWAS and genomic selection toward high statistical power.

4.1 Genomic BLUP

Genomic prediction is performed with the method based on genomic best linear unbiased prediction, (gBLUP)[7]. The method was extended to compressed best linear unbiased prediction (cBLUP) by using the CMLM approach that was proposed for GWAS[6]. The genetic potential for a group, which is derived from the BLUPs of group effects in the compressed mixed model, is used as a prediction for all individuals in the group.

The groups created from compression belong to either a reference (R) or an inference (I) panel. All groups in the reference panel have at least one individual with phenotypic data, and all groups in the inference panel have no individuals with phenotypic data. Genomic prediction for groups in the inference panel is based on phenotypic ties with corresponding groups in the reference panel.

The group kinship matrix is then partitioned into to R and I groups as follows:

$$k=\begin{bmatrix} k_{RR} & k_{RI} \\ k_{IR} & k_{II} \end{bmatrix} , \quad (4)$$

where $\mathbf{k_{RR}}$ is the variance-covariance matrix for all groups in the reference panel, $\mathbf{k_{RI}}$ is the covariance matrix between the groups in the reference and inference panels, $\mathbf{k_{IR}} = (\mathbf{k_{RI}})'$ is the covariance matrix between the groups inference and reference panels, and $\mathbf{k_{II}}$ is the variance-covariance matrix between the groups in the inference panels.

Solving of mixed linear model is performed on the reference individuals.

$$y_R = X_R\beta + Z_R u_R + e_R \quad , \quad (5)$$

where all terms are as defined in Equation (1), and the "R" subscript denotes that only individuals in the reference panel are considered.

The genomic prediction of the inference groups is derived Henderson's formula (1984) as follows:

$$u_I = K_{IR} K_{RR}^{-1} u_R \quad , \quad (6)$$

where $\mathbf{k_{IR}}$, $\mathbf{k_{RR}}$, and $u_\mathbf{R}$ are as previously defined, and $u_\mathbf{I}$ is the predicted genomic values of the individuals in the inference group.

The reliability of genomic prediction is calculated as follows:

$$\text{Reliability} = 1 - \frac{PEV}{\sigma_a^2}$$

(7)

where PEV is the prediction error variance which is the diagonal element in the inverse left-hand side of the mixed model equation, and σ_a^2 is the genetic variance.

4.2 Compressed gBLUP

The compressed MLM substitute individuals with their corresponding groups that were clustered based on the kinship among individuals. Research demonstrates that the compressed MLM had higher statistical power for GWAS. Research also demonstrated that compressed MLM also had higher prediction accuracy than the regular MLM, especially for traits with low heritability. As the regular MLM is an extreme case of compressed MLM, the compressed MLM has higher, or at least equal prediction accuracy as the regular MLM. When a compressed MLM is specified in GAPIT, individuals' breeding values are predicted by the breeding values of their corresponding groups.

4.3 SUPER gBLUP

The regular MLM uses the kinship derived from all the markers while SUPER uses the kinship derived from the associated markers. As the associated markers are selected from all the markers using the maximum likelihood method, the kinship used by SUPER has a better likelihood than the kinship used by the regular MLM. Research demonstrated that the estimated breeding values from SUPER had higher prediction accuracy than the estimated breeding values from the regular MLM.

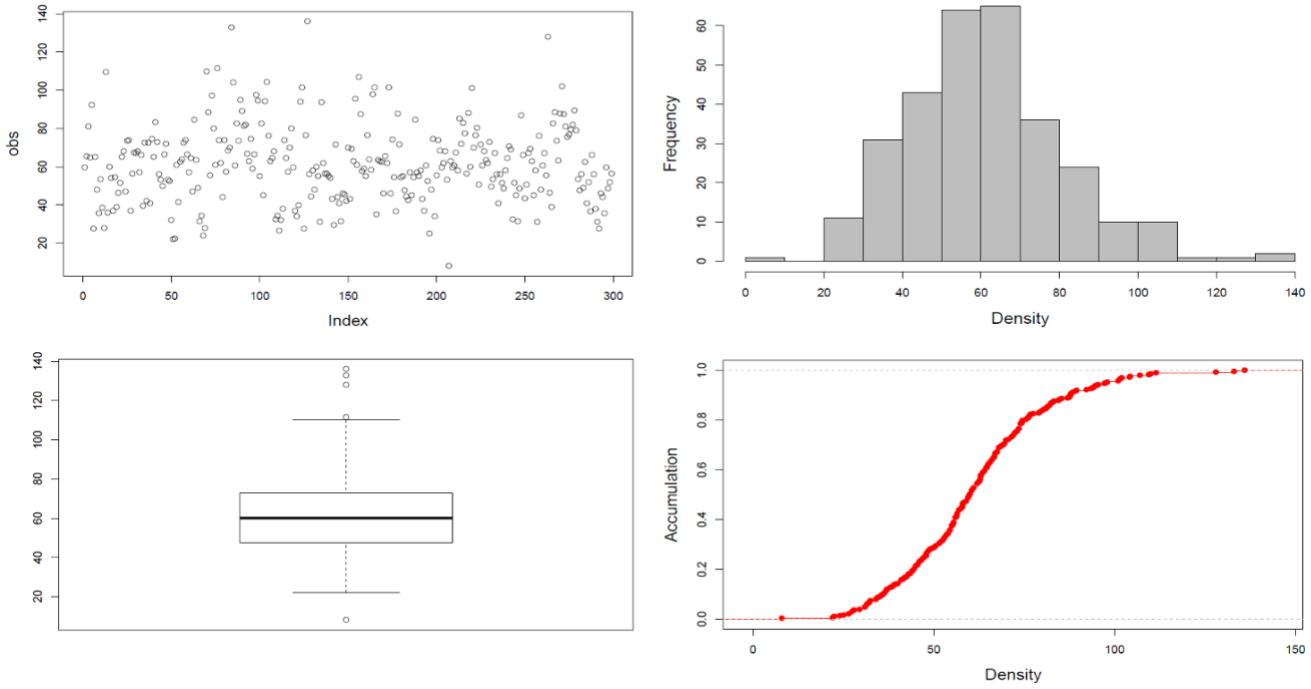
5 Output Results

GAPIT produces a series of output files that are saved in two formats. All tabular results are saved as comma separated value (.csv) files, and all graphs are stored as printable document format (.pdf) files. This section provides descriptions of these output files.

File name	Description	Type
Allelic_Effect_Estimates	Estimate allelic effect with method	CSV
Df.tValue.StdErr	Estimate allelic t-value	CSV
GWAS.Results	SNP information and P-Value	CSV
Log	Log of whole model	CSV
PRED	Genomic Prediction	CSV
ROC	Table for power and FDR	CSV
Kin.VanRaden	kinship with VanRanden method	CSV
PCA	Principle components analysis	CSV
PCA.eigenvalues	Eigenvalues of PCA	CSV
PCA.loadings	Rotation of PCA	CSV
Compression.multiple.group	Compress likelihood, heritability and variance.	PDF
MAF	Minimum Allelic Frequency	PDF
Manhattan.Plot.Chromosomewise	Chromosome Manhattan	PDF
Manhattan.Plot.Genomewise	Genome Manhattan	PDF
Optimum	Heritability and Variance components	PDF
phenotype_view	Phenotype analysis	PDF
QQ-Plot	QQ plot	PDF
ROC	Power and FDR in ROC	PDF
Heterozygosity	Heterozygosity of genotype	PDF
Kin.VanRaden	Heat map of kinship	PDF
Marker.Density	Marker Density	PDF
Marker.LD	LD of first 1000 markers	PDF
PCA.2D	2D PCA plot	PDF
PCA.3D	3D PCA plot	PDF
PCA.eigenValue	Eigenvalue and variance of PCA	PDF
NJtree.fan	Fan type NJ tree	PDF
NJtree.unrooted	Unrooted NJ tree	PDF
Manhattan.Mutiple.Plot	Manhattan plot for multiple traits or methods	PDF
Circular.Manhattan.Plot.	Circular Manhattan plot	PDF
Multraits.QQplot	QQ plot for multiple trait or method	PDF
Interactive.PCA	Interactive PCA plot	HTML
Interactive.Manhattan	Interactive Manhattan plot	HTML
Interactive.QQ	Interactive QQ plot	HTML

5.1 Phenotype diagnosis

GAPIT diagnosis phenotype in several ways, including scatter plot, histogram, box plot and accumulative distribution.



5.2 Marker density

Marker density is critical to establish Linkage Disequilibrium (LD) between markers and causal mutations. Comparison between the marker density and the LD decade over distance provides the indication if markers are dense enough to have good coverage of LD.

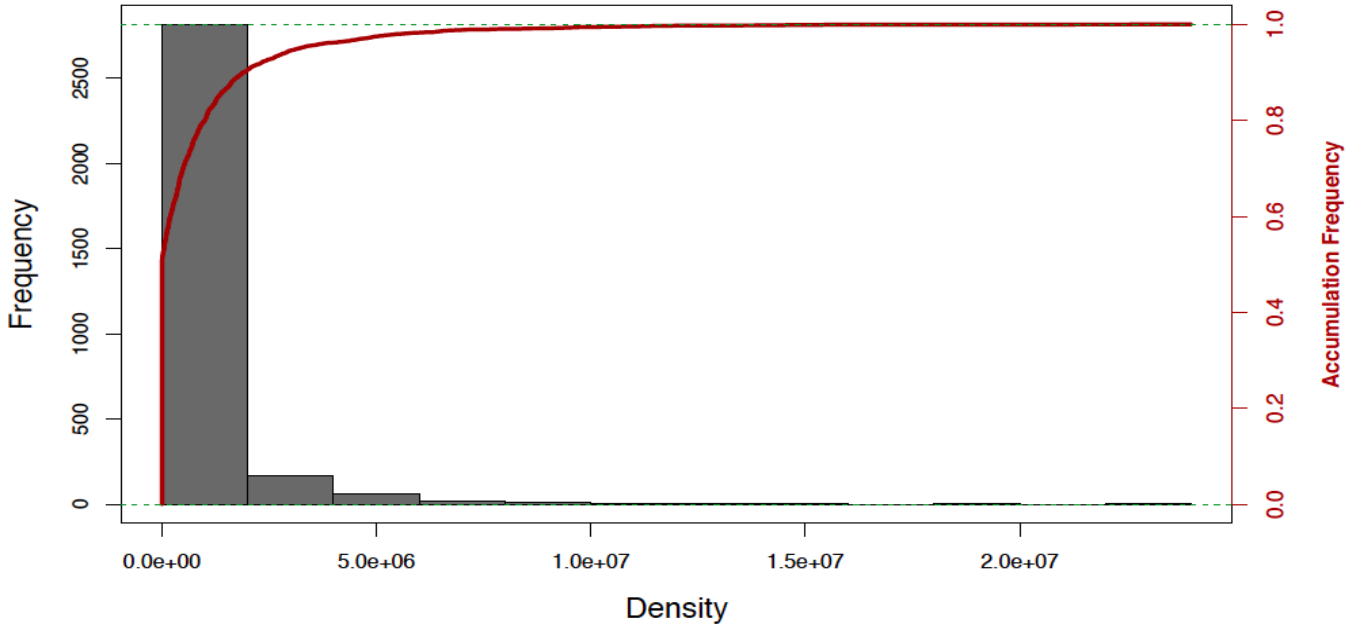


Figure 5.2 Frequency and accumulative frequency of marker density. Distribution of marker density is displayed as a histogram and an accumulative distribution.

5.3 Linkage Disequilibrium Decay

Linkage disequilibrium are measured as R square for pair wise markers and plotted against their distance. The moving average of adjacent markers were calculated by using a sliding windows with ten markers.

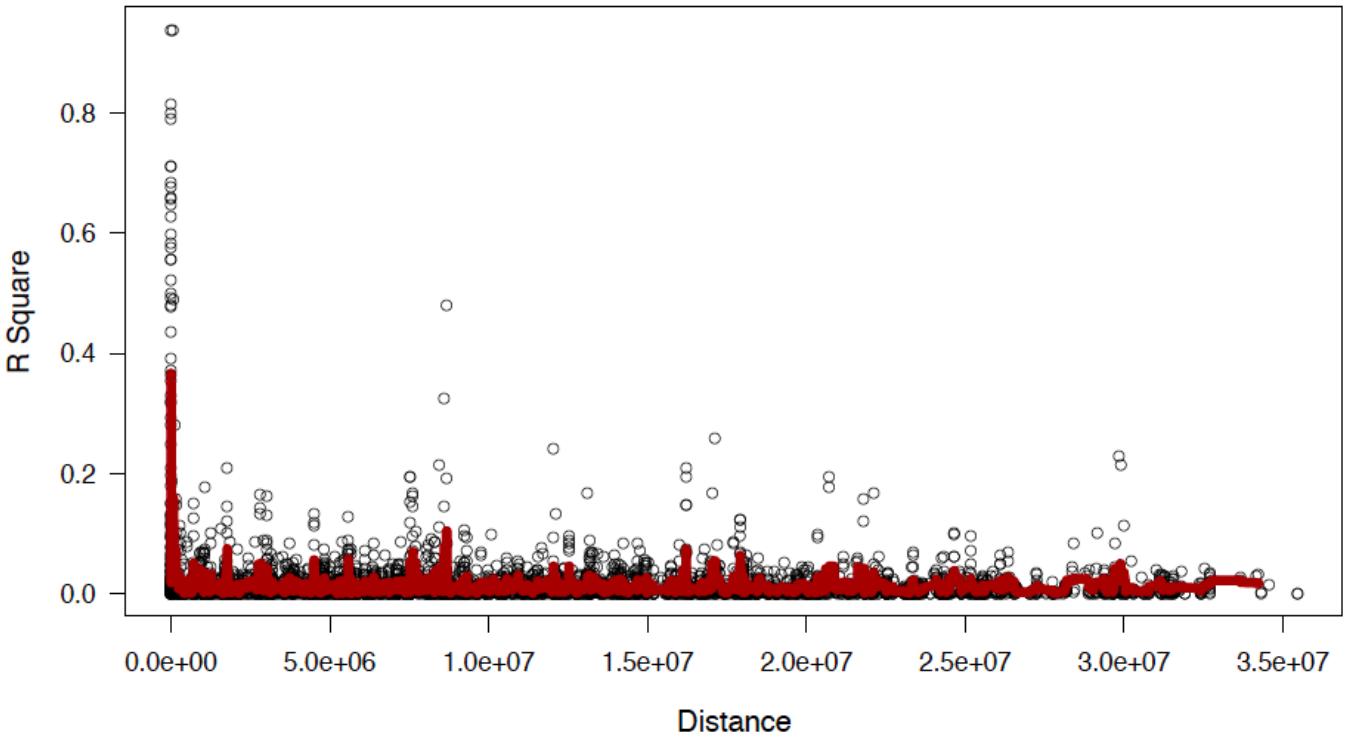
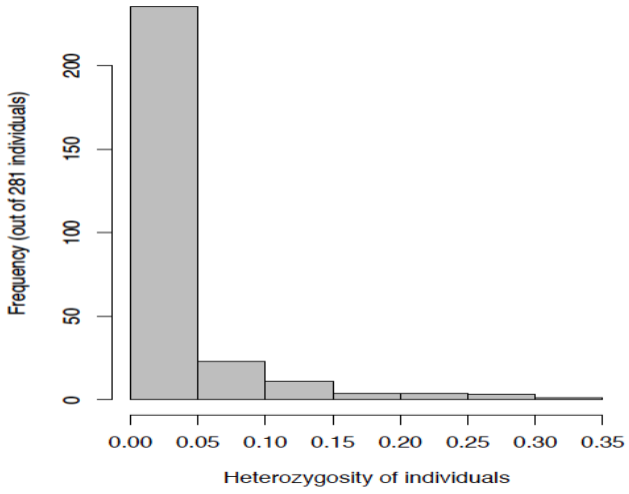
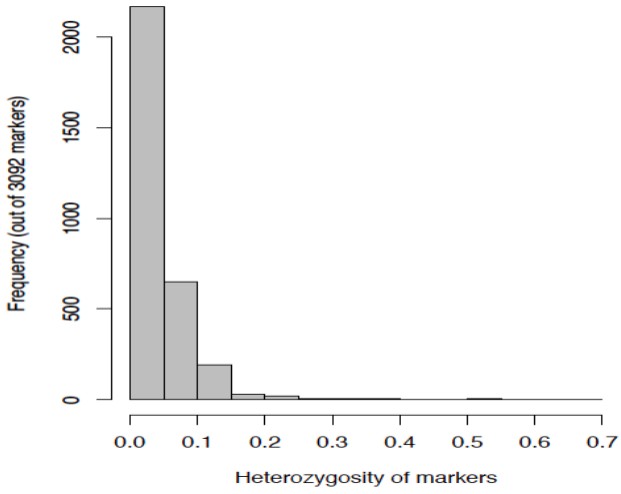


Figure 5.3 Linkage disequilibrium (LD) decay over distance. LDs were calculated on sliding windows with 100 adjacent genetic markers. Each dot represents a pair of distances between two markers on the window and their squared correlation coefficient. The red line is the moving average of the 10 adjacent markers.

5.4 Heterozygosis

The frequency of heterozygous were calculated for both individuals and markers. High level of heterozygosis indicated low quality. For example, over 50% of heterozygosis on inbred lines for some of markers suggested they problematic (see bottom right).





5.5 Principal Component (PC) plot

For each PC included in the GWAS and GPS models, the observed PC values are plotted.

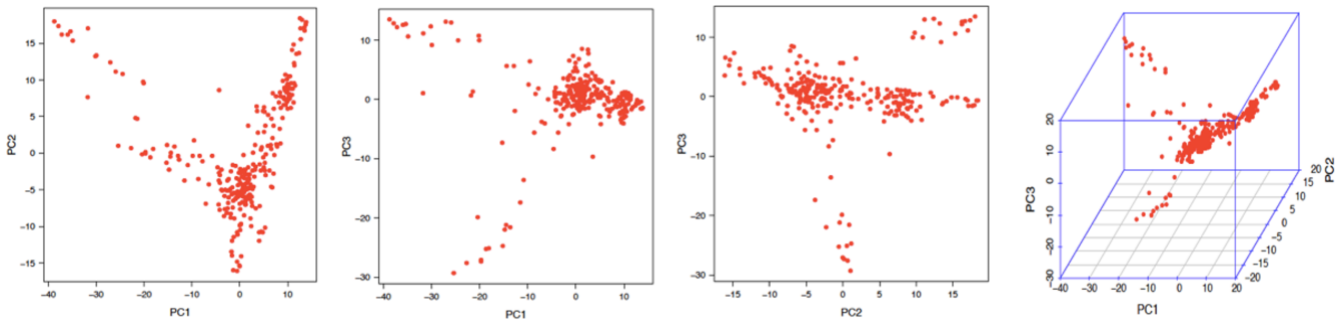


Figure 5.5 Pair-wise plots and 3D plots of principal component (PC).

5.6 Kinship plot

The kinship matrix used in GWAS and GPS is visualized through a heat map. To reduce computational burden, this graph is not made when the sample size exceeds 1,000.

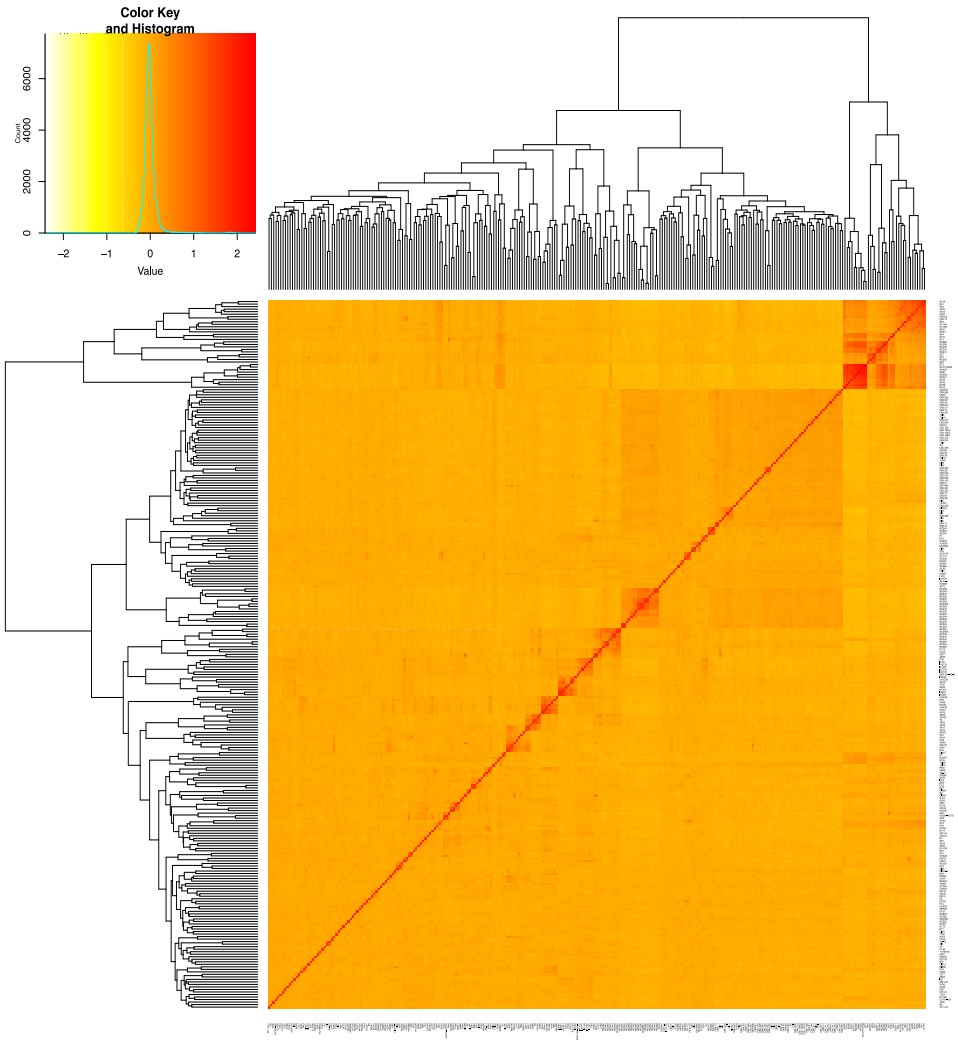


Figure 5.6 Kinship plot. A heat map of the kinship matrix is created to indicate the relationship between individuals.

5.7 Neighbor-Joining (NJ)-tree

Following the compress group, we classify individuals to explain population structure. We also can plot group PCA plot with previous group.

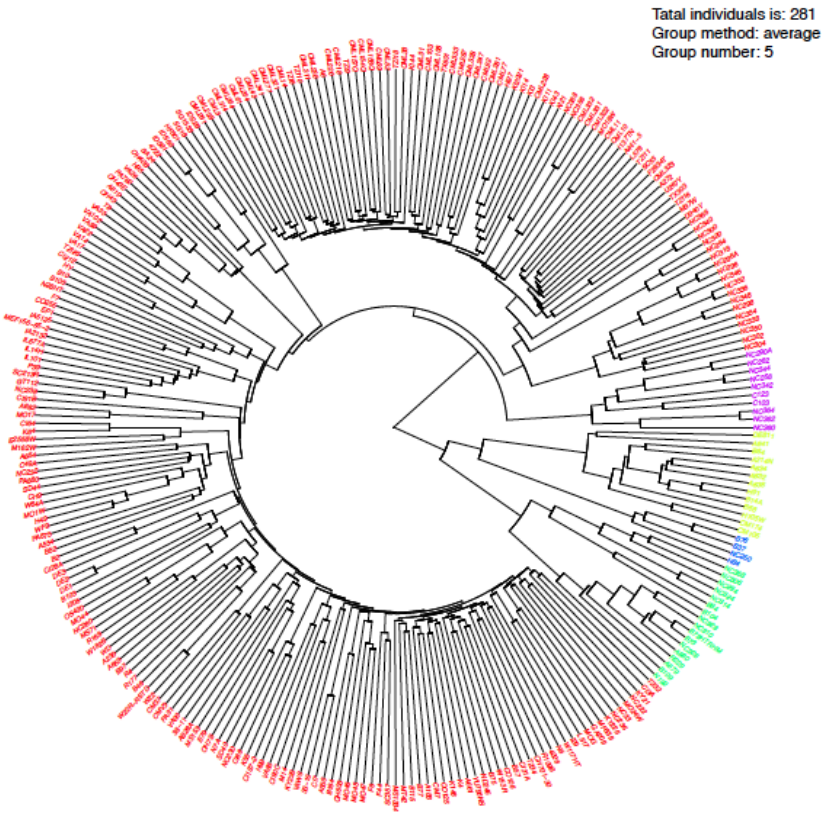


Figure 5.7 Neighbor-Joining (NJ)-tree The whole population was divided into 5 clusters with each colors.

5.8 QQ-plot

The quantile-quantile (QQ) –plot is a useful tool for assessing how well the model used in GWAS accounts for population structure and familial relatedness. In this plot, the negative logarithms of the *P*-values from the models fitted in GWAS are plotted against their expected value under the null hypothesis of no association with the trait. Because most of the SNPs tested are probably not associated with the trait, the majority of the points in the QQ-plot should lie on the diagonal line. Deviations from this line suggest the presence of spurious associations due to population structure and familial relatedness, and that the GWAS model does not sufficiently account for these spurious associations. It is expected that the SNPs on the upper right section of the graph deviate from the diagonal. These SNPs are most likely associated with the trait under study. By default, the QQ-plots in GAPIT show only a subset of the larger *P*-values (i.e., less significant *P*-values) to reduce the file size of the graph.

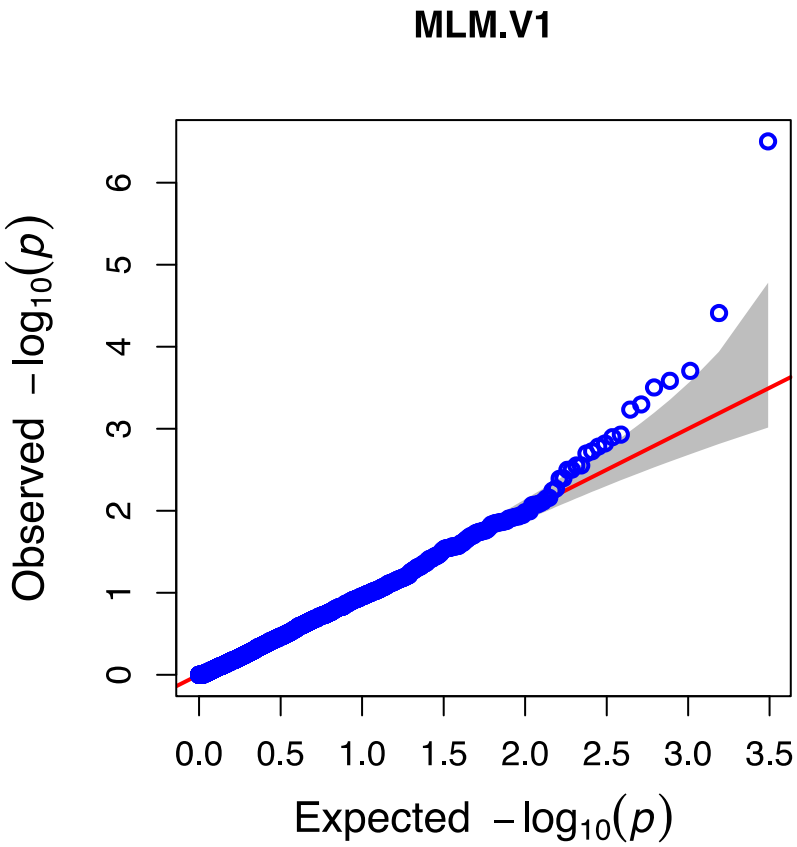


Figure 5.8 Quantile-quantile (QQ) –plot of *P*-values. The Y-axis is the observed negative base 10 logarithm of the *P*-values, and the X-axis is the expected observed negative base 10 logarithm of the *P*-values under the assumption that the *P*-values follow a uniform[0,1] distribution. The dotted lines show the 95% confidence interval for the QQ-plot under the null hypothesis of no association between the SNP and the trait.

5.9 Manhattan Plot

The Manhattan plot is a scatter plot that summarizes GWAS results. The X-axis is the genomic position of each SNP, and the Y-axis is the negative logarithm of the *P*-value obtained from the GWAS model (specifically from the *F*-test for testing H_0: No association between the SNP and trait). Large peaks in the Manhattan plot (i.e., "skyscrapers") suggest that the surrounding genomic region has a strong association with the trait. GAPIT produces one Manhattan plot for the entire genome (Figure 3.4) and individual Manhattan plots for each chromosome.

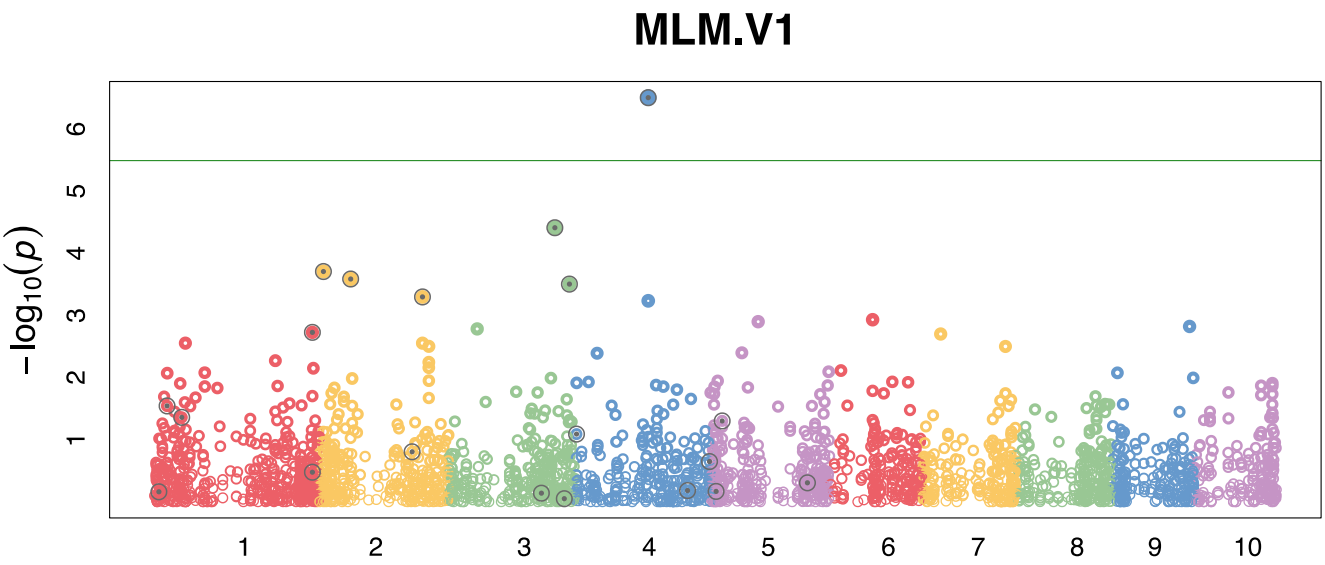


Figure 5.9 Manhattan plot. The X-axis is the genomic position of the SNPs in the genome, and the Y-axis is the negative log base 10 of the *P*-values. Each chromosome is colored differently. SNPs with stronger associations with the trait will have a larger Y-coordinate value.

5.10 Association Table

The GWAS result table provides a detailed summary of appropriate GWAS results. The rows display the results for each SNP above the user-specified minor allele frequency threshold. The SNPs sorted by their *P* values (from smallest to largest).

Table 5.10 GWAS results for all SNPs that were analyzed.

SNP	Chromosome	Position	P.value	maf	nobs	Rsquare.of.Model.without.SNP	Rsquare.of.Model.with.SNP	FDR_Adjusted_P-values
Fea2.4	4	132736424	3.13E-07	0.290035587	281	0.079004463	0.170450593	0.000966617
PZB01223.3	3	192865132	3.88E-05	0.346975089	281	0.079004463	0.137165003	0.059980668
PZA03748.1	2	7481079	0.000196769	0.195729537	281	0.079004463	0.126357611	0.193907122
PZA00394.11	2	56930271	0.00025959	0.145907473	281	0.079004463	0.124537393	0.193907122
PZA00219.6	3	219309117	0.000313461	0.379003559	281	0.079004463	0.123302809	0.193907122

This table provides the SNP id, chromosome, bp position, *P*-value, minor allele frequency (maf), sample size (nobs), R^2 of the model without the SNP, R^2 of the model with the SNP, and adjusted *P*-value following a false discovery rate (FDR)-controlling procedure[25].

5.11 Allelic Effects Table

A separate table showing allelic effect estimates is also included in the suite of GAPIT output files. The SNPs, presented in the rows, are sorted by their position in the genome.

Table 5.11 Information of associated SNPs.

SNP	Chromosome	Position	DF	t Value	std Error	effect
PZB00859.1	1	157104	276	0.255638527	0.501027645	0.128081969
PZA01271.1	1	1947984	276	-0.205390736	0.451124818	-0.092656858
PZA03613.2	1	2914066	276	-1.143780776	0.4887135	-0.558981106
PZA03613.1	1	2914171	276	1.194922452	0.540019499	0.645281424
PZA03614.2	1	2915078	276	0.277681398	0.489962989	0.136053608

This table provides the SNP id, chromosome, bp position, and the allelic effect estimate of each SNP analyzed.

5.12 Compression Profile

There are seven algorithms available to cluster individuals into groups for the compressed mixed linear model. There are also four summary statistics available for calculating the group kinship matrix. When only one group number (i.e., one dimension for the group kinship matrix) is specified, a column chart is created to illustrate the compression profile for 2*log likelihood function (the smaller the better), genetic variance, residual variance and the estimated heritability.

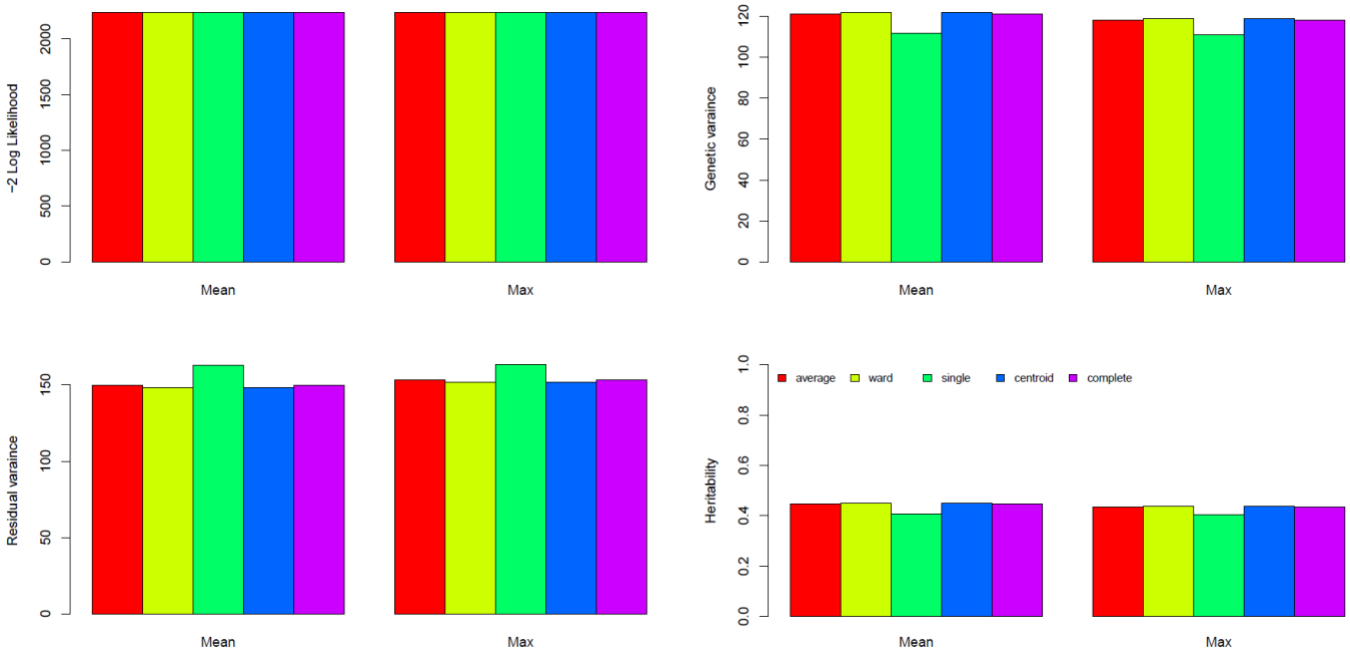


Figure 5.12.1. Compression profile with single group. The X-axis on all graphs display the summary statistic method used to obtain the group kinship matrix. The rectangles with different colors indicate the clustering algorithm used to group individuals.

Note: This graph is not created when multiple groups are specified.

When a range of groups (i.e., a range of dimensions for the group kinship matrix) is specified, a different series of graphs are created. In this situation, the X-axis displays the group number. Lines with different style and colors are used to present the combinations between clustering algorithm and the algorithm to calculate kinship among groups.

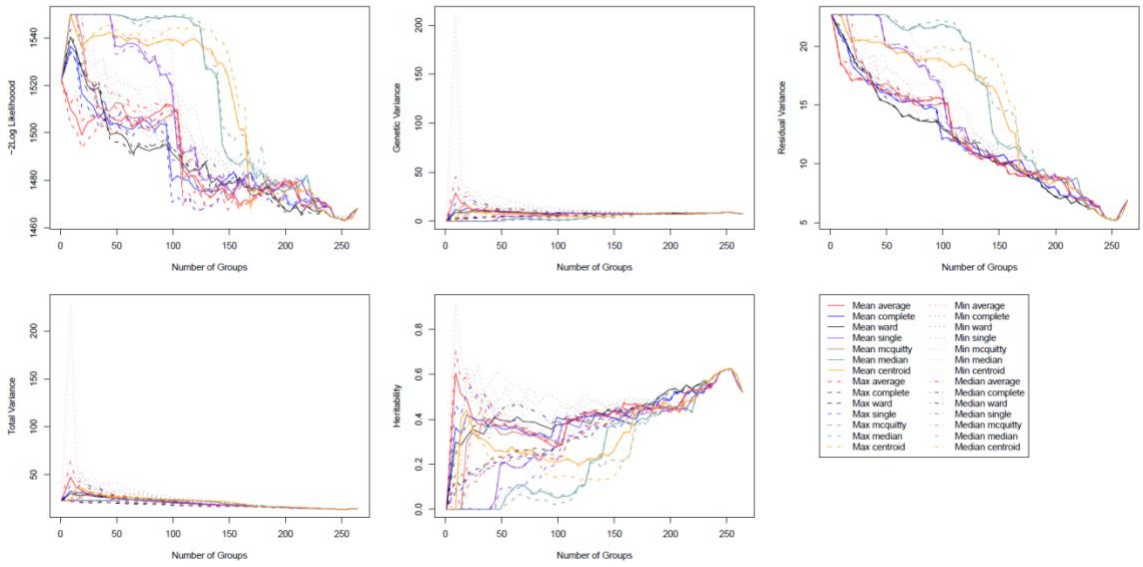


Figure 5.12.2. Compression profile over multiple groups. The X-axis on each graph is the number of groups considered, and the Y-axes on the graphs are the -2*log likelihood function, the estimated genetic variance components, the estimated residual variance component, the estimated total variance, and the heritability estimate. Each clustering method and group kinship type is represented as a line on each graph.

Notice: This graph is not created when only one group is specified.

5.13 The Optimum Compression

Once the optimal compression settings are determined, GAPIT produces a PDF file containing relevant detailed information. This information includes the optimal algorithm to calculate the group kinship matrix, the optimal clustering algorithm, the optimal number of groups, −2*log likelihood function and the estimated heritability.

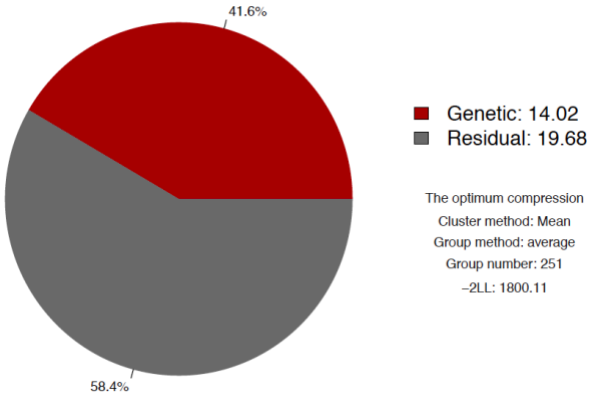


Figure 5.13 The profile for the optimum compression. The optimal method to calculate group kinship is "Mean", the optimal clustering method is "average", the number of groups (ie., the dimension of the group kinship matrix) is 251, the value of -2*log likelihood function is 1800.11, and the heritability is 41.6%.

5.14 Model Selection Results

By selecting "Model.selection = TRUE", forward model selection using the Bayesian information criterion (BIC) will be conducted to determine the optimal number of PCs/Covariates to include for each phenotype. The results summary (below) for model selection are stored in a .csv file called ".BIC.Model.Selection.Results".

Table 5.14.1 Summary for Bayesian information criterion (BIC) model selection results.

Number of PCs/Covariates	BIC (larger is better) - Schwarz 1978	log Likelihood Function Value
0	-816.3884646	-807.9578633
1	-810.4406992	-799.1998974
2	-796.269878	-782.2188759
3	-798.1539656	-781.292763

The number of PCs/Covariates, the BIC value, and the log Likelihood function value are presented. In this table, the optimal number of PCs to include in the GWAS model is 2.

5.15 Multiple traits, environments, or models

There are several new method integrated in. Furthermore, more than single method or trait result, we propose a multiple method or traits Manhattan and QQ plots for comparison with methods or traits, this part is based on MVP library to exhibit. There are two types of Manhattan plots (Orthogon and Roundness).

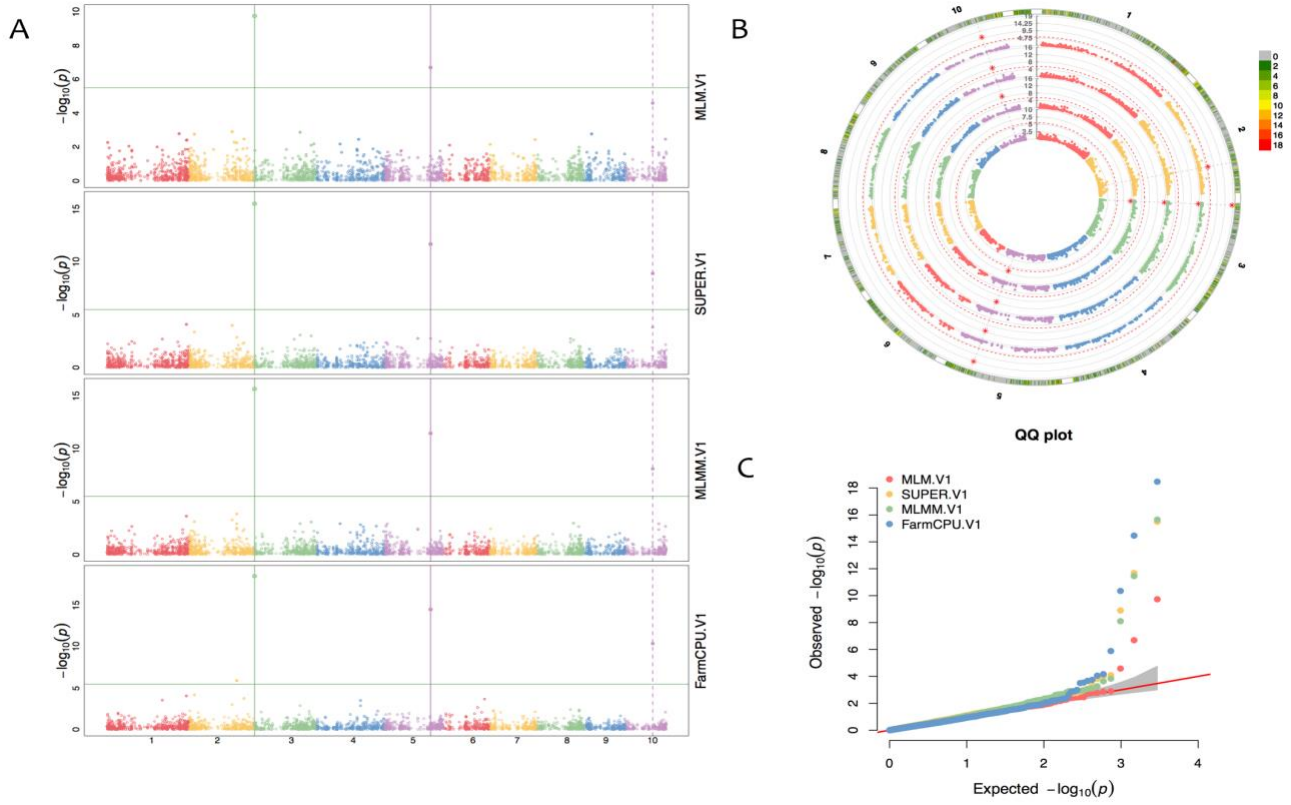


Figure 5.15.1 Multiple traits or methods Manhattan and QQ plot. The dash line in the Figure A indicated the common significant markers were detected by two methods or traits. The solid line in the Figure A indicated the common significant markers were detected by more than two methods or traits.

Another multiple Manhattan plot with multiple symphysis could be created in the GAPIT now. The square, triangle, circle, diamond, and inverted triangle indicated each GWAS results of multiple methods or traits. Users also can define the type of points by "allpch" parameter.

Figure 5.15.2 Multiple traits or methods symphysis Manhattan plot. The dash line indicated the common significant markers were detected by two methods or traits. The solid line indicated the common significant markers were detected by more than two methods or traits.

5.16 Genomic Prediction

The genomic prediction results are saved in a .csv file.

Table 4.16.1 Genomic Breeding values and prediction error variance.

Taxa	Group	RefInf	ID	BLUP	PEV	BLUE	Prediction
33-16	1	1	1	4.61889306	9.86652431	-9.2195121	-4.600619
38-11	2	1	2	4.81979176	10.1420079	-8.8888307	-4.069039
4226	3	1	3	-0.8609631	11.3324656	-8.8692075	-9.7301705
4722	4	1	4	5.68756264	9.29840109	-9.1363798	-3.4488171
A188	5	1	5	-4.404809	11.3810847	-9.1053928	-13.510202
A214N	6	1	6	0.15659848	17.6333752	-8.1102155	-7.9536171
A239	7	1	7	5.12523349	10.8215374	-8.9566662	-3.8314327
A272	8	1	8	-0.4981601	11.08823	-9.6011763	-10.099336

The individual id (taxa), group, RefInf which indicates whether the individual is in the reference group (1) or not (2), the group ID number, the BLUP, and the PEV of the BLUP.

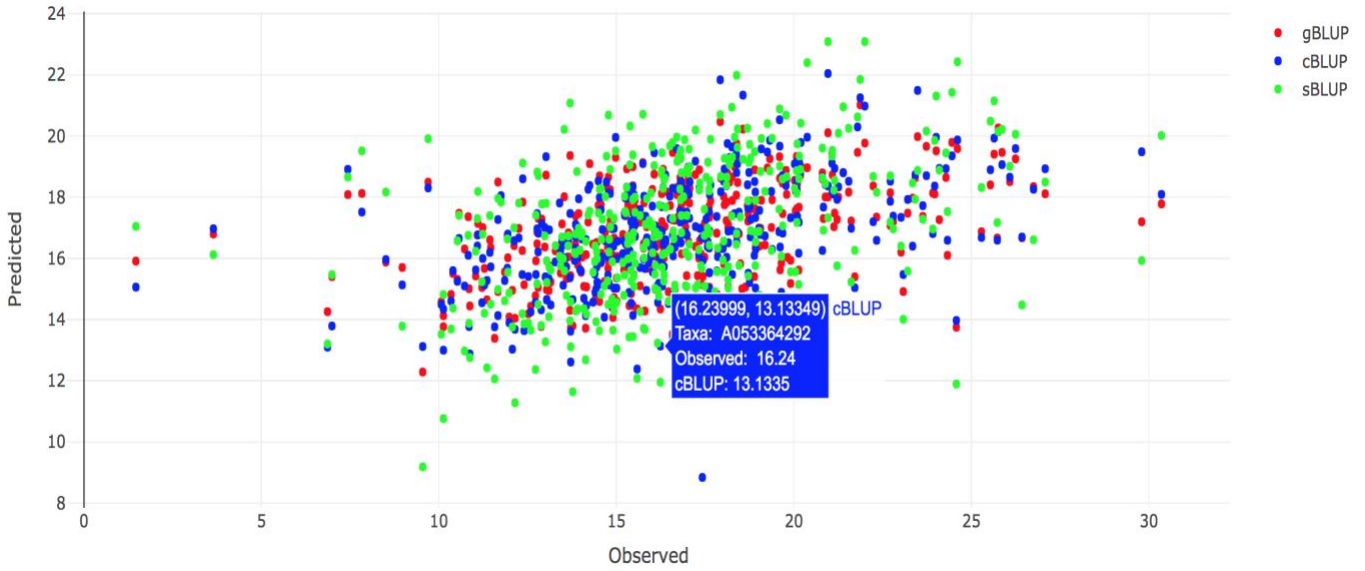


Figure 5.16 Interactive GS plot for gBLUP, cBLUP and sBLUP.

5.17 Distribution of BLUPs and their PEV.

A graph is provided to show the joint distribution of GBV and PEV. The correlation between them is an indicator of selection among the sampled individuals[26].

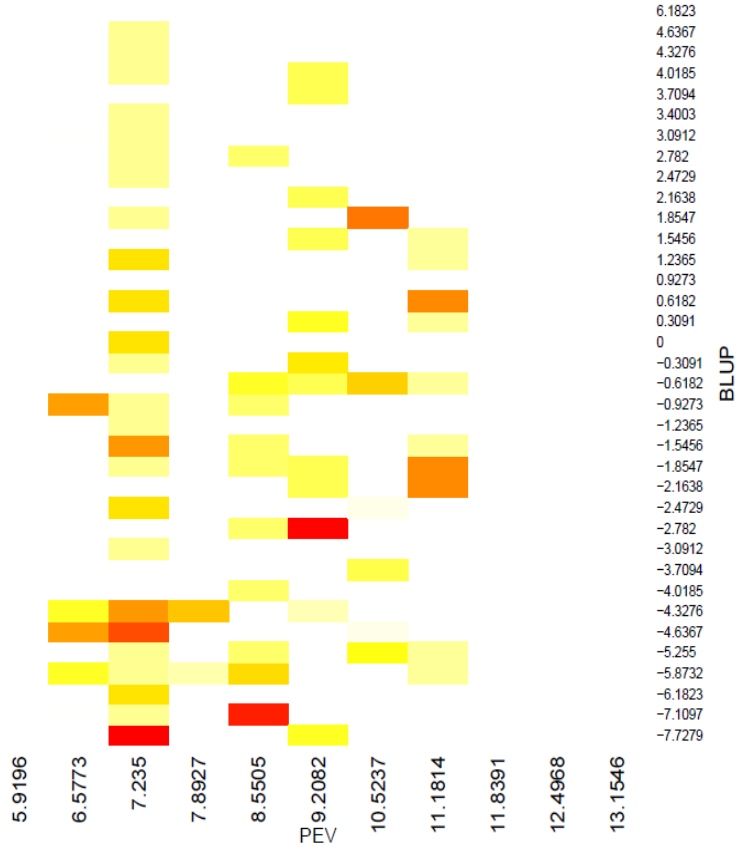


Figure 5.17 Joint distribution of genomic breeding value and prediction error variance.

5.18 Interactive GWAS plot

A graph is provided to show the Interactive Manhattan and QQ plot with “Inter.Plot=T”. These files are HTML and supported to act with mouse. The more details of SNP will be showed when the mouse moved on the point.

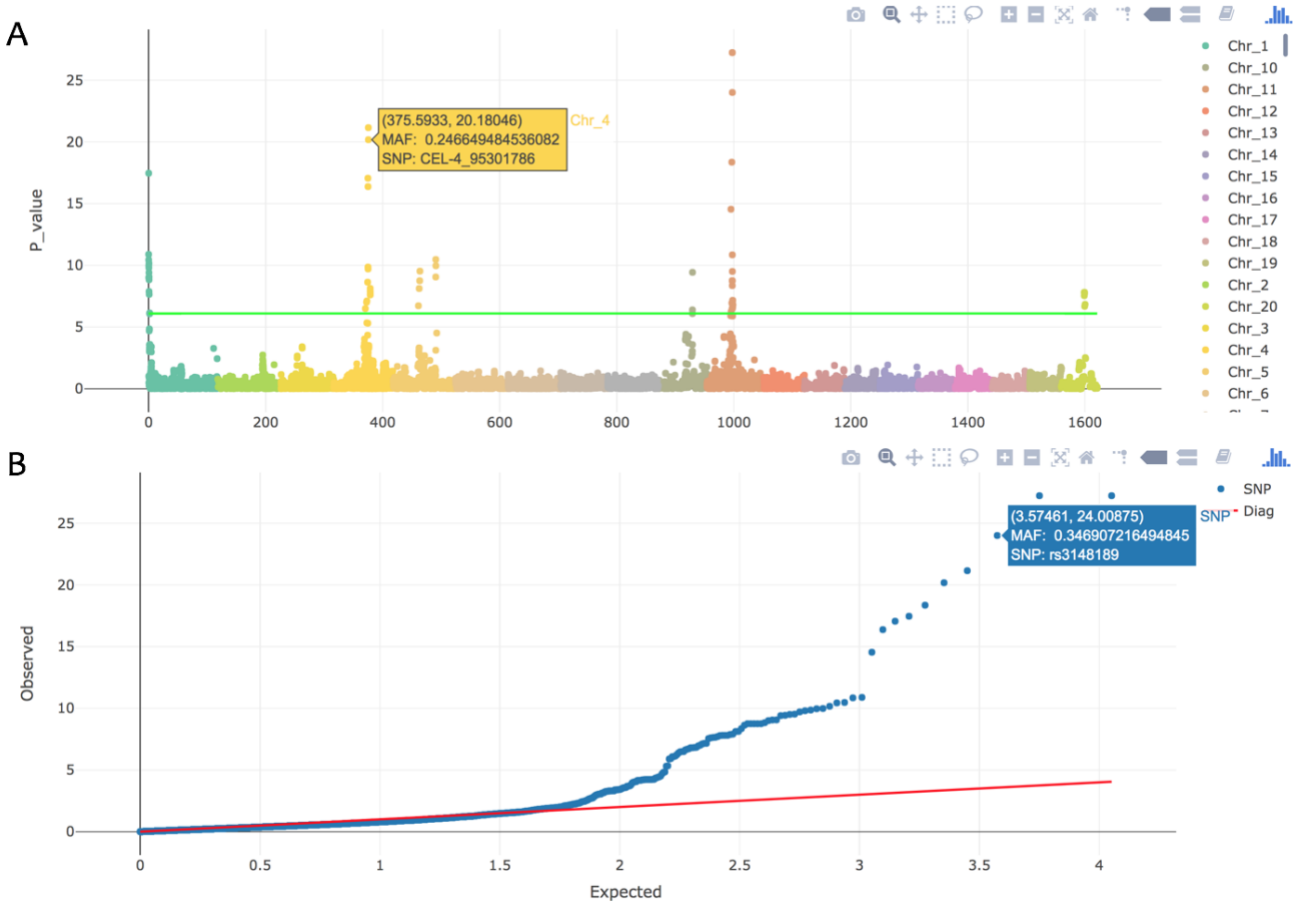


Figure 5.18 Interactive Manhattan and QQ plot.

6 Tutorials

The "Getting started" section should be reviewed before running these tutorials, and it is assumed that the GAPIT package and its required libraries have been installed. These tutorials begin with a scenario requiring minimal user input. Subsequent scenarios require a greater amount of user input. Each scenario involves two steps: reading in the data and then running the GAPIT() function. All tutorials are available on the GAPIT home page, which also contains the R source code and results for all the scenarios.

The GAPIT maize demonstration data (described at www.panzea.org) are from a maize association panel consisting of 281 diverse lines[27]. The genotypic data consist of 3,093 SNPs distributed across the maize genome, and are available in HapMap and numeric format. The three phenotypes included are ear height, days to pollination, and ear diameter. The kinship matrix was calculated using the method of Loiselle *et al.*[28] and the fixed effects used to account for population structure were obtained from STRUCTURE [29].

Notice: It is important that the correct paths to the directories are specified. Please note that two backward slashes ("\\") are necessary when specifying these paths.

6.1 A Basic Scenario

The user needs to provide two data sets (phenotype and genotype) and one input parameter. This parameter, "PCA.total", specifies the number of principal components (PCs) to include in the GWAS model. GAPIT will automatically calculate the kinship matrix using the VanRaden method[30], perform GWAS and genomic prediction with the optimum compression level using the default clustering algorithm (average) and group kinship type (Mean). The scenario assumes that the genotype data are saved in a single file in HapMap format. If the working directory contains the tutorial data, the analysis can be performed by typing these command lines:

```
#Step 1: Set data directory and import files
myY  <- read.table("mdp_traits.txt", head = TRUE)
myG <- read.table("mdp_genotype_test.hmp.txt", head = FALSE)

#Step 2: Run GAPIT
myGAPIT <- GAPIT(
Y=myY,
G=myG,
PCA.total=3
)
```

6.2 Enhanced Compression

In this scenario, the user can specify additional clustering algorithms (controlled by the "kinship.cluster" parameter) and kinship summary statistic (controlled by the "kinship.group" parameter). The default is kinship.cluster="average", and kinship.group="Mean". Their expansion, the Enriched CMLM[8] improves statistical power. Additionally, a specific range group numbers (i.e., dimension of the kinship matrix) can be specified. This range is controlled by the "group.from", "group.to", and "group.by" parameters. The analysis can be performed by typing these command lines:

```
#Step 1: Set data directory and import files
myY  <- read.table("mdp_traits.txt", head = TRUE)
myG <- read.table("mdp_genotype_test.hmp.txt", head = FALSE)

#Step 2: Run GAPIT
myGAPIT <- GAPIT(
Y=myY,
G=myG,
PCA.total=3,
kinship.cluster=c("average", "complete", "ward"),
kinship.group=c("Mean", "Max"),
model="CMLM"
)
```

6.3 User-inputted Kinship Matrix and Covariates

This scenario assumes that the user provides a kinship matrix and covariate file. The kinship matrix or covariates (e.g., PCs) may be calculated previously or from third party software. When the PCs are input in this way, the parameter "PCA.total" should be set to 0 (default). Otherwise, PCs will be calculated within GAPIT, resulting in a singular design matrix in all model fitted for GWAS. The analysis can be performed by typing these command lines:

```
#Step 1: Set data directory and import files
myY  <- read.table("mdp_traits.txt", head = TRUE)
myG <- read.table("mdp_genotype_test.hmp.txt", head = FALSE)
myKI <- read.table("KSN.txt", head = FALSE)
myCV <- read.table("mdp_PC", head = TRUE)

#Step 2: Run GAPIT
myGAPIT <- GAPIT(
Y=myY,
G=myG,
KI=myKI,
CV=myCV
)
```

6.4 Multiple Genotype Files

In this scenario, the HapMap genotypic data set from Scenario 1 is subdivided into multiple genotype files , one for each chromosome. This scenario mimics the situation where the genotype file is too large to be handled in R. When this situation arises, all genotype files need to have a common name and extensions, as well as a sequential number (e.g., "mdp_genotype_chr1.hmp.txt", "mdp_genotype_chr2.hmp.txt", …). The starting and ending file are indicated by the "file.from" and "file.to" parameters. The common file name (e.g., "mdp_genotype_chr") and file name extension (e.g., "hmp.txt") are passed to GAPIT through the "file.G", "file.Ext.G" parameters, respectively. When "file.path" is not provided, GAPIT try to get the data from the current working directory. The analysis can be performed by typing these command lines:

```
#Step 1: Set data directory and import files
myY  <- read.table("mdp_traits.txt", head = TRUE)

#Step 2: Run GAPIT
myGAPIT <- GAPIT(
Y=myY,
PCA.total=3,
file.G="mdp_genotype_chr",
file.Ext.G="hmp.txt",
file.from=1,
file.to=10,
file.path="C:\\myGAPIT\\"
)
```

The three genotype file used in these scenario are from the file used in Tutorial 5.1. Their results should be identical.

6.5 Numeric Genotype Format

In this scenario, the genotype data set from Scenario 1 is formatted differently, specifically in numerical format. Two genotype files are required. One file contains the genotypic data, and the other contains the chromosome and base pair position of each SNP. These are passed to GAPIT through the "GD" and "GM" parameters, respectively. The analysis can be performed by typing these command lines:

```
#Step 1: Set data directory and import files
myY  <- read.table("mdp_traits.txt", head = TRUE)
myGD <- read.table("mdp_numeric.txt", head = TRUE)
myGM <- read.table("mdp_SNP_information.txt" , head = TRUE)

#Step 2: Run GAPIT
myGAPIT <- GAPIT(
Y=myY,
GD=myGD,
GM=myGM,
PCA.total=3
)
```

6.6 Numeric Genotype Format in Multiple Files

In this scenario, the numeric genotype data set from Scenario 6 is subdivided into multiple genotype files. The common name and extension of genotype data file are passed to GAPIT through "file.GD" and "file.Ext.GD" parameters, respectively. Similarly, the common name and extension of genotype map file are passed to GAPIT through the "file.GM" and "file.Ext.GM" parameters, respectively. The analysis can be performed by typing these command lines:

```
#Step 1: Set data directory and import files
myY  <- read.table("mdp_traits.txt", head = TRUE)

#Step 2: Run GAPIT
myGAPIT <- GAPIT(
Y=myY,
PCA.total=3,
file.GD="mdp_numeric",
file.GM="mdp_SNP_information",
file.Ext.GD="txt",
file.Ext.GM="txt",
file.from=1,
file.to=3,
)
```

The three genotype file used in these scenario are the splits from the file used in the previous scenario. Their results should be identical.

6.7 Fractional SNPs for Kinship and PCs

The computations of kinship and PCs are extensive with large number of SNPs. Sampling a fraction of it would reduce computing time. More importantly, it would give very similar result with appropriate number of SNPs sampled. The fraction can be controlled by "Ratio" parameter in GAPIT. The sampling scheme is random. A line of "SNP.fraction=0.6" is added to the previous scenario which has 3,093 SNPs:

```
#Step 1: Set data directory and import files
myY  <- read.table("mdp_traits.txt", head = TRUE)

#Step 2: Run GAPIT
myGAPIT <- GAPIT(
Y=myY,
PCA.total=3,
file.GD="mdp_numeric",
file.GM="mdp_SNP_information",
file.Ext.GD="txt",
file.Ext.GM="txt",
file.from=1,
file.to=3,
SNP.fraction=0.6
)
```

6.8 Memory saving

With large amount of individuals, loading a entire large genotype dataset could be difficult. GAPIT load a fragment of it each time. The default of the fragment size is 512 SNPs. This number can be changed with "file.fragment" parameter in GAPIT. Here is an example of using "file.fragment =128".

```
#Step 1: Set data directory and import files
myY  <- read.table("mdp_traits.txt", head = TRUE)

#Step 2: Run GAPIT
myGAPIT <- GAPIT(
Y=myY,
PCA.total=3,
file.GD="mdp_numeric",
file.GM="mdp_SNP_information",
file.Ext.GD="txt",
file.Ext.GM="txt",
file.from=1,
file.to=3,
SNP.fraction=0.6,
file.fragment = 128
)
```

This scenario is the same as previous scenario except changing "file.fragment" from default (512) to 128. As SNPs (minimume of two) are sampled withing each fragment, the final SNPs sampled would be different for different length of fragment when the SNP sample fraction is less than 100%. The results in this scenario would be different from the previous one.

6.9 Model selection

The degree of correlation with population structure varies from trait to trait. Therefore, the full set of PCs selected to account for population structure in the GWAS model are not necessary for all traits. As such, GAPIT has the capability to conduct Bayesian information criterion (BIC)-based model selection to find the optimal number of PCs for inclusion in the GWAS models. Model selection is activated by selecting "Model.selection = TRUE". The results for the BIC model selection procedure are summarized in the ".BIC.Model.Selection.Results.csv" output file.

```
myY  <- read.table("mdp_traits.txt", head = TRUE)
myG <- read.table("mdp_genotype_test.hmp.txt", head = FALSE)

#Step 2: Run GAPIT
myGAPIT <- GAPIT(
Y=myY,
G=myG,
PCA.total=3,
Model.selection = TRUE
)
```

6.10 SUPER

GAPIT also implements the SUPER GWAS method[9], which extracts a small subset of SNPs and uses them in FaST-LMM. This method not only retains the computational advantage of FaST-LMM, but also increases statistical power.

```
#Step 1: Set data directory and import files
myY  <- read.table("mdp_traits.txt", head = TRUE)
myG <- read.table("mdp_genotype_test.hmp.txt" , head = FALSE)

#Step 2: Run GAPIT
myGAPIT_SUPER <- GAPIT(
Y=myY[,c(1,2)],
G=myG,
PCA.total=3,
model="SUPER"
)
```

6.11 MLMM

Multiple Loci Mixied linear Model is published by Segura in 2012. The code of MLMM in GAPIT is:

```
myY  <- read.table("mdp_traits.txt", head = TRUE)
myG <- read.table("mdp_genotype_test.hmp.txt", head = FALSE)

#Step 2: Run GAPIT
myGAPIT <- GAPIT(
Y=myY,
G=myG,
PCA.total=3,
model="MLMM"
)
```

6.12 Farm-CPU

Fixed and random model Circulating Probability Unification (FarmCPU) is published by Xiaolei in 2016. The code of Farm-CPU in GAPIT is:

```
myY  <- read.table("mdp_traits.txt", head = TRUE)
myG <- read.table("mdp_genotype_test.hmp.txt", head = FALSE)

#Step 2: Run GAPIT
myGAPIT <- GAPIT(
Y=myY,
G=myG,
PCA.total=3,
model="FarmCPU"
)
```

6.13 BLINK

Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK) is published by Meng in 2018. The code of BLINK used in GAPIT is:

```
myY  <- read.table("mdp_traits.txt", head = TRUE)
myG <- read.table("mdp_genotype_test.hmp.txt", head = FALSE)

#Step 2: Run GAPIT
myGAPIT <- GAPIT(
Y=myY,
G=myG,
PCA.total=3,
model="Blink"
)
```

The BLINK C version execute file should be in the working directory and change mod 777.
(chmod 777 blink_versions)
The execute file can be downloaded at https://github.com/Menggg/BLINK/blob/master/

6.14 Multiple model

The GAPIT provide an approach for comparison of multiple methods in GWAS. All GWAS methods in the GAPIT can be used in here:

```
myY  <- read.table("mdp_traits.txt", head = TRUE)
myGD=read.table("http://zzlab.net/GAPIT/data/mdp_numeric.txt",head=T)
myGM=read.table("http://zzlab.net/GAPIT/data/mdp_SNP_information.txt",head=T)

#Step 2: Run GAPIT
myGAPIT <- GAPIT(
Y=myY,
GD=myGD,
GM=myGM,
PCA.total=3,
Multiple_analysis=TRUE,
model=c("GLM","MLM","MLMM","FarmCPU","Blink")
)
```

6.15 gBLUP

The gBLUP is based on the Mixed linear Model to predict phenotype, BLUP, and BLUE value.

```
myY  <- read.table("mdp_traits.txt", head = TRUE)
myGD=read.table("http://zzlab.net/GAPIT/data/mdp_numeric.txt",head=T)
myGM=read.table("http://zzlab.net/GAPIT/data/mdp_SNP_information.txt",head=T)

#Step 2: Run GAPIT
myGAPIT <- GAPIT(
Y=myY[,c(1,2)],
GD=myGD,
GM=myGM,
PCA.total=3,
model=c("gBLUP")
)
```

6.16 cBLUP

The cBLUP is based on the compressed Mixed linear Model. It used optimum compression group kinship instead of individual kinship.

```
myY  <- read.table("mdp_traits.txt", head = TRUE)
myGD=read.table("http://zzlab.net/GAPIT/data/mdp_numeric.txt",head=T)
myGM=read.table("http://zzlab.net/GAPIT/data/mdp_SNP_information.txt",head=T)

#Step 2: Run GAPIT
myGAPIT <- GAPIT(
Y=myY[,c(1,2)],
GD=myGD,
GM=myGM,
PCA.total=3,
model=c("cBLUP")
)
```

6.17 sBLUP

The sBLUP is based on the SUPER Model. It used pseudo QTNs to build individual kinship.

```
myY  <- read.table("mdp_traits.txt", head = TRUE)
myGD=read.table("http://zzlab.net/GAPIT/data/mdp_numeric.txt",head=T)
myGM=read.table("http://zzlab.net/GAPIT/data/mdp_SNP_information.txt",head=T)

#Step 2: Run GAPIT
myGAPIT <- GAPIT(
Y=myY[,c(1,2)],
GD=myGD,
GM=myGM,
PCA.total=3,
model=c("sBLUP")
)
```

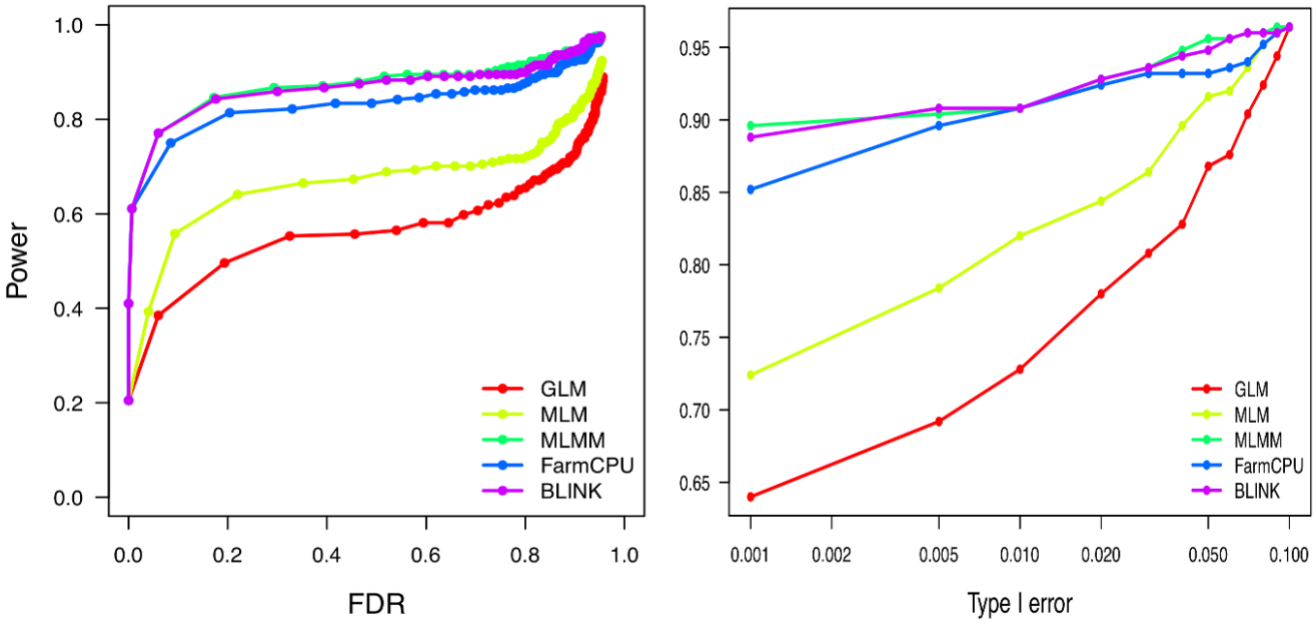
7 Prototype

The usage of GAPIT described in previous chapters barely require knowledge of R. Users can simply copy/paste the command lines from the user manual with a minimal keyboard typing such as changing file names and path. All the results are saved in the format of text files and PDF files. GAPIT also output R objects which can be used for advance purposes, including: 1) developing new statistical approaches or software package using GAPIT output as a starting point; 2) comparing GAPIT with other new or existing statistical methods or software packages; 3) studying a specific result from GAPIT. Using these objects require knowledge of R. This chapter give examples to use GAPIT output R objects.

7.1 Statistical power comparison among methods

GAPIT.Power.compare() is an example function use multiple functions in GAPIT and the output R objects to perform comparison of statistical power using different models. The deteils can be found in GAPIT source code. The following is an example to use the function. Note: running 100 replicates may take more than a day to finsh.

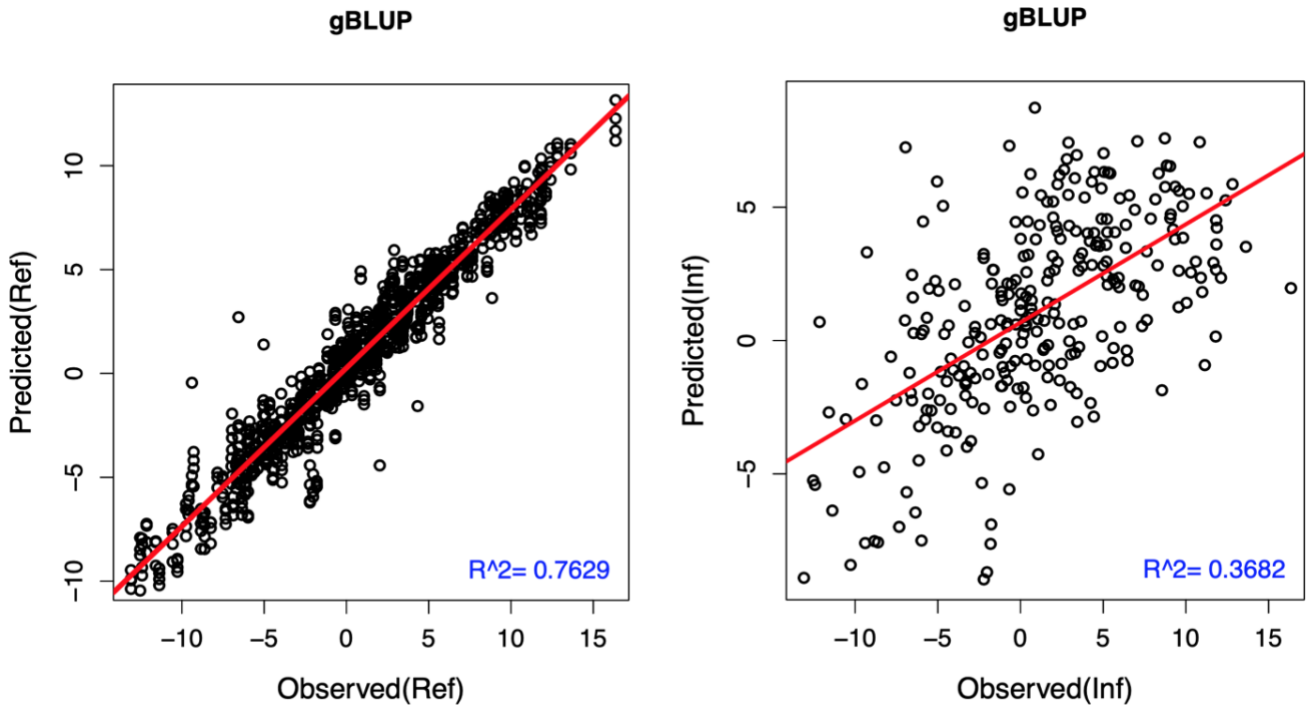
```
myGD <-read.table("mdp_numeric.txt", head = TRUE)
myGM <-read.table("mdp_SNP_information.txt", head = TRUE)
GAPIT.Power.compare(
myGD=myGD,
myGM=myGM,
nrep=100,
h2=0.9,
all.method=c("GLM","MLM","MLMM","FarmCPU","BLINK"),
NQTN=5)
```



7.2 Genomic selection

GAPIT.Prediction() is another example of using GAPIT subfunctions and R objects outputs for GS

```
myY<-read.table("mdp_traits.txt", head = TRUE)
myGD <-read.table("mdp_numeric.txt", head = TRUE
myGM <-read.table("mdp_SNP_information.txt", head = TRUE)
set.seed(99163)
GAPIT.Validation(
   Y=myY[,1:2],
   model=c("gBLUP"),
   GD=myGD,
   GM=myGM,
   PCA.total=3,
   file.output=T,
   nfold=5
)
```



7.3 Cross validation with replacement

Here we demonstrate an example of studying a specific output of GAPIT, specifically to investigate the accuracy of genome prediction through cross validation.

First, we randomly set 25% of original phenotype (Y) as missing (NA) and generate a genomic prediction model by using their kinship. Then we record the correlation between the predicted phenotypic values and the original phenotype. We repeat this process for 1000 times. The average correlation is used as the criteria of genome prediction accuracy. The corresponding R code is displayed in the following box. The accuracy (correlation coefficient) over the 100 replicates (for demonstration purposes only; we suggest

using 1000 replications) were 0.9203 and 0.6749 in the reference and inference (cross validation), respectively. The standard deviations were 0.078 and 0.0054 in the reference and inference, respectively.

R code for cross validation with replacement

```
#Import files
#############################################################################################
myY  <- read.table("mdp_traits.txt", head = TRUE)
myKI <- read.table("KSN.txt", head = FALSE)
myCV <- read.table("mdp_PC", head = TRUE)

#Initial
#############################################################################################
t=100 #total replicates
s=1/5 #sample of inference, e.g. set it to 1/5 for five fold cross validation
Y.raw=myY[,c(1,3)]#choos a trait
Y.raw=Y.raw[!is.na(Y.raw[,2]),] #Remove missing data
n=nrow(Y.raw)
n.missing=round(n*s)
storage.ref=matrix(NA,t,1)
storage.inf=matrix(NA,t,1)

#Loop on replicates
for(rep in 1:t){

#Set missing data
sample.missing=sample(1:n,n.missing)
if(n.missing>0){ Y0=Y.raw[-sample.missing,]
}else{Y0=Y.raw}

#Prediction
myGAPIT <- GAPIT(
Y=Y0,
KI=myKI,
CV=myCV,
model="gBLUP"
)
prediction=myGAPIT$Pred

#Separate reference (with phenotype) and inference (without phenotype)
prediction.ref=prediction[prediction[,3]==1,]
prediction.inf=prediction[prediction[,3]==2,]

#Merge prediction with original Y
YP.ref <- merge(Y.raw, prediction.ref, by.x = "Taxa", by.y = "Taxa")
YP.inf <- merge(Y.raw, prediction.inf, by.x = "Taxa", by.y = "Taxa")

#Calculate correlation  and store them
r.ref=cor(as.numeric(as.vector(YP.ref[,2])),as.numeric(as.vector(YP.ref[,6]) ))
r.inf=cor(as.numeric(as.vector(YP.inf[,2])),as.numeric(as.vector(YP.inf[,6]) ))
storage.ref[rep,1]=r.ref
storage.inf[rep,1]=r.inf
}#End of for (rep in 1:t)

storage=cbind(storage.ref,storage.inf)
colnames(storage)=c("Reference","Inference")
write.table(storage, "GAPIT.Cross.Validation.txt", quote = FALSE, sep = "\t", row.names = TRUE,col.names = NA)
```

7.4 Cross validation without replacement

Cross validation can also be performed by excluding one or a set of individuals in the reference to derive predicted phenotypic values from the genomic prediction model. Using this approach, this process can be

repeated until all the individuals have been excluded at least once. The correlation between the originals and the prediction (might be more than once) is used as the accuracy of prediction. The following demonstrate the process with the same data in previous section.

```
#Initial
#################################################################################################
nj= 200 # number of Jack Knifes, nj>0, nj!=1

Y.raw=myY[,c(1,3)]#choos a trait
Y.raw=Y.raw[!is.na(Y.raw[,2]),] #Remove missing data
n=nrow(Y.raw)

if(nj>=1){nLoop=nj
}else{
  nLoop=1/nj
}
assignment=ceiling((1:n)/(n/nLoop))
randomization=sample(1:n,n)
assignment=assignment[randomization]
nLoop=ceiling(nLoop)

#Loop on replicates
for(rep in 1:nLoop){

#Set missing data
if(nj>=1){Y0=Y.raw[assignment!=rep,]
}else{
  Y0=Y.raw[assignment==rep,]
}
#Prediction
myGAPIT <- GAPIT(
Y=Y0,
KI=myKI,
CV=myCV,
model=" gBLUP "
)
prediction=myGAPIT$Pred

#Separate reference (with phenotype) and inference (without phenotype)
if(rep==1){
  prediction.inf=prediction[prediction[,3]==2,]
  }else{
  prediction.inf=rbind(prediction.inf,prediction[prediction[,3]==2,] )
  }
}#End of for (rep in 1:t)

#Merge prediction with original Y
YP.inf <- merge(Y.raw, prediction.inf, by.x = "Taxa", by.y = "Taxa")

#Calculate correlation  and store them
r.inf=cor(as.numeric(as.vector(YP.inf[,2])),as.numeric(as.vector(YP.inf[,6]) ))
write.table(YP.inf, "GAPIT.Jack.Knife.txt", quote = FALSE, sep = "\t", row.names = TRUE,col.names = NA)
print(r.inf)
```

7.5 Convert HapMap format to numerical

Many software require genotype data in the numerical format. GAPIT can perform such conversion with a few lines of code as follows.

myG <- read.table("mdp_genotype_test.hmp.txt", head = FALSE)
myGAPIT <- GAPIT(G=myG, output.numerical=TRUE)
myGD= myGAPIT$GD
myGM= myGAPIT$GM

8 Appendix

8.1 GAPIT Biography

Date	Version	Event
11-May-11	1.2	First public release with following method implemented: • Principal component method to encounter population structure (Price). • Unified mixed model to encounter both population structure and kinship. • EMMA method to improve the speed to estimate variance components (ratio). • Compressed mixed model to improve statistical power and speed • P3D or EMMAx to improve speed by estimating population parameters (e.g. variances and grouping) only once.
13-Jun-11	1.22	Genotype in numerical format in addition to hapmap format
2-Sep-11	1.31	Reading fragment within single genotype file to save memory
17-Sep-11	1.36	Interface change for prototyping
24-Oct-11	1.41	Option to impute missing genotypes as middle, major, minor, present/absent.
1-Nov-11	1.42	Matrix partitioning to improve speed (5-10 fold faster)
7-Dec-11	2.01	FaST-LMM method implemented
19-May-12	2.19	Splitting big genotype file into small ones
8-Jul-12	NA	GAPIT Bioinformatics paper accepted for publication
8-Nov-12	2.2	Automatic sorting taxa in kinship for regular MLM
14-Feb-13	2.25	Confidence interval on QQ plot
26-Apr-13	2.26	Labeling QTNs on Manhattan plot
15-Jul-14	2.27	SUPER implemented
20-Aug-14	2.28	ECMLM implemented
25-Oct-14	2.29	VanRaden kinship algorithm with centralization
28-Feb-15	2.3.41	Enrichment on output figures and tables
5-Sep-15	by date	Uniform output across models (Version 3 initiation)
12-Oct-15		Simulation of category phenotype
22-Oct-15		Compression BLUP (cBLUP) for genomic prediction
1-Apr-16		GAPIT version 2 paper published by Plant Genome
4-Apr-16		Pedigree-like marker based kinship
4-Apr-16		SUPER BLUP (sBLUP) for genomic prediction
31-Mar-17		Indicating LD to the strongest associated marker above threshold
2-May-18		Add cBLUP and sBLUP demo script into user manual
Jul 17,2019		Add command for loading GAPIT from GitHub
Sep 4, 2021		GAPIT version 3 paper published on Genomics, Proteomics & Bioinformatics

8.2 Frequently Asked Questions

1. How to cite GAPIT?

A: Citation may vary based on the usage of versions (the first version[1] ,the second version[2], or the third version[3]), and methods involved, such as the regular MLM[20] methods, CMLM[6], ECMLM[8], SUPER[9], P3D[6], FarmCPU[31], and BLINK[13]. Here is an example: "The GWAS was conducted using the BLINK model[17] implemented in GAPIT R Software package (version 3)[3]".

2. What do I do if I get frustrated?

A: Try to go through this Q/A list and GAPIT Forum first before asking help from GAPIT team. If you need to contact GAPIT team, email to Dr. Xiaolei Liu (email: xll19870827@hotmail.com) on questions related to FarmCPU, Dr. Meng Huang (email: meng.huang.cn@gmail.com) for questions related to BLINK, or Dr. Jiabo Wang (email: wangjiaboyifeng@163.com) on rest questions. In all cases (Forum or Emails), please state your names and your institutions.

3. Why GAPIT has different results from other software?

A: The most common reasons to have different results is that these software packages use different genetic models (e.g. additive vs. additive + dominant), statistical models (e.g. GLM, MLM, CMLM, and ECMLM), and processing of missing data. The GAPIT Bioinformatics paper demonstrated that GAPIT and TASSEL gave identical results for inbred (additive only) without missing values for using MLM.

4. There are many methods implemented in GAPIT, which one should I use?

A: Literature demonstrated the order of statistical power: BLINK > FarmCPU> MLMM > SUPER > ECMLM > CMLM > MLM > GLM.

5. How many PCs to include?

A: There no clear answer for this question. However, here are the two ways most of people do. 1) The number of principal components (PCs) included in the GWAS models can be adjusted in GAPIT. To help determine the number of PCs that adequately explain population structure, a screen plot is provided in the GAPIT output (if at least one PC is selected for inclusion into the final model). Once the ideal number of PCs is determined, GAPIT should be reran with this number PCs included in the GWAS models; 2) Use BIC-based model selection (activated by writing Model.selection = TRUE in the GAPIT() function) to determine the "optimal" number of PCs. The optimal is in quotations because no evidence has been found for optimum statistical power.

6. Is it feasible I compare different models on my data?

A: Yes, you can compare different models implemented in GAPIT or other software packages through simulation. All you need is a genotype file. The demo source code is amiable at the Workshop of Assessment of Statistical Power in GWAS (http://zzlab.net/WorkshopISU).

7. How do I report an error?

A: In order to fix the problem, please copy and paste the error message from the R environment and attach your R source code and the dataset that allow us to repeat the error.

8. What should I do with "Error in file (file, "rt") : cannot open the connection"?

A: In most cases this error is caused by incorrect file name or number of file specific is more than exist.

9. What should I do with "Error in GAPIT (... : unused argument(s) ..."?

A: In most cases this error is caused by incorrect spelling of GAPIT key word such as upper or lower case, or missing a comma.

10. How deal with "Error in solve.default(crossprod(X, X)) : system is computationally singular"?

A: Check covariate variables and remove the ones that are linear dependent with others.

11. How to fix the error of using covariates from STRUCTURE as fixed effects?

A: This error is occurring because the covariates from STRUCTURE are linearly dependent. In particular, for a given individual/taxa, these covariates sum to 1. To circumvent this error, remove one of the STRUCTURE covariates from the corresponding input file.

12. Should I remove SNPs with MAF below 5%?

A: The answers are Yes/No. Rare SNPs with low Minor Allele Frequency (MAF) usually cause false positives, especially for small samples and traits that do not have normal distribution. However, many causal genetic variants are rare. A recommended practice is to not remove them, but interpret them with caution.

13. My trait was measured in multiple environments, how do I use them simultaneously?

A: They can be averaged across environments, and use means by GAPIT. The genetic and environmental interaction was implemented in a separated software package: GEMT (http://zzlab.net/GEMT).

14. Is it OK to analyze binary traits (case-control) with GAPIT?

A: Yes, there are many applications.

15. Does normality transformation help?

A: Yes, non-normality, rare variants and small samples jointly cause false positives. The transformation helps in case of small samples and SNPs with low MAF.

16. Should I use PCs or Q matrix?

A: Keyan Zhao and et al (PLoS Genetics, 2007) compared the two methods and demonstrated that they had similar statistical power.

REFERENCES

1. Lipka, A. E. *et al.* GAPIT: genome association and prediction integrated tool. *Bioinformatics* **28**, 2397–2399 (2012).
2. Tang, Y. *et al.* GAPIT Version 2: An Enhanced Integrated Tool for Genomic Association and Prediction. *Plant Genome* **9**, 1–9 (2016).
3. Wang, J. & Zhang, Z. GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction. *Genomics. Proteomics Bioinformatics* **19**, 1–12 (2021).
4. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904–909 (2006).
5. Yu, J. M. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**, 203–208 (2006).
6. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**, 355–360 (2010).
7. Zhang, Z., Todhunter, R. J., Buckler, E. S. & Van Vleck, L. D. Technical note: Use of marker-based relationships with multiple-trait derivative-free restricted maximal likelihood. *J. Anim. Sci.* **85**, 881–885 (2007).
8. Li, M. *et al.* Enrichment of statistical power for genome-wide association studies. *BMC Biol.* **12**, 73 (2014).
9. Wang, Q., Tian, F., Pan, Y., Buckler, E. S. E. S. & Zhang, Z. A SUPER Powerful Method for Genome Wide Association Study. *PLoS One* **9**, e107684 (2014).
10. Segura, V. *et al. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations*. *Nature Genetics* **44**, 825–830 (2012).
11. Liu, X., Huang, M., Fan, B., Buckler, E. S. & Zhang, Z. Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. *PLoS Genet.* **12**, e1005767 (2016).
12. Wang, J. *et al.* Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traits. *Heredity (Edinb).* **121**, 648–662 (2018).
13. Huang, M., Liu, X., Zhou, Y., Summers, R. M. & Zhang, Z. BLINK: A package for the next level of genome-wide association studies with both individuals and markers in the millions. *Gigascience* giy154 (2019). doi:10.1093/gigascience/giy154
14. Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature* **465**, 627–631 (2010).
15. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* **42**, 355–360 (2010).
16. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* **81**, 559-575. (2007).
17. Huang, M. *et al.* BLINK: A package for the next level of genome-wide association studies with both individuals and markers in the millions. *Gigascience* giy154 (2018). doi:10.1093/gigascience/giy154
18. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
19. Zhao, K. *et al.* An Arabidopsis example of association mapping in structured samples. *PLoS Genet.* **3**, 0071–0082 (2007).
20. Yu, J. M. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**, 203–208 (2006).
21. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**, 355–60 (2010).
22. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**, 203–208 (2006).
23. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559–575 (2007).
24. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).

25. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to mutliple testing. *J.R. Statis. Soc. B.* **57**, 289–300 (1995).

26. Guo, G. *et al.* Canine hip dysplasia is predictable by genotyping. *Osteoarthr. Cartil.* **19**, 420–429 (2011).

27. Flint-Garcia, S. A. *et al.* Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J* **44**, 1054–1064 (2005).

28. Loiselle, B. A., Sork, V. L., Nason, J. & Graham, C. Spatial Genetic-Structure of a Tropical Understory Shrub, Psychotria Officinalis (Rubiaceae). *Am. J. Bot.* **82**, 1420–1425 (1995).

29. FALUSH, D., STEPHENS, M. & PRITCHARD, J. K. Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol. Ecol. Notes* **7**, 574–578 (2007).

30. VanRaden, P. M. Efficient methods to compute genomic predictions. *J Dairy Sci* **91**, 4414–4423 (2008).

31. Liu, X., Huang, M., Fan, B., Buckler, E. S. E. S. & Zhang, Z. Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. *PLoS Genet.* **12**, e1005767 (2016).