Performing Genome-Wide Association Studies with Multiple Models Using GAPIT

Jiabo Wang, You Tang, and Zhiwu Zhang

Abstract

Genome-wide association study (GWAS) is based on the linkage disequilibrium (LD) between phenotypes and genetic markers covering the whole genome. Besides the genetic linkage between the genetic markers and the causal mutations, many other factors contribute to the LD, including selection and nonrandom mating formatting population structure. Many methods have been developed with accompany of corresponding software such as multiple loci mixed model (MLMM). There are software packages that implement multiple methods to reduce the learning curve. One of them is the Genomic Association and Prediction Integrated Tool (GAPIT), which implemented eight models including GLM (General Linear Model), Mixed Linear Model (MLM), Compressed MLM, MLMM, SUPER (Settlement of mixed linear models Under Progressively Exclusive Relationship), FarmCPU (Fixed and random model Circulating Probability Unification), and BLINK (Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway). Besides the availability of multiple models, GAPIT provides comprehensive functions for data quality control, data visualization, and publication-ready quality graphic outputs, such as Manhattan plots in rectangle and circle formats, quantile–quantile (QQ) plots, principal component plots, scatter plot of minor allele frequency against GWAS signals, plots of LD between associated markers and the adjacent markers. GAPIT developers and users established a community through the GAPIT forum (https://groups.google.com/g/gapit-forum) with over 600 members for asking questions, making comments, and sharing experiences. In this chapter, we detail the GAPIT functions, input data frame, output files, and example codes for each GWAS model. We also interpret parameters, functional algorithms, and modules of GAPIT implementation.

Key words Genomic selection, Mixed linear model, Population structure, Statistical power, Phenotype simulation

1 Introduction

GAPIT, stands for Genomic Association and Prediction Integrated Tool, is an R package to conduct both Genome-Wide Association Study (GWAS) and Genomic Prediction [1]. The first version was developed to implement the compressed mixed linear model (CMLM) to overcome both the p value inflation problem using General Linear Model (GLM) incorporating population structure

Davoud Torkamaneh and François Belzile (eds.), *Genome-Wide Association Studies*,
Methods in Molecular Biology, vol. 2481, https://doi.org/10.1007/978-1-0716-2237-7_13,

[2] and the over correction problem using the Mixed Linear Model (MLM) incorporating both population structure and kinship among individuals [3]. As GLM and MLM are the two special cases of CMLM, they were all implemented in GAPIT version 1 in 2012 [1]. With the availability of SUPER (Settlement of mixed linear models Under Progressively Exclusive Relationship) [4] and the Enriched CMLM models [5], GAPIT version 2 was released in 2016 for the implementations of these models [6]. After then, two distinct multiple loci model were developed with extraordinary computing speed and statistical power [7, 8], named FarmCPU (Fixed and random model Circulating Probability Unification) and BLINK (Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway). Besides Multiple Loci Mixed Model (MLMM) [9], the current GAPIT (version 3) implemented FarmCPU and BLINK to boost both computing speed and statistical power [10].

2 GAPIT Modules

GAPIT source code is available on GAPIT website (https://zzlab.net/GAPIT) and GitHub (www.github.com/jiabowang/GAPIT3). The whole GAPIT package includes the following five functional modules.

1. Data and Parameters (DP): In this module, based on the format and type of genotype data, phenotype data, and input parameters, GAPIT will determine what users want to do and prepare the necessary data and parameters. For example, based on the "model=MLM", the number of compressed groups will be set to the maximum number of individuals. The HapMap data will be converted to numeric data. The minor allele of homozygous genotype will be set 2, and the other genotype 0. The missing genotype will be imputed as 1 (heterozygote) for ease of utilization. Some logical adjustments will also be performed in this module. If parameters in such a method conflict with each other or there is a lack of necessary data, the GAPIT run will be stopped and a log with reminders will be issued. When genotype, heritability, and a number of Quantitative Trait Nucleotides (QTNs) are provided without phenotype data, GAPIT will conduct a phenotype simulation from the genotype data.
2. Quality Control (QC): The function of the QC module is to filter markers by MAF, sort, and match the taxa of genotype and phenotype files. The missing or NA values of individuals' traits in the phenotype file will be removed in the genotype file to keep common taxa in genotype and phenotype files.

3. Intermediate Components (IC): The kinship, Principal Component Analysis (PCA), phenotype distribution, MAF distribution, heterozygosity distribution, marker density, and Linkage Disequilibrium (LD) decay will be determined following the input parameters.
4. Sufficient Statistics (SS): The SS module applies an adapter for multiple GWAS methods. The adapter contains a data format converting function that will join with the input and command line of multiple GWAS methods, which include General Linear Model (GLM), Mixed Linear Model (MLM), Compressed MLM (CMLM), Factored Spectrally Transformed Linear Mixed Models (FaST-LMM), FaST-LMM-Select, and Settlement of mixed linear models Under Progressively Exclusive Relationship (SUPER), Multiple Locus Mixed Model (MLMM), Fixed and random model Circulating Probability Unification (FarmCPU), and Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK).
5. Interpretation and Diagnoses (ID): All static reports and relative results including a Manhattan plot, QQ plot, estimated heritability figure, and interactive outputs are created in this module. The ID module also can be used independently with some GWAS results from other software.

3 Quick Start

3.1 Import GAPIT Functions and Demo Data

There are two websites hosting the GAPIT source code, which can be accessed accordingly as follows.

Zhiwu Zhang Lab website:

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

GitHub:

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

The demo data include 281 maize lines and 3093 markers, which can be downloaded from these two websites. The phenotype file contains three traits: ear height (EarHT), days to pollination (dpoll), and ear diameter (EarDia).

3.2 Copy-Paste Command-Line Interface

GAPIT provides demo code for multiple scenarios. Users can select one of them with copy-paste into R environment to run GAPIT. Here is an example using five methods with demo data to do GWAS.

```
#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_nu-
meric.txt",head=T)
myGM=read.table(file="http://zzlab.net/GAPIT/data/
mdp_SNP_information.txt",head=T)

#GWAS with five methods
myGAPIT <- GAPIT(
Y=myY[,c(1,2)],  #The  phenotype  file,  fist  column  is
individual  ID.
GD=myGD,  #  the  numeric  genotype  file,  first  column  is
individual  ID
GM=myGM, # the genotype map file including 3 columns. The
third  column  contains  a  unique  position  in  each  chromo-
some.
PCA.total=3,  #  the  number  of  PCs  will  be  put  into  the
model.
model=c("GLM", "MLM", "MLMM", "FarmCPU", "Blink"), # here
can use only one method or multiple methods
Multiple_analysis=TRUE)
```

3.3 Output Files

The success of the analysis will place multiple files in the R working directory, including Manhattan plot (chromosome-wise), Manhattan plot (genome-wise), QQ plot, GWAS result table, and estimated effect table for each method. There are also several genotype and phenotype analysis results including marker LD, phenotype view, marker density, PCA, heterozygosity, and kinship. Some output files ending with .html extension are interactive Manhattan and QQ plots.

4 Input Data and Format

4.1 Genotypes in HapMap Format

HapMap stores markers as rows and individual as columns. The first column is the marker name. The second column is the allele type (variance/reference). The third and fourth columns are the chromosome number and the physical position. The 5th–11th columns are attributes of the SNP sequence and the remaining columns show the observed genotype at each SNP for each individual. The first row contains all header labels (1st to 11th columns) and taxa of individuals (remaining columns). When reading this HapMap file, “head=FALSE” should be included in the reading code. Missing values can be accepted in the HapMap file and should be coded “NN” (double bit) or “N” (single bit). The real genotype in HapMap format can be coded as either double bit or as the standard IUPAC code.

4.2 Genotypes in Numeric Format and Map

The major information in the HapMap file is divided into two files; the SNP map information is in the GM file, and the SNP genotype information is in the GD file. The GD file is a numeric format that has been used by EMMA [11]. SNP genotypes are coded in a column and individuals are coded in a row. The GM file contains the SNP name, chromosome number, and physical position. The column order of the GD file should be matched to the order of the GM file. Homozygotes are denoted by “0” (no MAF marker) or “2” (MAF marker), and the heterozygotes are denoted by “1” in the GD file. The first column of the GD file contains the individuals’ names. The first column’s name is “taxa” and the remaining columns’ names are the names of SNPs. The taxa of individuals and SNPs should not be NA nor duplicated. The first column of the GM file is the taxa of SNPs. The second column is the chromosome number and the third is the physical position on the chromosome.

4.3 Phenotype

GAPIT accepts single or multiple trait files. This is achieved by including all phenotypes in the text file of phenotypic data. 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. We suggest that the taxa of individuals should be coded as alphanumeric characters plus symbols. Missing data should be indicated by either “NaN” or “NA”. The phenotype file is tab delimited.

4.4 Covariate Variables and PCA

Covariates (CV) can be put into the model as fixed effects. They can include population structure (Q matrix or principal components), other related traits, and environmental variables. The first column of the CV file consists of individual 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. Importantly, the individuals’ taxa included in the GD file must also be in the CV file. Otherwise, GAPIT will remove the individuals not listed in the CV file. GAPIT also applies an option to perform PCA and incorporate the top 3 PCs into the model using PCA.total=3.

4.5 Kinship

The kinship (called “KI” in GAPIT) is formatted as an n by n+1 matrix where the first column contains the individuals’ taxa name, and the remaining columns contain a square kinship matrix value.

Unlike the other input data files, the first row of the kinship matrix file does not consist of headers. The kinship can be either provided in an input file using “KI=myKI” or selected a method for calculation by GAPIT using “kinship.algorithm="Zhang"”. The options for calculation methods include “Zhang” [12], “VanRaden” [13], “Loiselle” [14], and “emma” [11]. The kinship can be used as a random effect in the MLM, CMLM, SUPER, and MLMM methods.

5 Genome-Wide Association Study

As a GWAS and GS integrated tool, GAPIT implements multiple statistical models (Fig. 1), including the ones for GWAS only (e.g., GLM and BLINK) and for both GWAS and GS (e.g., MLM, CMLM, and SUPER). For the methods shared by GWAS and GS, the descriptions here refer to GWAS.

5.1 GLM

The fastest GWAS linear method, the GLM method can be performed simply by setting model=“GLM” in GAPIT. The population structure (Q matrix or PCs) is only one of the fixed effects in the total model. Therefore, the GLM is very sensitive to the population structure. The SNP effect is estimated by testing each marker in the model in a stepwise fashion. This is the method employed in PLINK [15], which is commonly used in human genetics studies. The GLM result in GAPIT should be the same as in PLINK.

5.2 MLM

A mixed linear model can be performed by setting model=“MLM”, which treats each individual as a cluster group in the random effect (kinship matrix, K). Combined with the fixed effect, MLM improves the statistical power over GLM with the “Q+K” approach rather than only the “Q” approach. GAPIT uses EMMAX/P3D to reduce computing time in MLM, CMLM, and SUPER [12]. The population and residual variance components will be estimated using the EMMA algorithm only once for each SNP effect estimated. Currently, MLM is the most popularly used or compared method for GWAS.

5.3 CMLM

Since kinship is derived from all the markers, incorporating kinship for testing markers in an MLM causes confounding between the tested 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 (CMLM) developed by Zhang et al. in 2010 [12]. 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. CMLM can be performed by setting model=“CMLM”. By default, the compressed groups will

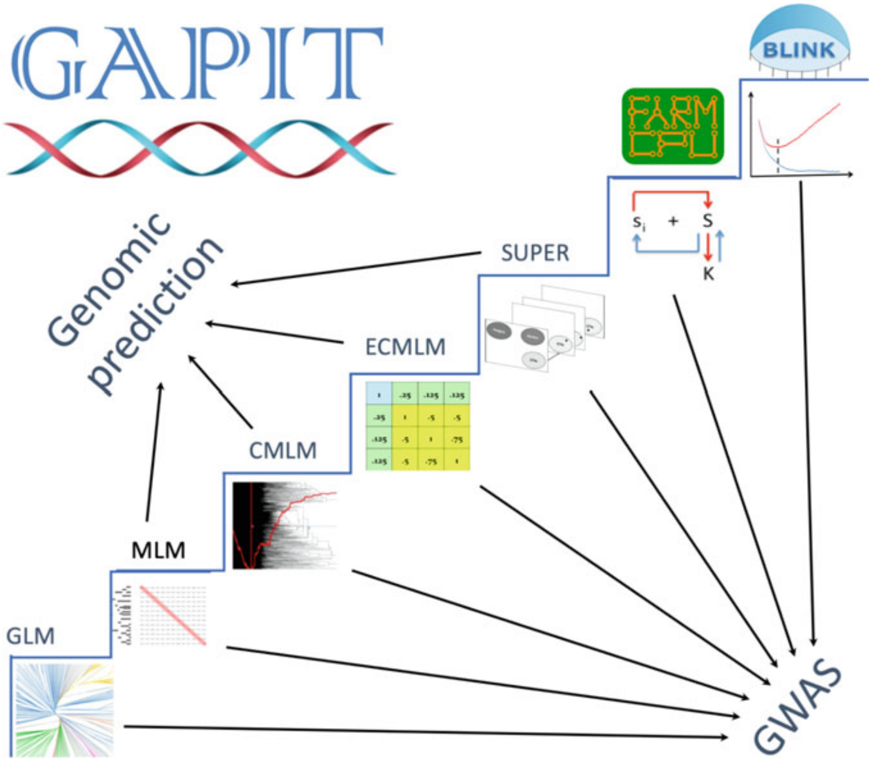

Fig. 1 Statistical Methods for Genome-Wide Association Studies and Genomic Selection in GAPIT. The GAPIT version 1 included GLM, MLM, and CMLM methods for GWAS, as well as gBLUP with MLM for GS. The GAPIT version 2 included ECMLM and SUPER for GWAS, and cBLUP (CMLM), and sBLUP (SUPER) for GS. Now GAPIT version 3 also includes FarmCPU and BLINK for GWAS

be from 1 to the number of all individuals, and the compressed step length will be 20. The compressed groups can be modified using group.from and group.to parameters; the compressed step length can be modified using group.by. The likelihood values will be used to estimate the best optimum compressed group, which often results in higher statistical power than GLM or MLM. The GLM also can be named a special CMLM, in which the number of compressed groups is set at 1. That means all individuals' relationship values are compressed as a cluster group so that the random effect in the mixed linear model is reduced to 0. Similarly, the MLM also can be named a special CMLM, in which the number of compressed groups is set at the maximum number of individuals.

5.4 ECMLM

Based on the CMLM, the clustered method can be divided into eight algorithms, including "ward.D", "ward.D2", "single", "complete", "average", "mcquitty", "median", and "centroid". The calculated kinship value in each clustered group can also be divided into four algorithms, including "mean", "max", "min", and "median". The different combinations (24) between clustered and grouped algorithms have been shown to be more powerful for a specific population and trait by Li et al. [5]. Now, “kinship. cluster” can be used to select clustered algorithms and “kinship. group” can be used to select grouped algorithms in CMLM of GAPIT.

5.5 SUPER

An advanced version of FaST-Select, SUPER was developed by Wang et al. [4]. SUPER uses the bin approach to select pseudo QTNs. The entire genome is divided into equal-sized bins and the most significant marker of the bin is selected to represent the bin type. The bin size and number of bins selected are optimized using the maximum likelihood method in a random model with the kinship derived from the selected bins. When each marker is tested in a new MLM, the pseudo QTNs are excepted from the kinship. Use model=“SUPER” to easily run SUPER in GAPIT. The bin. from, bin.to, and bin.by are used to optimize bin size and default values are all 10,000. The inclosure.from, inclosure.to, and inclosure.by are used to optimize the number of bins selected, and default values are all 10.

5.6 MLMM

Different from the above single-locus GWAS method, the multilocus mixed model (MLMM) uses a forward stepwise regression model to detect pseudo QTNs as covariates and a backward stepwise regression model to estimate the markers’ effect. In each step, the variances are estimated by generalized least-square (GLS), and the P-values are estimated by F-test. The significantly associated SNPs are added into the model as cofactors during the next step, and the P values of all new cofactors are reestimated together. In the MLMM, the extended Bayesian information criteria (BIC) is used to define the BIC penalty to test the model convergence. To run MLMM in GAPIT, simply specify model=“MLMM”.

5.7 FarmCPU

To solve the problem of false-positive control and the confounding between markers and covariates, an iterative and multilocus method called Fixed and random model Circulating Probability Unification (FarmCPU), was developed in 2016 [8]. The associated markers detected from the iterations are fitted as the cofactors to control for false positives when testing the remaining markers in a fixed effect model. To avoid the model overfitting problem in stepwise regression, a random effect model is used to select the associated markers using a maximum likelihood method. Use model=“FarmCPU” to directly run FarmCPU in GAPIT.

5.8 BLINK

BLINK is a new GWAS model, which was designed to have both high statistical power and computational efficiency. In BLINK, Bayesian information criteria (BIC) is used to select the optimum constriction of loops in the fixed effect model, replacing REML in the random effect model. This eliminates the assumption required by FarmCPU that causal genes are evenly distributed across the genome. The assumption can cause either inclusion of no causal genes or failure to identify causal genes that are in the same bin with another causal gene that has a stronger signal. The linkage disequilibrium information is used to replace the bin method for simplification of genome information. To run the BLINK R version in GAPIT, simply type model="Blink". To run the BLINK C version in GAPIT, first download the BLINK C executable file, then specify model="BlinkC".

6 Simulation and Assessment of Statistical Power

6.1 Simulating Complex Traits

The phenotype simulation is used to compare models and experimental design. Based on the heritability, number of QTNs, and the distribution of their additive genetic effects, GAPIT applies an approach to simulate a continued phenotype for each individual with a genotype file (either numeric or HapMap file). The parameter "h2=0.7" defines heritability. The parameter "NQTN=20" can be used to set the number of QTNs as 20. In the simulation situation, GAPIT will not accept the input phenotype file. Then GAPIT will give a list output, in which the QTN.position indicates the simulated QTN position ordered in the GM file, and Y indicates simulated phenotype value. The simulated phenotype order is the same as the individual order in the GD file.

```
mysimulation<-GAPIT(h2=0.7, NQTN=20, GD=myGD, GM=myGM)
posi=mysimulation$QTN.position
myY=mysimulation$Y
```

6.2 Simulating Ordinary Traits

By default, GAPIT samples the QTN effect from a standard normal distribution with mean 0 and variance 1 for complex trait simulations. Some simple traits are controlled by a few genes with large effect, so GAPIT applies the second approach for ordinary traits simulated using "QTNDist=geometry" and "effectunit=0.9". That means the first QTN contributes 0.9 effect, and the next QTN contributes 0.9^n effect. The n is the order of QTNs. For an ordinary trait, GAPIT applies multiple parameters to the simulation. The category is set, and the number indicates the group number of simulated category traits. When the category is set as 0, GAPIT will simulate a binary trait with the ratio (r).

The code below shows how to simulate an ordinary trait:

```
mygeometry<-GAPIT(h2=0.7, NQTN=20, GD=myGD, GM=myGM, QTNDist=”geometry”,
effectunit=0.9)
mycategory<-GAPIT(h2=0.7, NQTN=20, GD=myGD, GM=myGM, category=6)
mybinary<-GAPIT(h2=0.7, NQTN=20, GD=myGD, GM=myGM, category=0, r=0.8)
#Object: Simulated phenotype for multiple distribution of QTN effect with the demo data
  #Input: GD - The numeric SNP file (content format: 0,1 and 2)
  #Input: GM - SNP information file (content: SNP ID, chromosome, position)
  #input: QTNDist – assume distribution of QTN effect
  #input: h2 - heritability
  #input: effectunit – the first QTN effect in the geometry distribution
  #input: NQTN - number of QTNs
  #input: category – the number of category trait
  #input: r – the ratio of binary trait
```

6.3 Assessment of Statistical Power

GAPIT provides an integrated function for comparison of multiple GWAS models, named GAPIT.Power.Compare. In this function, GAPIT will simulate a trait based on the genetics parameters defined by the users, then calculate power, FDR, and Type I error. The average values of power, FDR, and Type I error in all replications are used for the comparison (Fig. 2).

```
GAPIT.Power.Compare(
GD=myGD,
GM=myGM,
rep=100,
h2=0.7,
Method=c("GLM","MLM",”FarmCPU”),
NQTN=5
)
 #Object: Compare to Power-FDR with different model method with the demo data
  #Input: GD - The numeric SNP file (content format: 0, 1 and 2)
  #Input: GM - SNP information file (content: SNP ID, chromosome, position)
  #input: rep - repeat 100 times
  #input: h2 - heritability
  #input: Method - Choosing multiple GWAS methods to compare
  #input: NQTN - number of QTNs
```

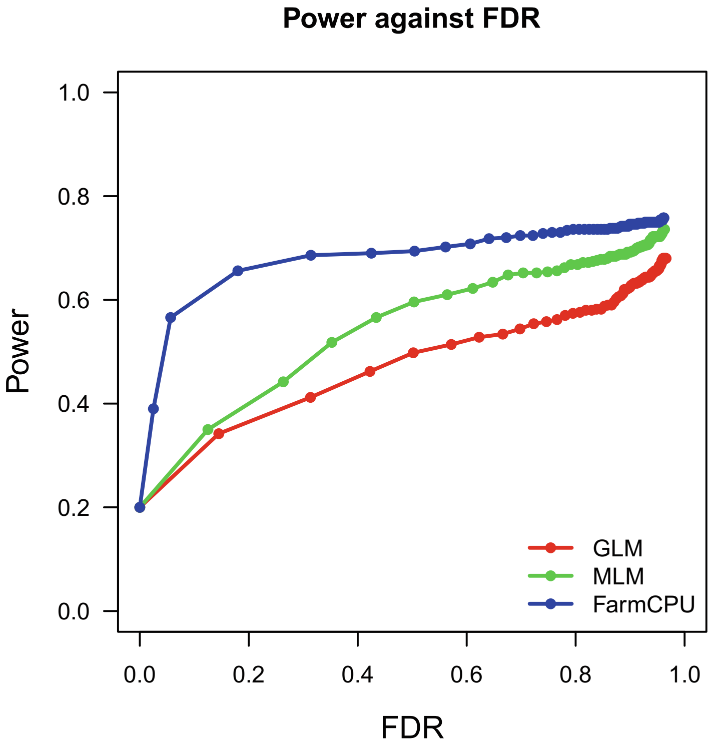

Fig. 2 The power vs. FDR for GLM, MLM, and FarmCPU in a trait simulation. The replication is 100, and the number of QTN is set at 5

7 Interpretation of Results

7.1 Manhattan Plot

The Manhattan plot is a scatter plot that summarizes GWAS results. The *X*-axis indicates the genomic physical position of each SNP, and the *Y*-axis displays the negative logarithm of the *P*-value obtained from the GWAS model (specifically from the *F*-test for testing H0: No association between the SNP and trait). Each chromosome is colored differently. Large peaks in the Manhattan plot (i.e., “skyscrapers”) suggest that the surrounding genomic region has a strong association with the trait. The green dashed line shows the FDR threshold cutoff ($0.05 \times o/m$, o is the order of difference value between cutoff and (all *P* value)/m (m is the number of total SNPs) [16], and the green solid line shows the Bonferroni threshold cutoff ($0.05/m$) [17]. In some simulation studies, the solid black points indicate simulated real QTN. GAPIT produces one Manhattan plot for the entire genome (Fig. 3) and individual Manhattan plots for each chromosome (Fig. 4). On the chromosome-wise Manhattan plot, the most significant marker genotype is used to calculate correlation with its neighboring markers' genotypes. The correlation levels are shown as a heatmap.

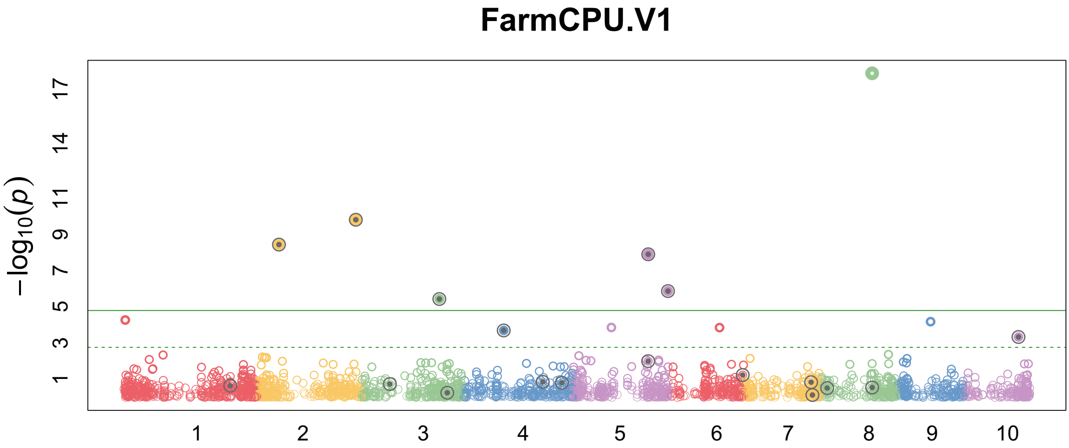

Fig. 3 Genome-wise Manhattan plot of FarmCPU with a simulated trait in GAPIT. The heritability is 0.5, and the number of QTN is 20. "V1" is the simulated trait's taxa. The data used in the simulation is 281 maize lines and 3093 markers

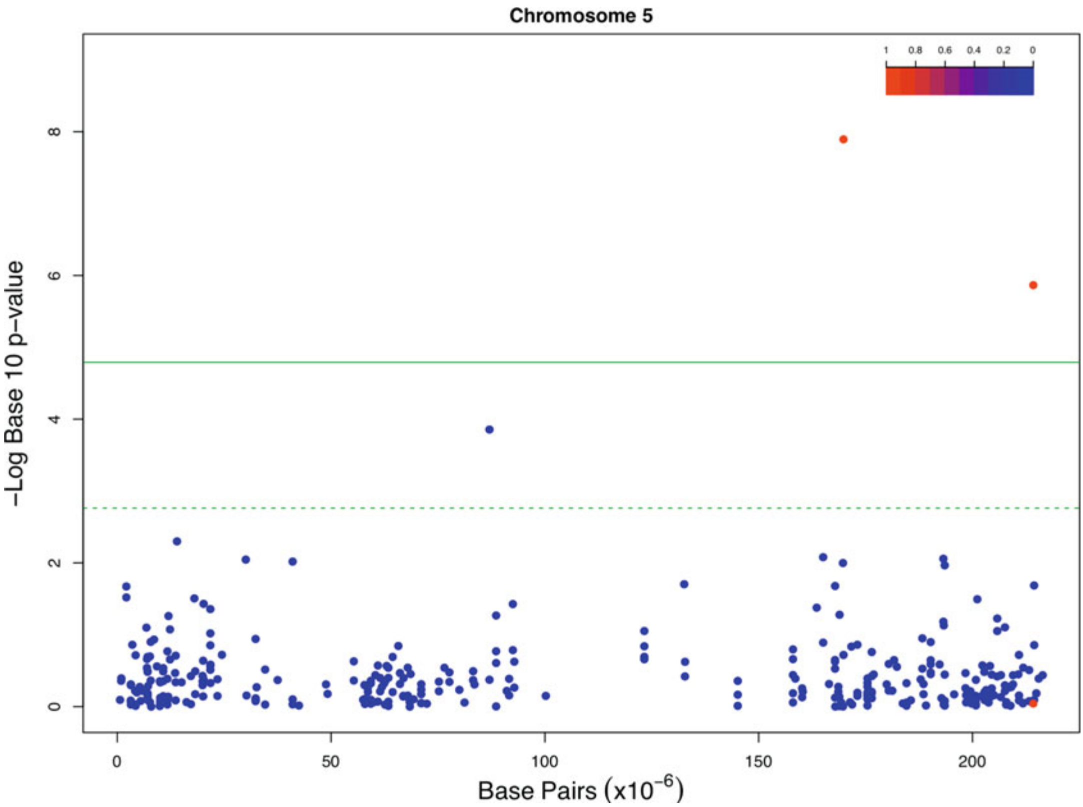

Fig. 4 Chromosome-wise Manhattan plot of FarmCPU with simulated trait on chromosome 5. The heritability is 0.5, and the number of QTN is 20

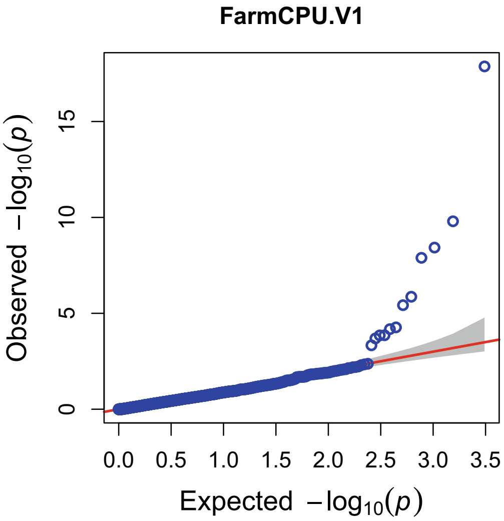

Fig. 5 QQ plot of FarmCPU with a simulated trait in GAPIT. The heritability is 0.5, and the number of QTN is 20

7.2 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 [18]. In this plot, the negative logarithms of the *P*-values from the models fitted in GWAS are plotted against their expected values under the null hypothesis of no association with the trait (Fig. 5). 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.

7.3 Interactive Plots

GAPIT provides interactive Manhattan and QQ plots (Fig. 6) with additional important information, such as minor allele frequency (MAF), estimated effect, neighboring gene names, and the ratio of markers explaining phenotypic variance. Users can use a mouse to select a subset, such as a chromosome, zoom in and out on the whole Manhattan plot, or filter markers based on the temporary cutoff. The R package "plotly" applies an approach from R to HTML [19]. Each marker in the Manhattan plot is linked to a pop-up containing detailed information. The plot can be displayed in web browsers with a folder named "library". The parameter "Inter.Plot=TRUE" is used to create interactive GWAS plots.

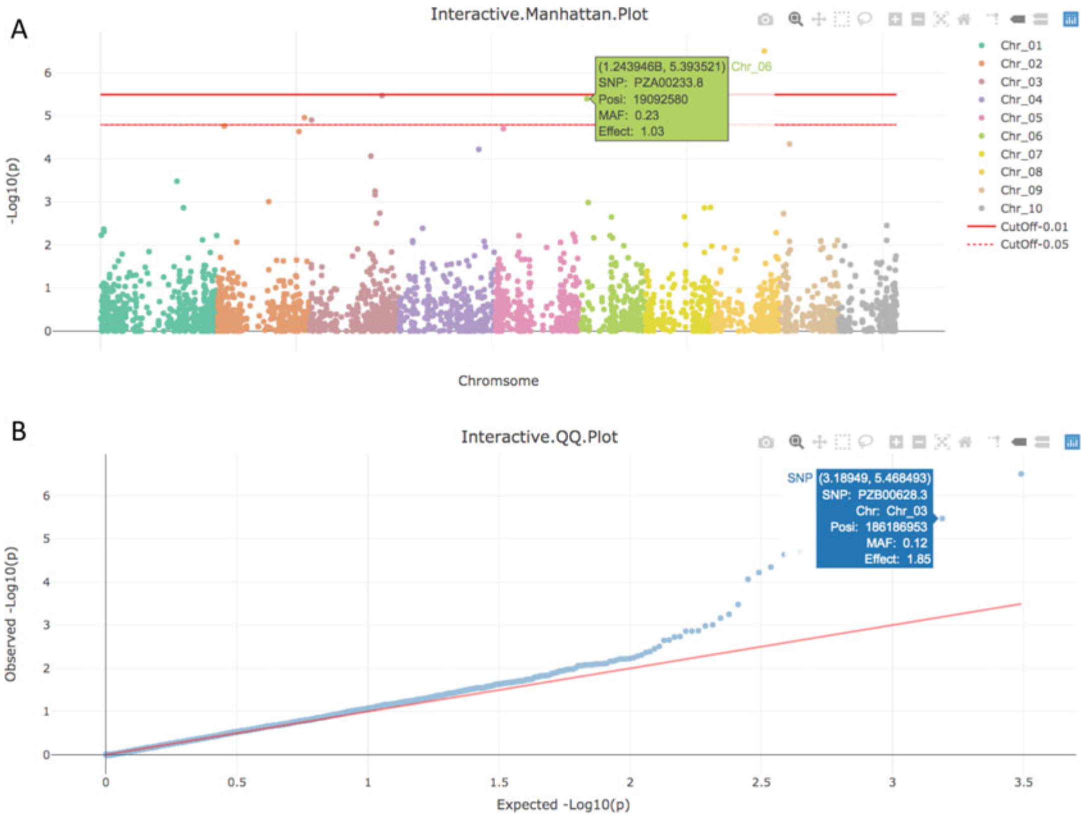

Fig. 6 Interactive Manhattan (A) and QQ plot (B) using "plotly" R package as HTML file

7.4 Table Reports

After GWAS analysis, GAPIT will output several Excel tables for interpretation of all calculations and GWAS results. All these tables are CSV format files, including GWAS result, t value table, LOG record, Prediction table, PCA value, PCA eigenvalue, and Kinship table. The dimension of the GWAS result table is m by 10; *m* is the number of SNPs. The first 3 columns are SNPID, Chromosome number, and Position of SNP. The remaining columns are *P*-value, MAF, DF, *R* square with SNP, *R* square without SNP, FDR-adjusted P-value, and estimated effect. The *t* value table contains SNP map information and t value with standard error.

7.5 Genotype Analysis

GAPIT provides multiple genotype and genome analyses, including marker density and distribution, linkage disequilibrium (LD) decay, and heterozygosity (Fig. 7). These analyses will help users to understand the genetic background from genotype and evaluate the results. By default, the top 100 markers are used to draw the density, distribution, and LD decay plots. Marker density is critical to establish LD between markers and causal mutations. Comparison between the marker density and the LD decay over distance

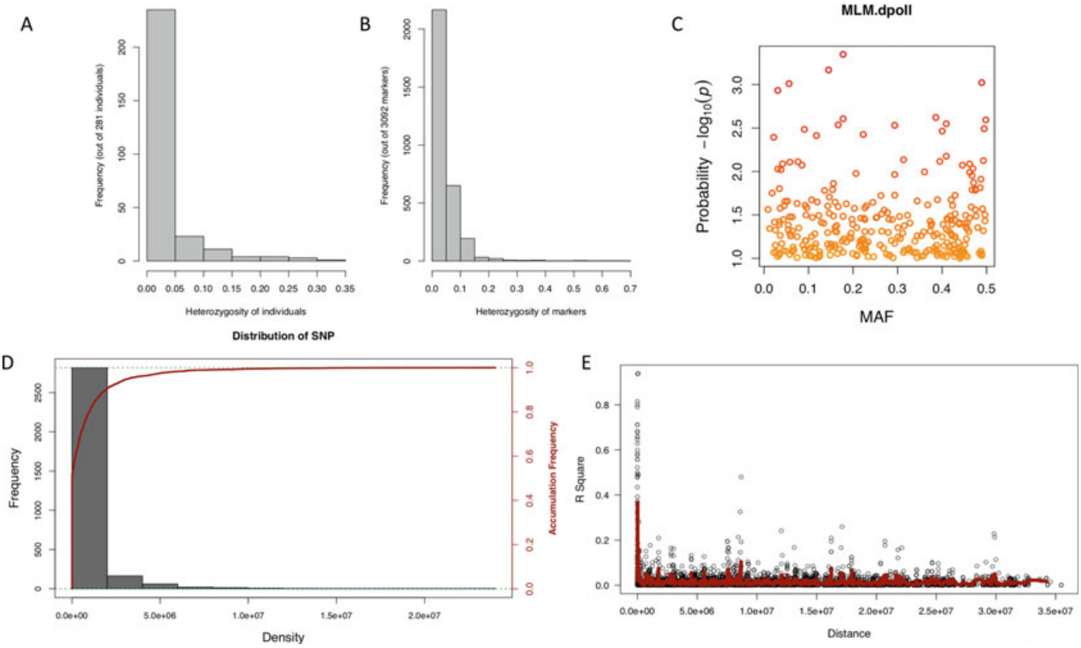

Fig. 7 Genotype analysis including Heterozygosity (**a** and **b**), MAF (**c**), SNP density (**d**), and SNP LD (**e**)

indicates if markers are dense enough to have good coverage of LD. Linkage disequilibrium is measured as R square for pairwise markers and plotted as the distance between them. The moving average of adjacent markers is calculated by using sliding windows of ten markers. All markers are used to draw a heterozygosis plot. The frequency of heterozygotes is calculated for both individuals and markers. A high level of heterozygosity indicates low quality. For example, over 50% heterozygosity for some of the markers in inbred lines is problematic. All genotype analyses can be skipped with "Geno.View.output=FALSE" when the user wants to simply run the GWAS program.

7.6 Population Structure and Kinship

There are two types of PCA that can be used to explain the population structure graph in GAPIT. The first type contains pairwise plots of PCs (Fig. 8a), and the second type gives an interactive three-dimensional plot (Fig. 8b). This interactive plot is created as an HTML file using the "rgl" R package. Tt can be zoomed in or out, displayed by subpopulation, and rotated to view from any angle. The PCA values are in the PCA.csv file, combined with individuals' taxa as the first column. The color of PCA points can be set using "PC.col=color" to set. The kinship also can be used to create two graphs; one is the heat map among individuals, and the other is the clustering neighbor-joining (NJ) tree (Fig. 9). In GAPIT, either the parameter "NJtree.group" can be used to set the subpopulation in the PCA and NJ-tree plot, or the clustered

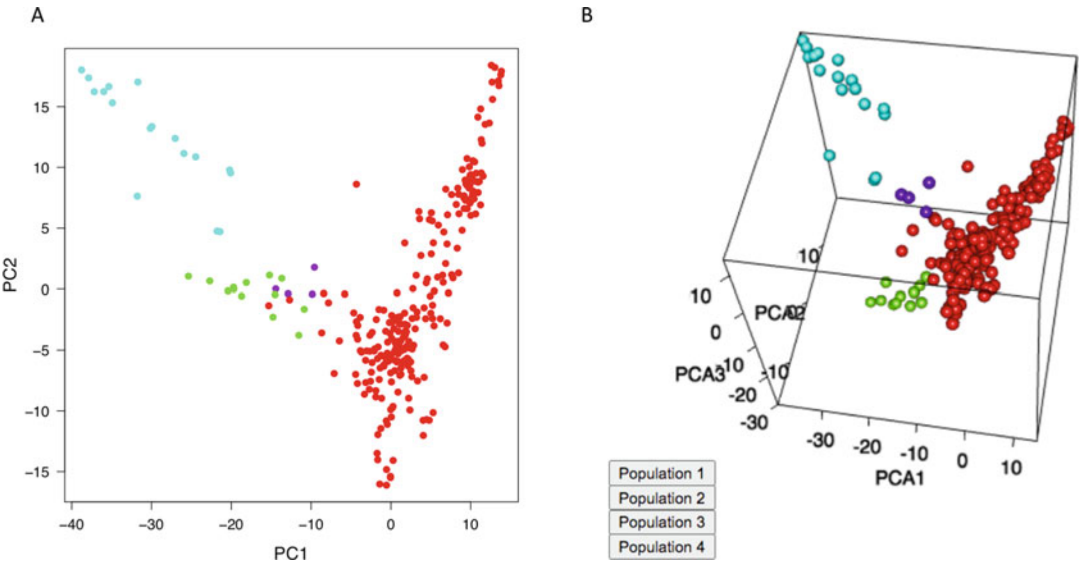

Fig. 8 Population structure with 2D (**a**) and 3D (**b**) PCA plot. 281 individuals are clustered into four groups based on compressed kinship

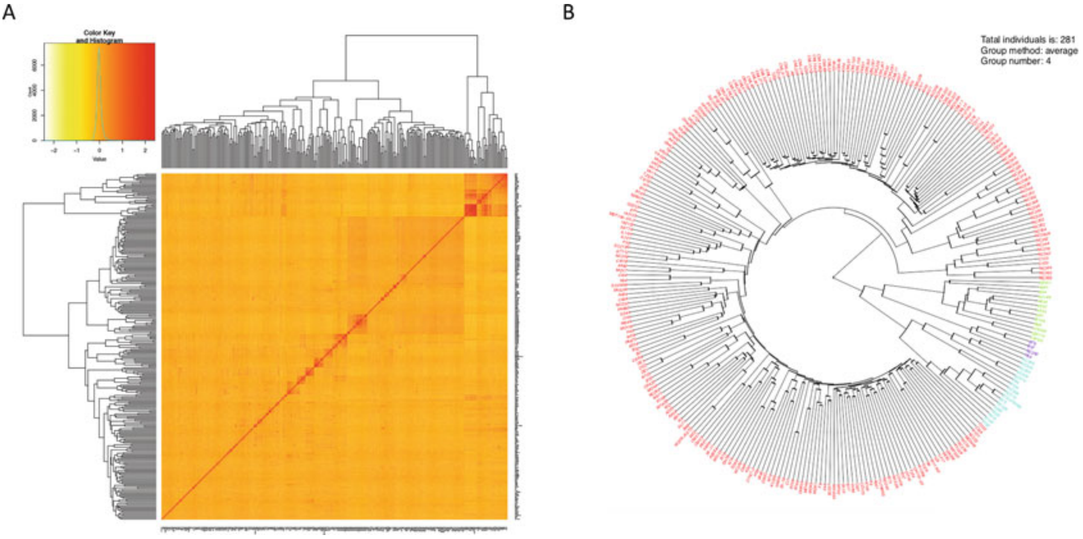

Fig. 9 The cluster heatmap (**a**) and NJ-tree plot (**b**) used kinship to present the relationships between individuals

kinship can be used to select the subpopulation by default. In the CMLM and MLM, the kinship was estimated and presented with pie plot (Fig. 10). The plot shows the ratio of genetic variance to total phenotypic variance.

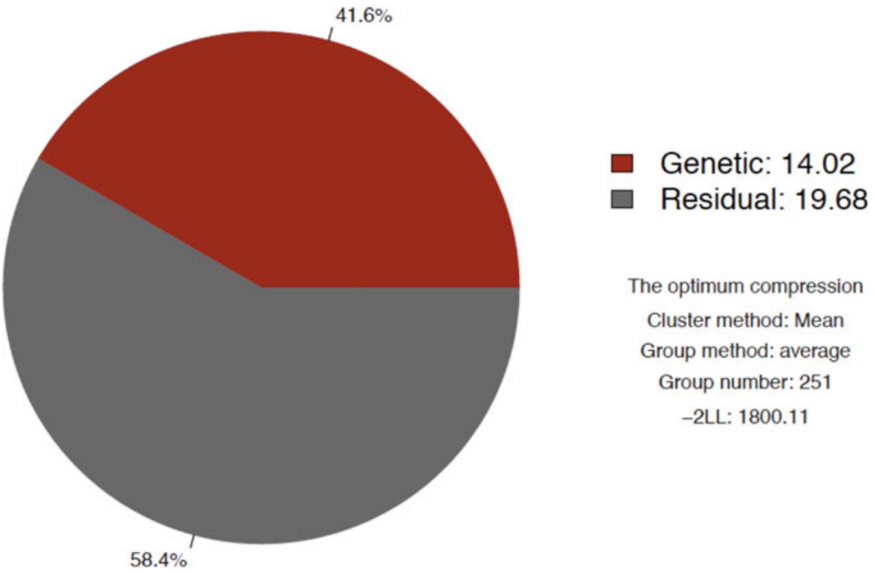

Fig. 10 Estimated heritability of optimum compression. The optimal method to calculate group kinship is "Max", the optimal clustering method is "average", the number of groups (i.e., the dimension of the group kinship matrix) is 251, the value of −2*log likelihood function is 1800.11, and the heritability is 41.6%

The PCA and NJ tree can be created with the following codes.

```
myGAPIT=GAPIT(
GD=myGD,
GM=myGM,
PCA.total=3,
PCA.3d=TRUE,
NJtree.group=4,
NJtree.type=c("fan"),
)
#Object: Create interactive PCA and NJ tree plot with the demo data
  #Input: GD - The numeric SNP file (content format: 0,1 and 2)
  #Input: GM - SNP information file (content: SNP ID, chromosome, position)
  #input: rep - repeat 100 times
  #input: h2 - heritability
  #input: Method - Choosing multiple GWAS methods to compare
  #input: NQTN - number of QTNs
```

8 Further Help

8.1 GAPIT User Manual

To guide how to use GAPIT, we created a user manual at zzlab.net/GAPIT. Now GAPIT has been developed to version 3 with multiple GWAS and Genomic Prediction methods. The previous methods and cited papers are listed in the user manual to help user understand the principles and choose the appropriate method. Following additional updates to GAPIT functions, the user manual will be continually revised and updated. The log of each updated function is registered in the GAPIT Biography section. Some representative questions from GAPIT users are answered in the Frequently Asked Questions section.

8.2 User Communication

GAPIT has received over a thousand citations for version 1 and 2 [1, 6]. The GAPIT website (https://zzlab.net/GAPIT) has received over 20,000 page views since 2016. The GAPIT forum (https://groups.google.com/g/gapit-forum) has over 600 members with over 800 conversations. The forum was viewed over 3000 times by the GAPIT community between 2016 and 2019.

8.3 GAPIT Team

The first version of GAPIT was published in Bioinformatics by Dr. Alex Lipka et al. in 2012 [1]. The second version of GAPIT was published in The Plant Genome by Dr. You Tang et al. in 2016 [6]. The third version is in the publication process by Genomics, Proteomics, and Bioinformatics [10]. Currently, Dr. Jiabo Wang is maintaining the software and leading the development of GAPIT version 4. A full usage of the forum is encouraged to ask questions, finding answers, and making comments for benefit of GAPIT user community before contacting the current leading author (Dr. Jiabo Wang, e-mail: wangjiaboyifeng@163.com).

Acknowledgments

This project was partially funded by the National Science Foundation of the United States (Award # DBI 1661348 and ISO 2029933), the United States Department of Agriculture—National Institute of Food and Agriculture (Hatch project 1014919, Award #s 2018-70005-28792, and 2019-67013-29171, and 2020-67021-32460), the Washington Grain Commission, the United States (Endowment and Award #s 126593 and 134574), Sichuan Science and Technology Program, China (Award #s 2021YJ0269 and 2021YJ0266), the Program of Chinese National Beef Cattle and Yak Industrial Technology System, China (Award # CARS-37), and Fundamental Research Funds for the Central Universities, China (Southwest Minzu University, Award # 2020NQN26).

Author Contributions Jiabo Wang: programming, data curation, writing the manuscript draft, and software testing.

You Tang: software testing.

Zhiwu Zhang: conceptualization and writing.

Availability The GAPIT source code, demo script, and demo data are freely available on the GAPIT website (zzlab.net/GAPIT) and Github (www.github.com/jiabowang/GAPIT3).

References

1. Lipka AE, Tian F, Wang Q et al (2012) GAPIT: genome association and prediction integrated tool. Bioinformatics 28:2397–2399. https://doi.org/10.1093/bioinformatics/bts444
2. Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. Am J Hum Genet 69:1–14. https://doi.org/10.1086/321275
3. Yu J, Pressoir G, Briggs WH et al (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208. https://doi.org/10.1038/ng1702
4. Wang Q, Tian F, Pan Y et al (2014) A SUPER powerful method for genome wide association study. PLoS One 9:e107684. https://doi.org/10.1371/journal.pone.0107684
5. Li M, Liu X, Bradbury P et al (2014) Enrichment of statistical power for genome-wide association studies. BMC Biol 12:73. https://doi.org/10.1186/s12915-014-0073-5
6. Tang Y, Liu X, Wang J et al (2016) GAPIT version 2: an enhanced integrated tool for genomic association and prediction. Plant Genome 9
7. Huang M, Liu X, Zhou Y, Summers RM, Zhang Z (2019) BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. Gigascience 8(2):399–404
8. Liu X, Huang M, Fan B et al (2016) Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLoS Genet 12:e1005767. https://doi.org/10.1371/journal.pgen.1005767
9. Korte A, Vilhjálmsson BJ, Segura V et al (2012) A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet 44:1066–1071. https://doi.org/10.1038/ng.2376
10. Wang J, Zhang Z (2020). GAPIT version 3: boosting power and accuracy for genomic association and prediction. BioRxiv 2020.11.29.403170. https://doi.org/10.1101/2020.11.29.403170
11. Kang HM, Zaitlen NA, Wade CM et al (2008) Efficient control of population structure in model organism association mapping. Genetics 178:1709–1723. https://doi.org/10.1534/genetics.107.080101
12. Zhang Z, Ersoz E, Lai CQ et al (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42:355–360. https://doi.org/10.1038/ng.546
13. VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423. https://doi.org/10.3168/jds.2007-0980
14. Loiselle BA, Sork VL, Nason J et al (1995) Spatial genetic structure of a tropical understory shrub, Psychotria officinalis (Rubiaceae). Am J Bot 82:1420–1425. https://doi.org/10.2307/2445869
15. Purcell S, Neale B, Todd-Brown K et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575. https://doi.org/10.1086/519795
16. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 57:289–300. https://doi.org/10.2307/2346101
17. Bonferroni CE (1936) Teoria statistica delle classi e calcolo delle probabilità. Pubbl Del R Ist Super Di Sci Econ e Commer Di Firenze. Seeber
18. Pleil JD (2016) QQ-plots for assessing distributions of biomarker measurements and generating defensible summary statistics. J Breath Res 10(3):035001. https://doi.org/10.1088/1752-7155/10/3/035001
19. Carson S, Chris P, Toby H, et al. plotly: Create Interactive Web Graphics via "plotly. js." R Packag Version 2016