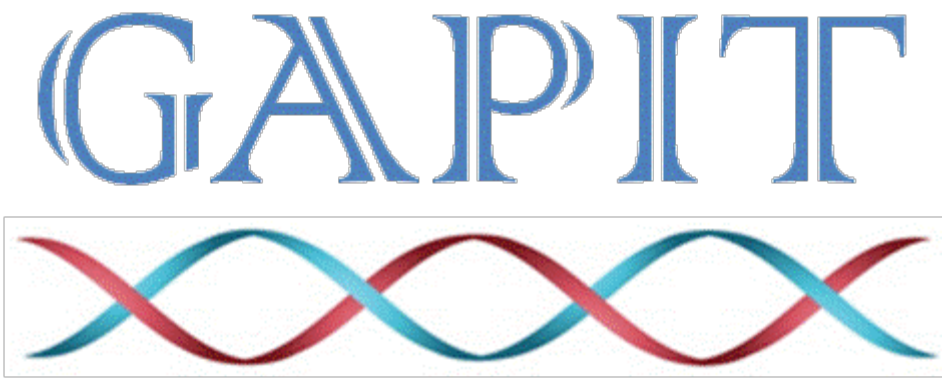
User Manual for



Genomic Association and Prediction Integrated Tool

by

Alexander E. Lipka, Feng Tian, Qishan Wang, Jason Peiffer, Meng Li, Peter J Bradbury, Michael Gore, Edward S Buckler and Zhiwu Zhang

Institute for Genomic Diversity
Cornell University

(Last updated: January 30, 2013)

Please send questions, comments, and suggestions to Alex Lipka (ael54@cornell.edu)

http://www.maizegenetics.net/GAPIT

Disclaimer: While extensive testing has been performed by the Buckler Lab at Cornell University and results are, in general, reliable, correct or appropriate, 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, TASSEL and EMMA.

Further help: Users are welcome to report bugs and request new features. Questions are also welcome. Please email your questions, comments and suggestion to the authors.

The GAPIT project is supported by the National Science Foundation and the USDA-ARS.

Contents

1 INTRODUCTION

1.1 Why GAPIT?

The gap between developing statistical methods and applying them to solve bioinformatics problems is increasing because of recent advances in genotyping technologies. It is now more affordable than ever to obtain millions of SNPs on a large number of individuals. Although software programs have been developed to implement these statistical methods, many of them cannot accommodate large data sets. Additionally, many of these computer programs require a complex user interface, and thus may not be easily accessible to researchers.

The Efficient Mixed Model Association (EMMA) R package (Kang *et al*, 2008) and Trait Analysis by Association, Evolution, and Linkage (TASSEL; Bradbury *et al*, 2010) are two relatively easy to use programs. The EMMA R package uses the EMMA algorithm to reduce computational time when estimating variance components in the mixed model. TASSEL was written in Java and can be operated through a simple graphic user interface. In addition to the EMMA algorithm, TASSEL uses the compressed mixed linear model (CMLM; Zhang *et al*, 2010) and population parameters previously determined (P3D; Zhang *et al*, 2010) to speed up computational time and optimize statistical performance. The P3D method was independently developed by Kang *et al* (2010) and is implemented in a C program called EMMA expedited (EMMAx).

Identifying the genetic potential of individuals is one of the ultimate goals of genetic researchers. In plants and animals, this genetic potential can be incorporated into selection programs. In humans, this can be used for genomic prediction of an individual's likelihood for having a disease, and medical treatments can be applied accordingly. A recent study (Zhou *et al*., unpublished) showed that the CMLM method developed for GWAS can be used for genomic prediction and selection, and can provide accurate predictions with a dramatic reduction in computational time.

Genomic Association and Prediction Integrated Tool (GAPIT) is a computer package that uses EMMA, CMLM, and P3D to conduct GWAS and make genomic predictions. This package is operated in an R environment and uses a minimal amount of code. Large data sets with many SNPs can be analyzed in GAPIT by subdividing the genotypic data into multiple files. GAPIT can read in genotypic data in either HapMap format or in the numerical format required for the EMMA R package. GAPIT reports detailed results in a series of tables and graphs.

This help document and corresponding tutorials should assist with getting a new user familiar with GAPIT. Additional questions may be directed towards the authors.

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/. GAPIT package can be installed by typing this command line:

```
source("http://www.maizegenetics.net/images/stories/bioinformatics/GAPIT/gapit_functions.txt")
```

GAPIT uses five R libraries: multtest, gplots, LDheatmap, genetics, and a modified version of the EMMA R package. Library multtest can be installed by typing these command lines:

```
source("http://www.bioconductor.org/biocLite.R")
biocLite("multtest")
```

Library gplots, LDheatmap and genetics can be installed by typing these commend lines (you are required to choose a cite to download):

```
install.packages("gplots")
install.packages("LDheatmap")
install.packages("genetics")
```

Once the above packages are installed, these libraries can be imported to R environment by typing these commands:

```
library(multtest)
library("gplots")
library("LDheatmap")
library("genetics")
```

The EMMA library was developed by Kang *et al*. (2007). One line was added to the library to handle the grid search with "NA" likelihood. The modified library can be installed by typing this command line:

```
source("http://www.maizegenetics.net/images/stories/bioinformatics/GAPIT/emma.txt")
```

Now create a directory under C drive and set it as your working directory in R. Download the GAPIT tutorial dataset to the directory:

```
setwd("C:\\myGAPIT")
```

The easiest way of using GAPIT is to COPY/PASTE GAPIT tutorial code. Here is the code from the first tutorial. After a few minute, the GWAS and genomic prediction results will be saved in the above working directory.

```
#Step 1: Set working directory and import data
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
)
```

These results are also included in the tutorial data set. All tables, including GWAS *P*-values and genomic breeding values are stored in comma separated value (.csv) files. All graphs, including Manhattan plots and QQ plots, are saved as portable document format (.pdf) files. These results are presented in detail in

Chapter 4 (Results). Before reading next three chapters, we recommend you go directly to the tutorial chapter and run other tutorials.

1.3 How to use GAPIT user manual?

The next three chapters (2-4) describe details on the input data, type of analysis and output of results. The following chapter (5) is for users to use GAPIT for prototyping. The chapter 6 presents scenarios to demonstrate the applications. The last chapter (7) lists frequently questions and answers.

2 Data

The phenotypic data are the only data that must be directly provided by the user. Although the kinship is essential, it can be either provided by the user, or estimated automatically from the genotypic data. Genotypic data may not be needed for genomic prediction if the kinship matrix is provided by the user. Covariate variables (fixed effects), such as population structure represented by the Q matrix (subpopulation proportion) or principal components (PCs), is optional. GAPIT provides the option to calculate PCs from the genotypic data. All input files should be saved as a "Tab" delimited text file, and we recommend sorting the taxa in alphabetical order.

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.*

2.1 Phenotypic Data

The user has the option of performing GWAS on multiple phenotypes in GAPIT. 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. 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 genomic prediction. In the later case, genomic prediction is performed using a kinship matrix provided by the user.
GAPIT accepts genotypic data in either standard 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 genotype of each taxa to be stored in one 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 is 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 alleleic effect estimate will indicate that the minor allele is favorable.

2.2.2 Numeric format

GAPIT also accepts the numeric format used by EMMA. Columns are used for SNPs and rows are used for taxa. 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 (physical or genetic) 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.

Example file (X_3122_SNPs.txt from tutorial data set):

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

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

```
myGD <- read.table("X_3122_SNPs.txt", head = TRUE)
```

The "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.

Example file (IMPORTANT_Chromosome_BP_Location_of_SNPs.txt from tutorial data set):

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("IMPORTANT_Chromosome_BP_Location_of_SNPs.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.

Example (KSN.txt from tutorial data set):

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 presented in Section 1.1. 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.

Example file (mdp_population_structure.txt from tutorial data set):

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)
```

2.5 Import Genotype by File Names

Genotype data can be too large that it does not fit memory requirement. It can also be saved as multiple files, such as each from a chromosome. GAPIT is capable to import genotype by their file names. The file names must be named sequentially (e.g., "mdp_genotype_chr1.hmp.txt", "mdp_genotype_chr2.hmp.txt", …). For the Hapmap format. the common file name (e.g. "mdp_genotype_chr"), file name extension (e.g. "hmp.txt") are passed to GAPIT through the "file.G", "file.Ext.G". The starting file and the ending file are specified by file.from and file.to parameters. When the file is not in the working directory, the file path can be passed to GAPIT through file.path parameter.

For numeric format, 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.

2.6 Other GAPIT Input Parameters

In addition to the input parameters defined in previous sections, GAPIT has more parameters to define compression, PCA, GWAS and GPS (See Gallery of GAPIT Input Parameters).

Gallery of GAPIT input parameters

Parameter	Default	Options	Description
Y	NULL	User	Phenotype
KI	NULL	User	Kinship of Individual
CV	NULL	User	Covariate Variables
G	NULL	User	Genotype in hapmap format
GD	NULL	User	Genotype Data in numeric format
GM	NULL	User	Genotype Map for numeric format
file.Ext.G	NULL	User	file extention for Genotype in hapmap format
file.Ext.GD	NULL	User	file extention for Genotype Data in numeric format
file.Ext.GM	NULL	User	file extention for Genotype Map for numeric format
file.fragment	NULL	User	the fragment size to read each time within a file
file.G	NULL	User	The common name of file for genotype in hapmap format
file.GD	NULL	User	The common name of file for genotype map for numeric format
file.GM	NULL	User	The common name of file for genotype data in numeric format
file.path	NULL	User	Path for genotype files
file.from	0	>0	The first genotype files named sequentially
file.to	0	>0	The last genotype files named sequentially
group.by	10	>0	The grouping interval of compression
group.from	1	>1	The starting number of groups of compression
group.to	10000000	>1	The ending number of groups of compression
kinship.algorithm	VanRaden	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 centeral cloaction 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 Analysis

GAPIT is designed to accurately perform GWAS and genomic prediction on large data sets. It accomplishes this through implementation of state-of-the-art statistical methods, including compressed mixed linear model (CMLM), P3D (or EMMAx), and genomic prediction through CMLM. This section illustrates the options for using these methods.

3.1 Compressed Mixed Linear Model

A mixed linear model (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 or conducting a principal component analysis), the "Q+K" approach improves statistical power compared to "Q" only[2]. An MLM can be described using Henderson's matrix notation[6] as follows:

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e}, \tag{1}$$

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 random residual. The **u** and **e** vectors are assumed to be normally distributed with a null mean and a variance of:

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

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).

$$h^2 = \frac{\sigma^2_a}{\sigma^2_a + \sigma^2_e}, \tag{3}$$

We use cluster analysis 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. These parameters are group.from, group.to and group.by. Their defaults are 0, N and 10, where N is the total number of individuals.

3.2 Regular mixed Linear Model

Regular MLM 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.

3.3 General Linear Model

GAPIT has the option to a General Linear Model (GLM), which does not include the lines as random effects. This option can be simply performed by setting the number of groups to zero e.g. group.from=0 and group.to=0.

3.4 P3D/EMMAx

In addition to implementing compression, GAPIT can conduct EMMAx/P3D (Zhang *et al*., 2010; Kang *et al*., 2010). 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.

3.5 Genomic Prediction

Genomic prediction is performed with the method developed by Zhou et al (2011, in process), which is based on the CMLM approach that was proposed for GWAS [2,3]. The average genetic potential for a group, which is derived from the best linear unbiased predictions (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:

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

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.

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

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:

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

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:

$$\text{Reliability} = 1 - \frac{\text{PEV}}{\sigma_a^2} \tag{7}$$

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 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.

4.1 Principal Component (PC) plot

For each PC included in the GWAS and GPS models, the observed PC values are plotted. Every possible pair of these PCs are plotted against each other.

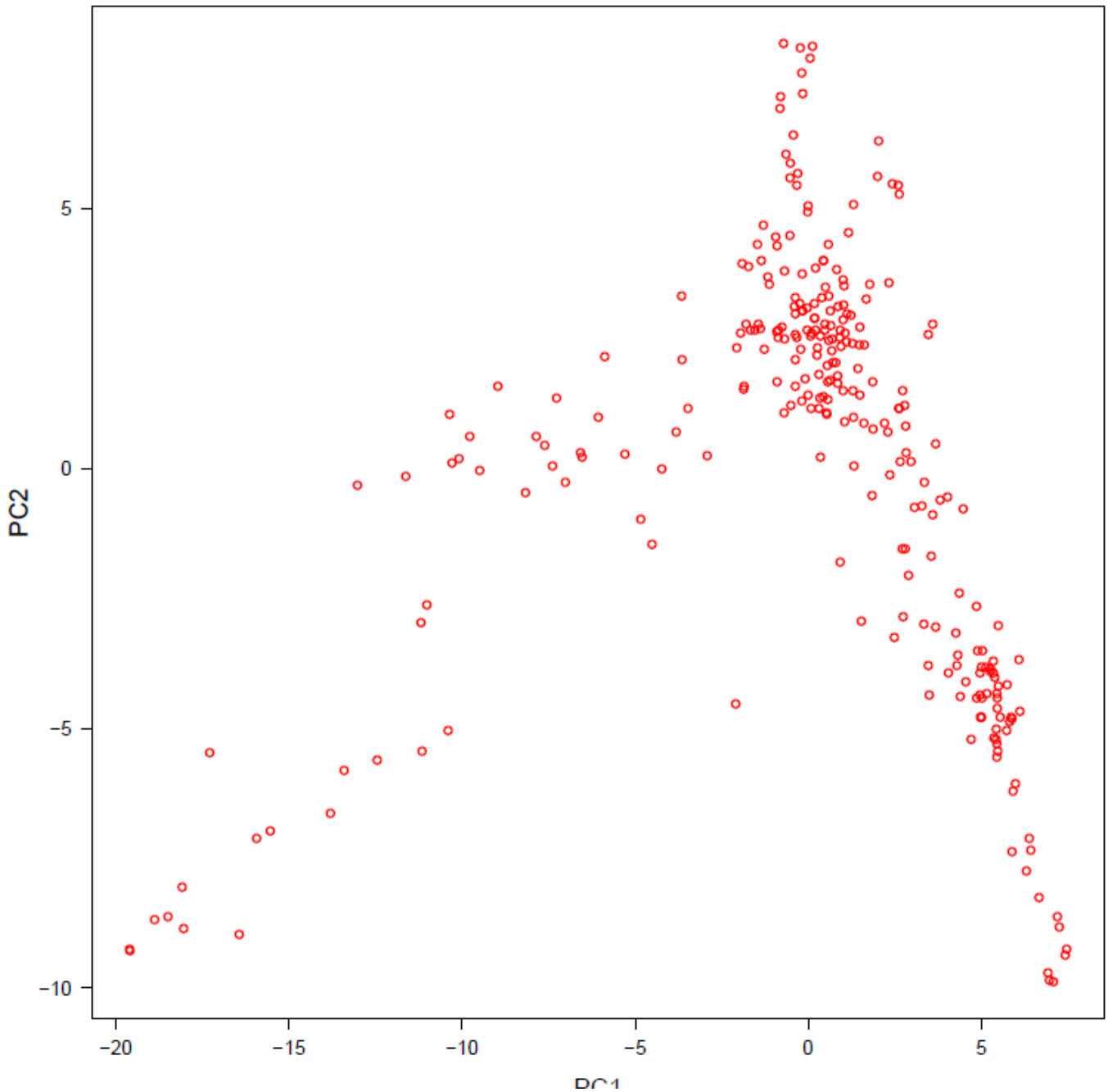


Figure 4.1. Plot of the second principal component (PC2) against the first principal component (PC1).

4.2 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.

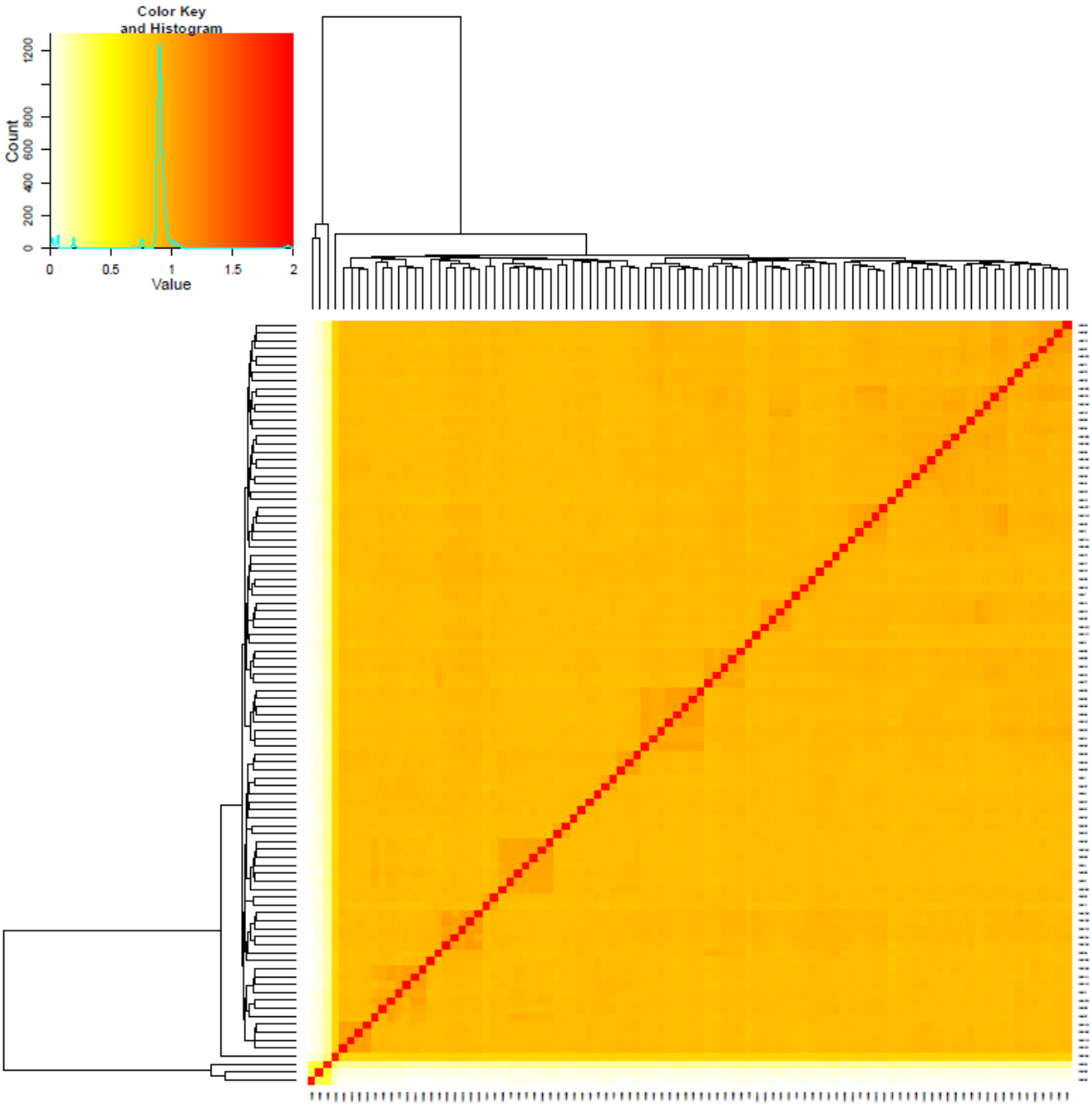


Figure 4.2. Kinship plot. A heat map of the values in the values in the kinship matrix is created.

4.3 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.

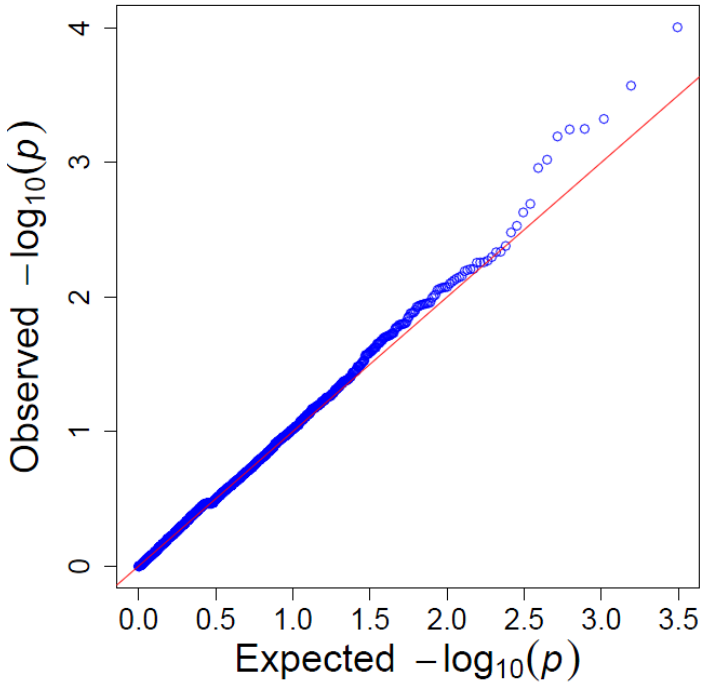


Figure 4.3. Quantile-quantile (QQ) –plot of *P*-values. The y-axis is the observed negative base 10 logarithm of the *P*-values, and the x-axis is the expected observed negative base 10 logarithm of the *P*-values under the assumption that the *P*-values follow a uniform[0,1] distribution.

4.4 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 4.4) and individual Manhattan plots for each chromosome.

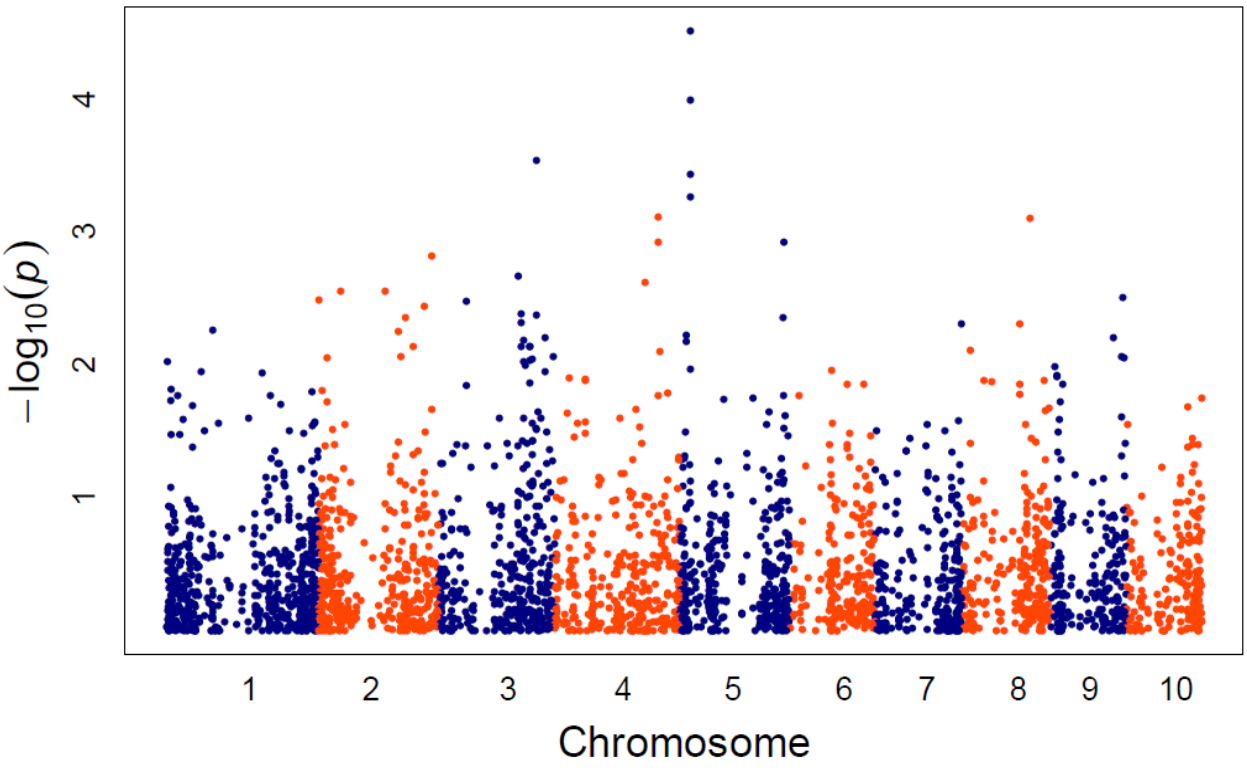


Figure 4.4. Manhattan plot. The X-axis is the genomic position of the SNPs in the genome, and the Y-axis is the negative log base 10 of the P-values. The chromosomes are alternatively colored. SNPs with stronger associations with the trait will have a large Y-coordinate value.

4.5 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 4.5. GWAS results for all SNPs that were analyzed.

SNP	Chromosome	Position	P.value	maf	nobs	Rsquare.without.SNP	Rsquare.with.SNP	FDR.adj.P.values
PZA03591.1	8	134813437	2.52E-05	0.174242424	264	0.437587182	0.477416352	0.077918424
PZB00149.1	1	188047564	0.000452725	0.21780303	264	0.437587182	0.4648972	0.341075833
PZB02516.2	3	193566873	0.000506399	0.170454545	264	0.437587182	0.464422237	0.341075833

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 (Benjamini and Hochberg, 1995).

4.6 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 4.6. Information of associated SNPs.

SNP	Chromosome	Position	Allelic Effect Estimate
PZB00859.1	1	157104	0.14601387
PZA01271.1	1	1947984	0.769014005
PZA03613.2	1	2914066	-0.272428161
PZA03613.1	1	2914171	-0.343080555

This table provides the SNP id, chromosome, bp position, and the allelic effect estimate of each SNP analyzed. When genotypic data in HapMap format are used, the sign of the allelic effect is relative to the minor allele.

4.7 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.

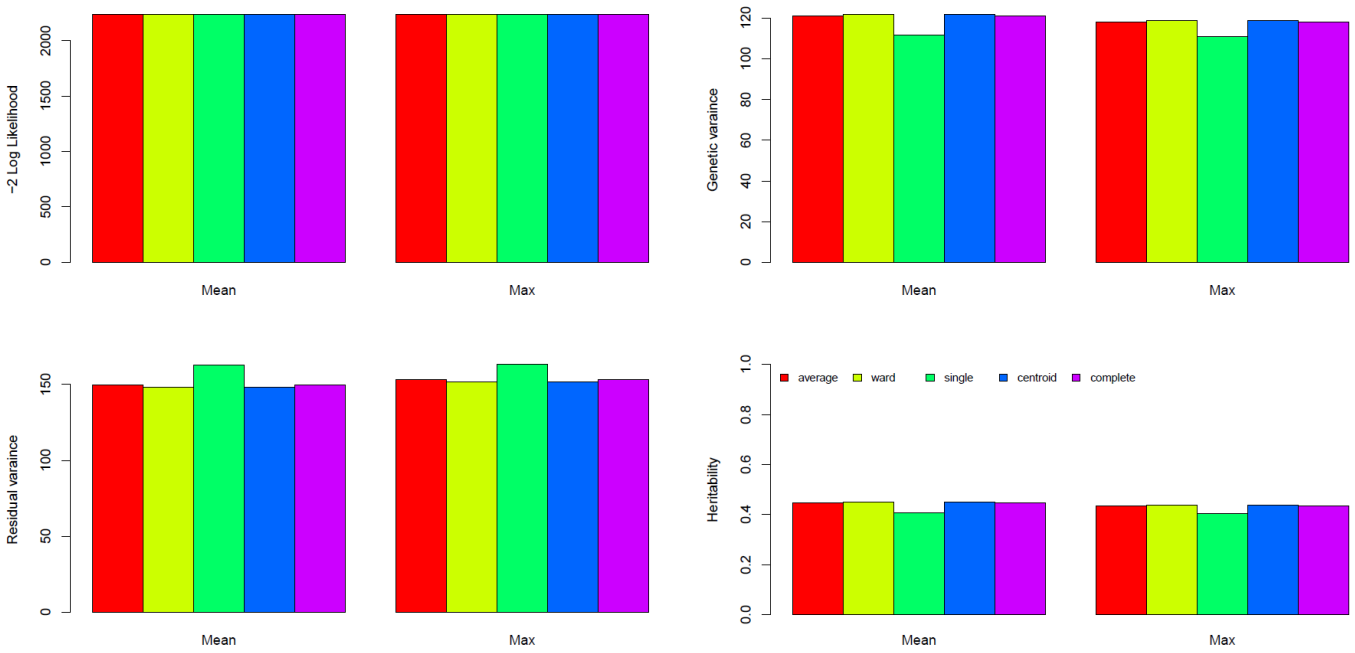


Figure 4.6.1. Compression profile with single group. The X-axis on all graphs display the summary statistic method used to obtain the group kinship matrix. The rectangles with different colors indicate the clustering algorithm used to group individuals.

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.

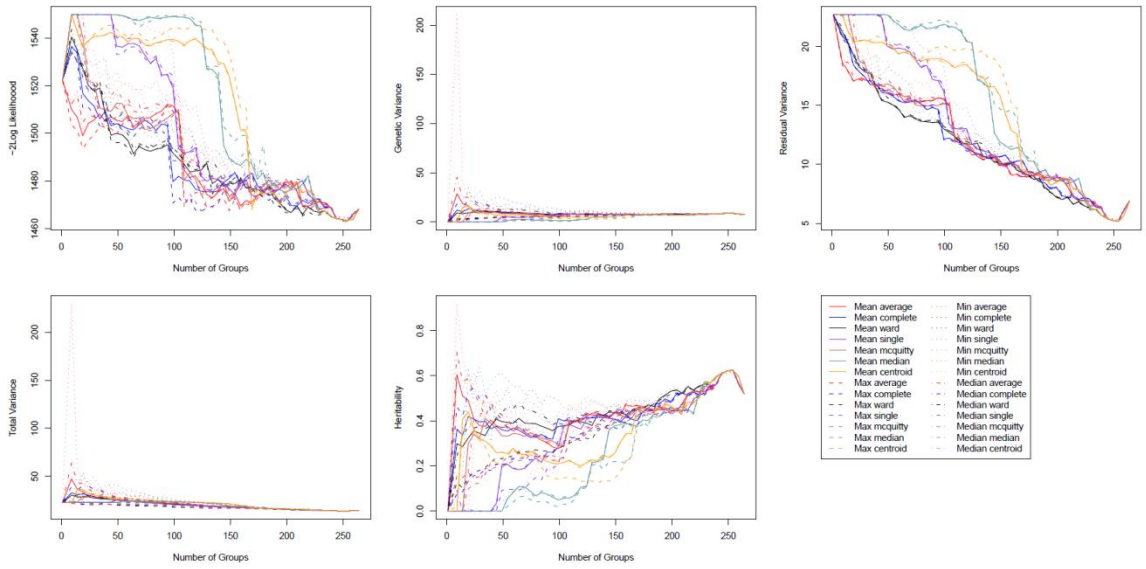


Figure 4.6.2. Compression profile over multiple groups. The X-axis on each graph is the number of groups considered, and the Y-axes on the graphs are the -2*log likelihood function, the estimated genetic variance components, the estimated residual variance component, the estimated total variance, and 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.

4.8 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.

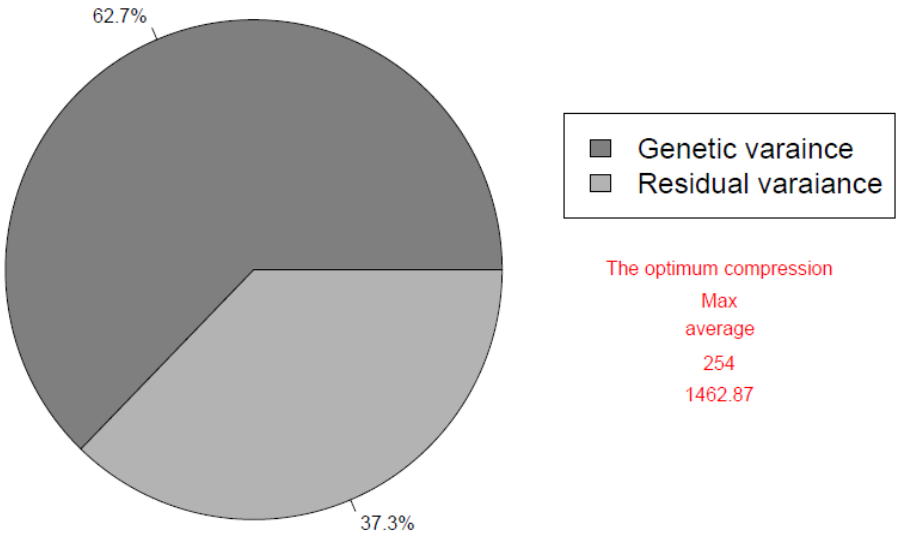


Figure 4.7. The profile for the optimum compression. The optimal method to calculate group kinship is "Max", the optimal clustering method is "average", the number of groups (ie., the dimension of the group kinship matrix) is 254, the value of -2*log likelihood function is 1462.87, and the heritability is 0.627.

4.9 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".

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.

4.10 Genomic Prediction

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

Table 4.6. Genomic Breeding values and prediction error variance.

Taxa	Group	RefInf	ID	BLUP	PEV
33-16	1	1	1	-2.29712	5.290034
38-11	2	1	2	2.741647	5.163874
4226	3	1	3	-5.05334	5.485718
4722	4	1	4	1.08699	4.763591
A188	5	1	5	-4.29157	5.476398
A214N	6	1	6	1.689654	5.124789
A239	7	1	7	-2.59947	6.183472
A272	8	1	8	0.795983	6.273644
A441-5	9	1	9	-0.20236	5.622717
A554	10	1	10	-0.98338	5.135618
A556	11	1	11	-1.16366	5.053223
A6	12	1	12	7.659746	6.701243

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.

4.11 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.

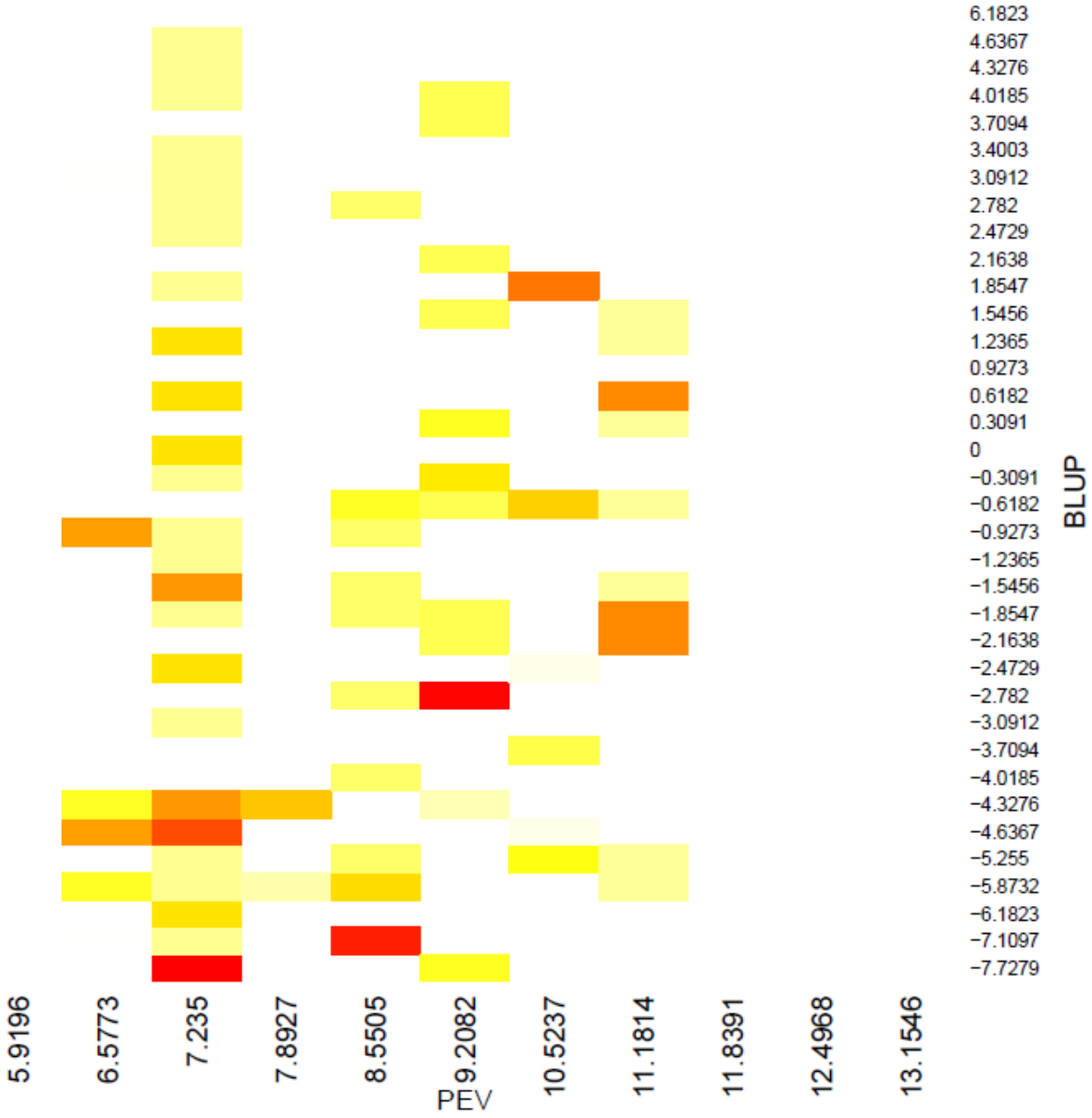


Figure 4.8. Joint distribution of genomic breeding value and prediction error variance.

5 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 (Flint-Garcia *et al*., 2005). 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*. (1995) and the fixed effects used to account for population structure were obtained from STRUCTURE (Pritchard *et al*., 2000).

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.

5.1 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 (2008) method, 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
)
```

5.2 *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). 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"),
group.from=200,
group.to=1000000,
group.by=10
)
```

5.3 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("Copy of Q_First_Three_Principal_Components.txt", head = TRUE)

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

5.4 Genomic Prediction

Genomic prediction can be performed without running GWAS. Thus, GWAS can be turned off by using "SNP.test=FALSE" option. 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)
myKI <- read.table("KSN.txt", head = FALSE)

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

5.5 *Multiple Genotype Files*

In this scenario, the HapMap genotypic data set from Scenario 1 is subdivided into multiple genotype files (e.g. 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.

5.6 *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
)
```

5.7 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.

5.8 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
)
```

5.9 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 "num.read" 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="X_3122_SNPs",
file.GM="IMPORTANT_Chromosome_BP_Location_of_SNPs",
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 (9999) 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.

5.10 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 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 (or only tutorials) 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.

This chapter describes the output of R object from GAPIT (see gallery of GAPIT output). Using these objects require knowledge of R. This is designed users whose are interested in: 1) to develop new method or software package on top of GAPIT; 2) to compare GAPIT with other new or existing statistical methods or software packages; 3) to study a specific topic with GAPIT.

Gallery of GAPIT output

Object	Dimension	Description
GWAS	mX7	GWAS results (SNPID,CHR,BP,P,MAF,DF,Effect)
GPS	nX6	GPS results (lineID,GroupID,Vategoty,BLUP,PEV)
compression	kX7	Compression result(VA,VE,REML,h2,group,cluster,method)
kinship	nXn	kinship among individuals
kinship.optimum	sXs	The optimum group kinship
PCA	NA	Principal components

6.1 Cross validation with replacement

Here we demonstrate an example from the last category. The way f using GAPIT for the first two categories can follow the same scheme. The scenario we used here is to investigate the accuracy of genome prediction through cross validation.

First we randomly set 25% of original phenotype (Y) as missing (NA) and generate genome prediction by using their kinship with the ones with phenotype. Then we record correlation between the genome prediction and the original phenotype. We repeat this process for 1000 times. The average of the correlation is used as the criteria of genome prediction accuracy. The R code of this study is displayed in the following box. The accuracy (correlation coefficient) over the 100 replicates 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("Copy of Q_First_Three_Principal_Components.txt", 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,
group.from=254,
group.to=254,
group.by=10,
kinship.cluster=c("average"),
kinship.group=c("Mean")
)
prediction=myGAPIT$GPS

#Seprate 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)
```

6.2 Cross validation without replacement

The validation can also be performed by excluding one or a set of individuals in the reference to derive genomic prediction of these individuals. Then repeat the process until all the individual 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))
randomdization=sample(1:n,n)
assignment=assignment[randomdization]
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,
group.from=254,
group.to=254,
group.by=10,
kinship.cluster=c("average"),
kinship.group=c("Mean")
)
prediction=myGAPIT$GPS

#Seprate 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)
```

6.3 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
```

6.4 Compile SNPs from multiple GAPIT analyses into one set of results

To further expedite completion of the analysis for a large amount of SNPs, one may want to split the SNPs into multiple subgroups, and simultaneously run GAPIT on multiple clusters or R consoles. The code below reads in GWAS results and allelic effect estimates from the GAPIT analysis run on each subgroup, then combines them, recalculates the adjusted *P*-values using the Benjamini-Hochberg (1995) FDR-controlling procedure, and produces a combined GWAS summary table, allelic effect estimates, Manhattan plots, and QQ-plot for all SNPs.

```
#Step 0: In your working directory, create three folders, and name them "Results_1",
#"Results_2", and "Results_Combined". Put all of the GAPIT output files from one set up SNPs in
# "Results_1", and all of the GAPIT output files from the second set of SNPs in "Results_2".
#  Finally, suppose that two traits named "trait1" and "trait2" are being analyzed
#############################################################################################
library('MASS')
library(multtest)
library(gplots)

setwd("C:\\Folder")

#Import the source code
source("http://www.maizegenetics.net/images/stories/bioinformatics/GAPIT/emma.txt")
source("http://www.maizegenetics.net/images/stories/bioinformatics/GAPIT/gapit_functions.txt")

#Step 1: Set data directory and import files
#############################################################################################
mydataPath.Results.1="C:\\Folder\\Results_1\\"
mydataPath.Results.2="C:\\Folder\\Results_2\\"

name <- c("trait1", "trait2")

#Step 2: Set the result directory to where the combined data will go
#############################################################################################
setwd("C:\\Folder\\Results_Combined")

for(i in 1:length(name)){

#Read in the GWAS and Allelic effect estimates
GWAS.Results.1  <- read.csv(paste(mydataPath.Results.1,"GAPIT.",name[i],".GWAS.Results.csv",sep=""), head = TRUE)
GWAS.Results.2 <- read.csv(paste(mydataPath.Results.2,"GAPIT.",name[i],".GWAS.Results.csv",sep=""), head = TRUE)
Effect.Estimates.1  <- read.csv(paste(mydataPath.Results.1,"GAPIT.",name[i],".Allelic_Effect_Estimates.csv",sep=""), head = TRUE)
Effect.Estimates.2  <- read.csv(paste(mydataPath.Results.2,"GAPIT.",name[i],".Allelic_Effect_Estimates.csv",sep=""), head = TRUE)

#Append the the GWAS results and alleleic effect estimates onto one folder
GWAS.Results <- rbind(GWAS.Results.1, GWAS.Results.2)
Effect.Estimates <- rbind(Effect.Estimates.1, Effect.Estimates.2)

#Remove the last column, which is the FDR adjusted P-values
GWAS.Results <- GWAS.Results[,-ncol(GWAS.Results)]
#Run the B-H procedure on the combined data, and append the FDR-adjusted P-values to the GWAS.Results
Conduct.FDR <- GAPIT.Perform.BH.FDR.Multiple.Correction.Procedure(PWI = GWAS.Results,
                    FDR.Rate = 0.05, FDR.Procedure = "BH")
GWAS.Results.with.FDR <- Conduct.FDR$PWIP

#Make QQ-plots
GAPIT.QQ(P.values = Conduct.FDR$PWIP[,4], name.of.trait = name[i],DPP=50000)

#Make new Manhattan plots with the combined results
GAPIT.Manhattan(GI.MP = GWAS.Results.with.FDR[,2:4], name.of.trait = name[i],
             DPP=50000, plot.type = "Genomewise",cutOff=0.00)
GAPIT.Manhattan(GI.MP = GWAS.Results.with.FDR[,2:4], name.of.trait = name[i],
             DPP=50000, plot.type = "Chromosomewise",cutOff=0.00)

#Export the combined GWAS results
write.table(GWAS.Results.with.FDR, paste("GAPIT.", name[i], ".GWAS.Results.csv", sep = ""),
            quote = FALSE, sep = ",", row.names = FALSE,col.names = TRUE)

#Export the combined Alleleic effect estiamtes
write.table(Effect.Estimates, paste("GAPIT.", name[i], ".Allelic_Effect_Estimates.csv", sep = ""),
            quote = FALSE, sep = ",", row.names = FALSE,col.names = TRUE)

rm(GWAS.Results.1)
rm(GWAS.Results.2)
rm(GWAS.Results)
rm(GWAS.Results.with.FDR)
rm(Effect. Estimates.1)
rm(Effect. Estimates.2)
rm(Effect. Estimates)
rm(Conduct.FDR)
} #End for(i in 1:length(name))
```

7 Appendix

7.1 Tutorial Data sets

The data set contains 9 files and can be downloaded at:
http://www.maizegenetics.net/images/stories/bioinformatics/GAPIT/gapit_tutorial_data.zip

Table 7.1. Properties of tutorial files

Name	Date	Time	Bytes
Copy of Q_First_Three_Principal_Components.txt	6/8/2011	11:00 PM	35,604
mdp_SNP_information.txt	6/14/2012	1:13 PM	73,728
mdp_SNP_information1.txt	6/14/2012	1:35 PM	23,325
mdp_SNP_information2.txt	6/14/2012	1:36 PM	23,539
mdp_SNP_information3.txt	6/14/2012	1:37 PM	26,011
KSN.txt	5/3/2011	2:56 PM	724,004
mdp_genotype_chr1.hmp.txt	9/2/2011	6:18 PM	489,479
mdp_genotype_chr2.hmp.txt	9/2/2011	6:19 PM	356,629
mdp_genotype_chr3.hmp.txt	9/2/2011	6:20 PM	322,390
mdp_genotype_chr4.hmp.txt	9/2/2011	6:20 PM	289,859
mdp_genotype_chr5.hmp.txt	9/2/2011	6:21 PM	324,154
mdp_genotype_chr6.hmp.txt	9/2/2011	6:22 PM	194,119
mdp_genotype_chr7.hmp.txt	9/2/2011	6:23 PM	223,920
mdp_genotype_chr8.hmp.txt	9/2/2011	6:23 PM	232,994
mdp_genotype_chr9.hmp.txt	9/2/2011	6:24 PM	194,047
mdp_genotype_chr10.hmp.txt	9/2/2011	6:25 PM	183,469
mdp_genotype_test.hmp.txt	4/28/2011	12:16 PM	2,796,282
mdp_genotype_test1.hmp.txt	6/2/2011	2:00 PM	904,942
mdp_genotype_test2.hmp.txt	6/2/2011	2:00 PM	905,156
mdp_genotype_test3.hmp.txt	6/2/2011	1:58 PM	989,468
mdp_genotype_test5.hmp.txt	5/3/2011	2:56 PM	724,004
mdp_traits.txt	5/20/2011	11:34 PM	6,586
mdp_numeric.txt	6/19/2012	9:40 AM	1,834,937
mdp_numeric1.txt	6/19/2012	9:41 AM	594,195
mdp_numeric2.txt	6/19/2012	9:43 AM	594,134
mdp_numeric3.txt	6/19/2012	9:44 AM	649,748

These file can be classified into following categories. We use "?" and "*" for a string with any length and one character respectively.

1. "Copy of *" is a covariate data set consisting of three PCs in maize.
   mdp_SNP_information*" are genotype map files containing SNP name, chromosome, and base pair position. The ones with numeric number are the ones separated from the one does not have numeric number.

2. "KSN.txt" is a kinship matrix of a subset of themaize diversity panel calculated using the method of Loiselle *et al*. (1995).
3. "mdp_genotype_*.hmp.txt" is genotype in format of hapmap containing 3093 *SNPs*. The files with numeric numbers are the ones separated from "mdp_genotype_test.hmp.txt". The files with "chr" are separated by chromosome and the ones with "test" are separated by block (1000,1000 and 1093 SNPs).
4. " mdp_traits.txt" is phenotype containing three traits: ear height (EarHT), days to pollinating (dpoll) and ear diameter (EarDia).
5. "mdp_numeric?.txt" is numeric genotype data files paired with "mdp_SNP_information?.txt". The one with numeric number are the ones separated from the without the numeric numbers. The first two files contain 1000 SNPs and the last one contain 1094 SNPs.

7.2 Typical ways of reading data

```
#Phenotypic Data
#myY  <- read.table("Phenotype_dpoll.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("Copy of Q_First_Three_Principal_Components.txt", head = TRUE)
```

7.3 Frequently Asked Questions

1. What do I do if I get frustrated?

A: The GAPIT team makes the effort to provide suggestions to any errors that users might run into. Try to find the answer from this question list. If not there, email the problem to the corresponding author.

2. What happen if the magnitude of my trait is too small or large?

A: This will cause problem due to rounding error. Multiple or divide by a value to make it into a reasonable range.

3. How many PCs to include?

A: 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 scree 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.

4. 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 sent your R source code and the dataset that allow us to repeat the error.

5. 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.

6. 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.

7. What should I do with "Error in solve.default(crossprod(X, X)) : system is computationally singular"?

A: Check covariate variables and remove the ones that are linear independent.

8. How to cite GAPIT?

A: GAPIT can be cite as follows:
Alexander E. Lipka, Feng Tian, Qishan Wang, Jason Peiffer, Meng Li, Peter J Bradbury, Michael Gore, Edward S Buckler and Zhiwu Zhang (2012). GAPIT: Genome Association and Prediction Integrated Tool. *Bioinformatics* doi: 10.1093/bioinformatics/bts444.

9. **Is it possible to analyze case-control studies in GAPIT?**

A: GAPIT was designed to use mixed model approaches for performing GWAS and GS on quantitative phenotypes in association panels. This program may be used for case-control studies, where a set of individuals with an affliction are matched up with healthy individuals with similar demographic characteristics. Thus, case-control studies use binary phenotypes (i.e., case versus control). Because of this, the approaches used in GAPIT are not necessarily statistically appropriate for case-control studies. However, the effectiveness of the mixed model for controlling population structure is still relevant for finding biologically real associations with diseases. Because of this, we recommend augmenting GAPIT GWAS results with fitting logistic regression models at the most significant SNPs. To illustrate the analysis of case-control studies in GAPIT, we included the "for.exercise" case –control data set from the snpStats R program (http://bioc.ism.ac.jp/2.8/bioc/html/snpStats.html) plus results from running the CMLM approach in GAPIT in our tutorial data. These data consist of 500 cases and 500 controls that are simulated on human chromosome 10.

7.4 GAPIT Biography

Date	Version	Event
May 11, 2011	1.20	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.
June 13, 2011	1.22	Genotype in numerical format in addition to hapmap format
August 15, 2011	1.25	Van Raden algorithm for kinship
September 2, 2011	1.31	Reading fragment within single genotype file to save memory
September 9, 2011	1.34	LD heat map
September 17, 2011	1.36	Interface change for prototyping
October 4, 2011	1.40	Genotype, kinship and PC process optimization to improve speed for multiple traits
October 24, 2011	1.41	Missing data imputation as middle, major, minor, present/absent. Heterozygous coding as the middle, left (one of the homozygous) and right (the other homozygous)
November 1, 2011	1.42	Matrix participation to improve speed (5-10 folds)
December 7, 2011	2.01	FaST-LMM method

REFERENCES

1. Kang, H.M. et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet **42,** 348-354 (2010).
2. Bradbury, P.J. et al.TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics **23:** 2633-2635 (2010).
3. Pritchard, J.K., Stephens, M., Rosenberg, N.A. & Donnelly, P. Association mapping in structured populations. *American Journal of Human Genetics* **67**, 170-181 (2000).
4. Yu, J.M. et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**, 203-208 (2006).
5. Zhang, Z. et al. Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* **42**, 355-60 (2010).
6. Kang, H.M. et al. Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics* **178**, 1709-1723 (2008).
7. Loiselle, B.A., *et al*. Spatial Genetic-Structure of a Tropical Understory Shrub, Psychotria Officinalis (Rubiaceae), *Am J Bot*, **82**, 1420-1425 (1995).
8. Flint-Garcia SA, Thuillet AC, Yu J, Pressoir G, Romero SM, Mitchell SE, Doebley J, Kresovich S, Goodman MM, Buckler ES Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J*. 44: 1054–1064 (2005).
9. Henderson, C.R. Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics* **31**, 423-447 (1975).