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

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

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

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Method paper</th>
<th>GAPIT Implementation</th>
</tr>
</thead>
</table>

Note: These references are listed in section of Reference.

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

1 INTRODUCTION .......................................................................................................................................................... 5
  1.1 WHY GAPIT? .............................................................................................................................................................. 5
  1.2 GETTING STARTED ...................................................................................................................................................... 6
  1.3 HOW TO USE THE GAPIT USER MANUAL? .................................................................................................................. 6

2 INPUT DATA .................................................................................................................................................................... 8
  2.1 PHENOTYPIC DATA ..................................................................................................................................................... 8
  2.2 GENOTYPIC DATA ......................................................................................................................................................... 9
    2.2.1 HAPMAP FORMAT .................................................................................................................................................. 9
    2.2.2 NUMERIC FORMAT .............................................................................................................................................. 10
  2.3 KINSHIP ......................................................................................................................................................................... 11
  2.4 COVARIATE VARIABLES ............................................................................................................................................... 12
  2.5 IMPORT GENOTYPE BY FILE NAMES .......................................................................................................................... 12

3 ANALYSIS ...................................................................................................................................................................... 13
  3.1 MIXED LINEAR MODEL (MLM) .................................................................................................................................. 15
  3.2 COMPRESSED MLM (CMLM) ................................................................................................................................... 16
  3.3 GENERAL LINEAR MODEL (GLM) ............................................................................................................................... 16
  3.4 P3D/EMMAX ................................................................................................................................................................. 17
  3.5 SUPER ............................................................................................................................................................................ 17
  3.6 MULTIPLE LOCUS MIXED LINEAR MODEL (MLMM) .................................................................................................... 17
  3.7 FARMCPU ........................................................................................................................................................................ 17
  3.8 BLINK ............................................................................................................................................................................... 18
  3.9 GENOMIC BLUP ............................................................................................................................................................ 18
  3.10 COMPRESSED GBLUP ............................................................................................................................................... 19
  3.11 SUPER GBLUP ........................................................................................................................................................... 19

4 OUTPUT RESULTS .......................................................................................................................................................... 20
  4.1 PHENOTYPE DIAGNOSIS .............................................................................................................................................. 20
  4.2 MARKER DENSITY ......................................................................................................................................................... 20
  4.3 LINKAGE DISEQUILIBRIUM DECADE .......................................................................................................................... 21
  4.4 HETEROZYGOSIS ............................................................................................................................................................ 22
  4.5 PRINCIPAL COMPONENT (PC) PLOT ............................................................................................................................... 22
  4.6 KINSHIP PLOT ............................................................................................................................................................... 23
  4.7 NEIGHBOR-JOINING (NJ)-TREE .................................................................................................................................. 25
  4.8 QQ-plot ............................................................................................................................................................................. 25
  4.9 MANHATTAN PLOT .......................................................................................................................................................... 27
  4.10 ASSOCIATION TABLE .................................................................................................................................................... 28
  4.11 ALLELIC EFFECTS TABLE ............................................................................................................................................ 28
  4.12 COMPRESSION PROFILE ............................................................................................................................................ 28
  4.13 THE OPTIMUM COMPRESSION .................................................................................................................................. 31
  4.14 MODEL SELECTION RESULTS .................................................................................................................................. 32
  4.15 MULTIPLE TRAITS OR METHODS .................................................................................................................................. 32
  4.16 GENOMIC PREDICTION .................................................................................................................................................. 33
  4.17 DISTRIBUTION OF BLUPs AND THEIR PEV .................................................................................................................. 34
  4.18 INTERACTIVE PLOT ...................................................................................................................................................... 35

5 TUTORIALS ....................................................................................................................................................................... 36
5.1 A BASIC SCENARIO ................................................................................................................. 36
5.2 ENHANCED COMPRESSIoN .................................................................................................. 36
5.3 USER-INPUTTED KINSHIP MATRIX AND COVARIATES ....................................................... 37
5.4 GENOMIC PREDICTION ...................................................................................................... 37
5.5 MULTIPLE GENOTYPE FILES .......................................................................................... 38
5.6 NUMERIC GENOTYPE FORMAT ....................................................................................... 38
5.7 NUMERIC GENOTYPE FORMAT IN MULTIPLE FILES ...................................................... 39
5.8 FRACTIONAL SNPs FOR KINSHIP AND PCs ...................................................................... 39
5.9 MEMORY SAVING .............................................................................................................. 40
5.10 MODEL SELECTION .......................................................................................................... 40
5.11 SUPER ............................................................................................................................... 40
5.12 MLMM ............................................................................................................................... 41
5.13 FARM-CPU ....................................................................................................................... 41
5.14 MULTIPLE MODEL ............................................................................................................ 41

6 PROTOTYPE ...................................................................................................................................... 43
6.1 PREPARATION ....................................................................................................................... 43
6.2 GALLERY OF GAPIT GRAPHIC OUTPUT .......................................................................... 43
6.3 GALLERY OF GAPIT FILE OUTPUT ................................................................................ 49
6.4 CROSS VALIDATION WITH REPLACEMENT .................................................................... 49
6.5 CROSS VALIDATION WITHOUT REPLACEMENT .......................................................... 51
6.6 CONVERT HMap FORMAT TO NUMERICAL .................................................................. 52
6.7 COMPILE SNPS FROM MULTIPLE GAPIT ANALYSES INTO ONE SET OF RESULTS ....... 52
6.8 GENOMIC PREDICTION .................................................................................................... 54

7 APPENDIX .................................................................................................................................... 55
7.1 TUTORIAL DATA SETS ....................................................................................................... 55
7.2 TYPICAL WAYS OF READING DATA ................................................................................ 56
7.3 GAPIT OUTPUT FILES ...................................................................................................... 57
7.4 FREQUENTLY ASKED QUESTIONS .................................................................................. 58
1. HOW TO CITE GAPIT? ....................................................................................................... 58
2. WHAT DO I DO IF I GET FRUSTRATED? ............................................................................ 58
3. WHY GAPIT HAS DIFFERENT RESULTS FROM OTHER SOFTWARE? ......................... 58
4. THERE ARE MANY METHODS IMPLEMENTED IN GAPIT, WHICH ONE SHOULD I USE? . 58
5. HOW MANY PCs TO INCLUDE? .......................................................................................... 58
6. IS IT FEASIBLE I COMPARE DIFFERENT MODELS BY MYSELF? ................................ 58
7. HOW DO I REPORT AN ERROR? ......................................................................................... 58
8. WHAT SHOULD I DO WITH “ERROR IN FILE (FILE, "RT") : CANNOT OPEN THE CONNECTION”? 59
9. WHAT SHOULD I DO WITH “ERROR IN GAPIT (... : UNUSED ARGUMENT(S) ...”)? .......... 59
10. HOW DEAL WITH “ERROR IN SOLVE.DEFAULT(CROSSPROD(X, X)) : SYSTEM IS COMPUTATIONALLY SINGULAR” ? ....................................................... 59
11. HOW TO FIX THE ERROR OF USING COVARIATES FROM STRUCTURE AS FIXED EFFECTS? 59
12. SHOULD I REMOVE SNPS WITH MAF BELOW 5%? ......................................................... 59
13. MY TRAIT WAS MEASURED IN MULTIPLE ENVIRONMENTS, HOW DO I USE THEM SIMULTANEOUSLY? ................................................................. 59
14. IS IT OK TO ANALYZE BINARY TRAITS (CASE-CONTROL) WITH GAPIT? ............... 59
15. WILL NORMALITY TRANSFORMATION HELP? .................................................................. 59
16. SHOULD I USE PCs OR Q MATRIX? ................................................................................ 59
7.5 GAPIT BIOGRAPHY .............................................................................................................. 60
REFERENCES .................................................................................................................................... 61
1 INTRODUCTION

1.1 Why GAPIT?

GAPIT implemented a series of methods for Genome Wide Association (GWAS) and Genomic Selection (GS). The software and related methods received thousands of citations:

<table>
<thead>
<tr>
<th>Publication</th>
<th>Description</th>
<th>Year</th>
<th>Citations*</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAPIT version 1</td>
<td>Implemented GLM, MLM, CMLM and gBLUP</td>
<td>2012</td>
<td>850</td>
</tr>
<tr>
<td>GAPIT version 2</td>
<td>Implemented ECMLM and SUPER</td>
<td>2016</td>
<td>80</td>
</tr>
<tr>
<td>MLM</td>
<td>Known as QK model to control both population (Q matrix) structure and kinship (K).</td>
<td>2005</td>
<td>2,800</td>
</tr>
<tr>
<td>Compressed MLM</td>
<td>Individual is compressed into groups to reduce confounding between kinship and testing markers.</td>
<td>2010</td>
<td>1,200</td>
</tr>
<tr>
<td>SUPER</td>
<td>Derived kinship from associated markers (QTNs) instead of all markers and use QTNs that are complementary to testing markers.</td>
<td>2014</td>
<td>80</td>
</tr>
<tr>
<td>Enriched compressed MLM</td>
<td>Optimization of group kinship</td>
<td>2014</td>
<td>50</td>
</tr>
<tr>
<td>MLMM</td>
<td>Step wise regression to include associated markers in MLM</td>
<td>2012</td>
<td>500</td>
</tr>
<tr>
<td>FarmCPU</td>
<td>Select associated markers as cofactor to control false positives using likelihood in MLM to avoid overfitting and test markers without kinship to eliminate confounding between kinship and testing markers in iterative fashion.</td>
<td>2016</td>
<td>220</td>
</tr>
<tr>
<td>BLINK</td>
<td></td>
<td>2019</td>
<td>4</td>
</tr>
<tr>
<td>gBLUP</td>
<td>Use RFLP markers to derive kinship among maize lines to estimate breeding values</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marker based MTDFREML</td>
<td>Add second path to use kinship derived from markers</td>
<td>2006</td>
<td>40</td>
</tr>
<tr>
<td>Efficient kinship</td>
<td>Define kinship as marker matrix multiplied by its transpose.</td>
<td>2008</td>
<td>3,600</td>
</tr>
<tr>
<td>sBLUP</td>
<td>Derive kinship from QTNs from SUPER method to estimate breeding values. The method has higher prediction accuracy than gBLUP for traits controlled with less number of genes.</td>
<td>2019</td>
<td>7</td>
</tr>
<tr>
<td>cBLUP</td>
<td>Estimate individuals’ breeding values by the breeding values of their corresponding groups. The method has higher prediction accuracy than gBLUP for traits with low heritability.</td>
<td>2019</td>
<td>7</td>
</tr>
</tbody>
</table>

*Google Scholar by April 4, 2020,
In addition to the multiple available methods, GAPIT is user friendly and produces comprehensive reports to interpret data and results in publication ready formats. For examples, GAPIT accept both numeric and hapmap genotype formats. The individuals in the phenotype files do not have to be in the same order as genotype files, or even can be partially different. GAPIT produces the distribution of marker density and decay of linkage equilibrium to inform user if the markers are dense enough. When multiple GWAS methods are selected, GAPIT produces the corresponding Manhattan plots with overlapped associated markers highlighted as illustrated on the right for the GAPIT demo data from maize. The results show that GLM method identified association signals above the Bonferroni threshold (dash red lines). However, the association signals are inflated across the genome (the red dots on the QQ plots). BLINK method identified two associated markers, including the marker close to a flowering time gene, VGT1 on chromosome 8. The QQ plot suggests that 99% of the markers have p values under expectation indicated by the solid red line.

1.2 Getting Started

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

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

The easiest way to GAPIT is to COPY/PASTE some code from GAPIT tutorial script and edit to fit your data and analysis. Here is an example using five methods on the data from Zhiwu Zhang Lab website.

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

#GWAS with five methods
myGAPIT_MLM <- GAPIT(
  Y=myY[,c(1,3)], #fist column is individual ID, the third columns is days to pollination
  GD=myGD,
  GM=myGM,
  PCA.total=3,
  model=c("GLM", "MLM", "CMLM", "FarmCPU", "Blink"),
  Multiple_analysis=TRUE)
```

The analysis should be completed within a few minutes. In your current R working directory, you should find multiples files with three types of extensions: pdf, csv, and txt. The pdf files include the Manhattan plot and the QQ plot displayed above.

1.3 How to use the GAPIT user manual?

The next three chapters (2-4) describe details on the input data, type of analysis and output of results. Chapter 5 presents scenarios to demonstrate the applications. Chapter 6 is for users to use GAPIT for
prototyping. The last chapter (7) lists frequently questions and answers. Before reading the next three chapters, we recommend you go directly to the tutorial chapter and run other tutorials.
2 Input Data

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

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

Table 2.1 Gallery of GAPIT input data

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Default</th>
<th>Options</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>NULL</td>
<td>User</td>
<td>Phenotype</td>
</tr>
<tr>
<td>KI</td>
<td>NULL</td>
<td>User</td>
<td>Kinship Matrix</td>
</tr>
<tr>
<td>CV</td>
<td>NULL</td>
<td>User</td>
<td>Covariate Variables</td>
</tr>
<tr>
<td>G</td>
<td>NULL</td>
<td>User</td>
<td>Genotype Data in Hapmap Format</td>
</tr>
<tr>
<td>GD</td>
<td>NULL</td>
<td>User</td>
<td>Genotype Data in Numeric Format</td>
</tr>
<tr>
<td>GM</td>
<td>NULL</td>
<td>User</td>
<td>Genotype Map for Numeric Format</td>
</tr>
<tr>
<td>file.Ext.GD</td>
<td>NULL</td>
<td>User</td>
<td>File Extension for Genotype Data in Numeric Format</td>
</tr>
<tr>
<td>file.Ext.GM</td>
<td>NULL</td>
<td>User</td>
<td>File Extension for Genotype Map for Numeric Format</td>
</tr>
<tr>
<td>file.fragment</td>
<td>NULL</td>
<td>User</td>
<td>The Fragment Size to Read Each Time within a File</td>
</tr>
<tr>
<td>file.G</td>
<td>NULL</td>
<td>User</td>
<td>The Common Name of File for Genotype in Hapmap Format</td>
</tr>
<tr>
<td>file.GD</td>
<td>NULL</td>
<td>User</td>
<td>The Common Name of File for Genotype Map for Numeric Format</td>
</tr>
<tr>
<td>file.GM</td>
<td>NULL</td>
<td>User</td>
<td>The Common Name of File for Genotype Data in Numeric Format</td>
</tr>
<tr>
<td>file.path</td>
<td>NULL</td>
<td>User</td>
<td>Path for Genotype Files</td>
</tr>
<tr>
<td>file.from</td>
<td>0</td>
<td>&gt;0</td>
<td>The First Genotype Files Named Sequentially</td>
</tr>
<tr>
<td>file.to</td>
<td>0</td>
<td>&gt;0</td>
<td>The Last Genotype Files Named Sequentially</td>
</tr>
</tbody>
</table>

Notice: The inputs in italic are supported by GAPIT version 1 and 2 (before 2017) only.

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:
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:

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

### 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.

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

```r
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:

<table>
<thead>
<tr>
<th>Genotype Code</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>R</th>
<th>Y</th>
<th>S</th>
<th>W</th>
<th>K</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>A</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>R</td>
<td>Y</td>
<td>S</td>
<td>W</td>
<td>K</td>
<td>M</td>
</tr>
<tr>
<td>CC</td>
<td>A</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>R</td>
<td>Y</td>
<td>S</td>
<td>W</td>
<td>K</td>
<td>M</td>
</tr>
<tr>
<td>GG</td>
<td>A</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>R</td>
<td>Y</td>
<td>S</td>
<td>W</td>
<td>K</td>
<td>M</td>
</tr>
<tr>
<td>TT</td>
<td>A</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>R</td>
<td>Y</td>
<td>S</td>
<td>W</td>
<td>K</td>
<td>M</td>
</tr>
<tr>
<td>AG</td>
<td>A</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>R</td>
<td>Y</td>
<td>S</td>
<td>W</td>
<td>K</td>
<td>M</td>
</tr>
<tr>
<td>CT</td>
<td>A</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>R</td>
<td>Y</td>
<td>S</td>
<td>W</td>
<td>K</td>
<td>M</td>
</tr>
<tr>
<td>CG</td>
<td>A</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>R</td>
<td>Y</td>
<td>S</td>
<td>W</td>
<td>K</td>
<td>M</td>
</tr>
<tr>
<td>AT</td>
<td>A</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>R</td>
<td>Y</td>
<td>S</td>
<td>W</td>
<td>K</td>
<td>M</td>
</tr>
<tr>
<td>GT</td>
<td>A</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>R</td>
<td>Y</td>
<td>S</td>
<td>W</td>
<td>K</td>
<td>M</td>
</tr>
<tr>
<td>AC</td>
<td>A</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>R</td>
<td>Y</td>
<td>S</td>
<td>W</td>
<td>K</td>
<td>M</td>
</tr>
</tbody>
</table>

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

### 2.2.2 Numeric format

GAPIT also accepts the numeric format 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 (mdp_numeric.txt from tutorial data set):**

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

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

```r
myGD <- read.table("mdp_numeric.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 (mdp_SNP_information.txt from tutorial data set):

<table>
<thead>
<tr>
<th>Name</th>
<th>Chromosome</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>PZB00859.1</td>
<td>1</td>
<td>157104</td>
</tr>
<tr>
<td>PZA01271.1</td>
<td>1</td>
<td>1947984</td>
</tr>
<tr>
<td>PZA03613.2</td>
<td>1</td>
<td>2914066</td>
</tr>
<tr>
<td>PZA03613.1</td>
<td>1</td>
<td>2914171</td>
</tr>
<tr>
<td>PZA03614.2</td>
<td>1</td>
<td>2915078</td>
</tr>
</tbody>
</table>

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

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

### 2.3 Kinship

The kinship matrix file (called “KI” in GAPIT) is formatted as an n by n+1 matrix where the first column is the taxa name, and the rest is a square symmetric matrix. Unlike the other input data files, the first row of the kinship matrix file does not consist of headers.

Example (KSN.txt from tutorial data set):

```
33-15 2 0.228837 0.229322 0.268842 0.237145 0.0781 0.347107
38-11 2 0.249635 0.293708 0.175211 0.079276 0.295606
4226 2 0.249635 0.244965 0.214859 0.236153 0.082693 0.283713
4722 2 0.268842 0.293708 0.214859 0.259352 0.061573 0.160104
A188 2 0.237145 0.175211 0.236153 0.259352 0.061469 0.227900
A214N 2 0.0781 0.079276 0.082693 0.061573 0.061469 0.110364
A239 2 0.347107 0.295606 0.283713 0.160104 0.232799 0.110364
```

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

```r
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 2.1. 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):

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

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

```r
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.

This option is supported by GAPIT version 1 and 2 (before 2017) only. Historical source code can be found at GAPIT website (http://zzlab.net/GAPIT/previous). For large datasets that GAPIT can not handle, BLINK (http://zzlab.net/blink) is recommended.
3 Analysis

GAPIT implemented a series of methods for GWAS and genomic selection toward high statistical power, high prediction accuracy, and high computing speed, see Figure 3.1. The methods include General Linear Model (GLM), Mixed Linear Model (MLM), Compressed MLM (CMLM), Enriched CMLM (ECMLM), Settlement of MLM Under Progressively Exclusive Relationship (SUPER), Fixed and random model Circulating Probability Unification (FarmCPU), and Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK).

Figure 3.1. Methods implemented in GAPIT for GWAS and genomic selection. All the methods support GWAS, including General Linear Model (GLM), Mixed Linear Model (MLM), Compressed MLM (CMLM), Enriched CMLM (ECMLM), Settlement of MLM Under Progressively Exclusive Relationship (SUPER), Fixed and random model Circulating Probability Unification (FarmCPU), and Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK). Some of these methods support genomic selection, including MLM, CMLM, ECMLM, SUPER, and FarmCPU. The remaining (GLM and BLINK) can be used for breeding through marker assisted selection (MAS).
The similarity and difference among these methods are summarized in the figure on the right. The most simplest model is to directly detect the association between a phenotype (y) and markers (Si) one at a time, where i=1 to m, and m is number of markers. When a cofactor, such as population structure (Q) is introduced through a general linear model (GLM), the cofactor may not only account residuals (e) partially, but also adjust some effect that does not belong to the testing markers and consequently reduce false positives. The mixed linear model (MLM) applies the same principle by adding individuals’ genetic effects as random cofactor effects with variance structure defined by the kinship (K) among individuals. In both Q or Q+K models, Q and K stay the same. There are no cofactors that are adjusted by the marker tests.

The situation was changed by the multiple loci mixed model (MLMM). Through marker association tests, the associated markers are fitted as the cofactors for marker test. The cofactors are adjusted through forward and backward stepwise regression of mixed model. However, both Q and K remain unchanged.

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

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

The random effect model in FarmCPU to select associated markers using maximum likelihood method remains a high computing cost for large number of individuals. BLINK approximate the maximum
likelihood using Bayesian Information Content (BIC) in a fixed effect model to eliminate the computational burden. BLINK also eliminates the assumption that causal genes are evenly distributed across genome required by SUPER and FarmCPU methods. Consequently, BLINK is not only superior to FarmCPU for computing speed, but also for high statistical power and less false positives.

Among these methods, researches demonstrated that method on higher stars has higher statistical power for GWAS than the one on lower stars. These differences disappear in some cases suggesting interactions between the methods and other factors, including populations, environments, and traits. For examples, all methods work well for cases with extreme high heritability and major genes, or do not work for the opposite situation. In all conditions, the order of statistical powers stay the same among all the methods. The inversion of the order has not been found.

However, the interactions have been found for genomic selection. The genomic selection based on SUPER, named SUPER BLUP, has higher prediction accuracy than the genomic selection based on MLM known as genomic BLUP (gBLUP) for traits controlled with a smaller number of genes. The prediction accuracies are reversed for traits controlled by a large number of genes. The genomic selection based on CMLM, named Compressed BLUP, has higher accuracy for traits with low heritability than gBLUP.

### Table 3.1. GAPIT input parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Default</th>
<th>Options</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>model</td>
<td>Blink</td>
<td>GLM, MLM, CMLM, SUPER, MLM, FarmCPU, and Blink</td>
<td>Choose model to conduct GWAS</td>
</tr>
<tr>
<td>kinship.algorithm</td>
<td>VanRaden</td>
<td>Zhang, Loiselle and EMMA</td>
<td>Algorithm to Derive Kinship from Genotype</td>
</tr>
<tr>
<td>kinship.cluster</td>
<td>average</td>
<td>complete, ward, single, mcquitty, median, and centroid</td>
<td>Clustering algorithm to group individuals based on their kinship</td>
</tr>
<tr>
<td>kinship.group</td>
<td>Mean</td>
<td>Max, Min, and Median</td>
<td>Method to derive kinship among groups</td>
</tr>
<tr>
<td>group.from</td>
<td>1</td>
<td>User</td>
<td>The starting number of groups of iteration</td>
</tr>
<tr>
<td>group.to</td>
<td>1000000</td>
<td>User</td>
<td>The ending number of groups of iteration, if larger than number of individuals (n), n is used.</td>
</tr>
<tr>
<td>group.by</td>
<td>10</td>
<td>User</td>
<td>Iteration interval for number of groups</td>
</tr>
<tr>
<td>LD.chromosome</td>
<td>NULL</td>
<td>User</td>
<td>Chromosome for LD analysis</td>
</tr>
<tr>
<td>LD.location</td>
<td>NULL</td>
<td>User</td>
<td>Location (center) of SNPs for LD analysis</td>
</tr>
<tr>
<td>LD.range</td>
<td>NULL</td>
<td>User</td>
<td>Range around the Central Location of SNPs for LD Analysis</td>
</tr>
<tr>
<td>PCA.total</td>
<td>0</td>
<td>&gt;0</td>
<td>Total Number of PCs as Covariates</td>
</tr>
<tr>
<td>PCA.scaling</td>
<td>None</td>
<td>Scaled, Centered.and.scaled</td>
<td>Scale And Or Center And Scale The SNPs Before Conducting PCA</td>
</tr>
<tr>
<td>SNP.FDR</td>
<td>1</td>
<td>&gt;0 and &lt;1</td>
<td>Threshold to Filter SNP on FDR</td>
</tr>
<tr>
<td>SNP.MAF</td>
<td>0</td>
<td>&gt;0 and &lt;1</td>
<td>Minor Allele Frequency to Filter SNPs in GWAS Reports</td>
</tr>
<tr>
<td>SNP.effect</td>
<td>Add</td>
<td>Dom</td>
<td>Genetic Model</td>
</tr>
<tr>
<td>SNP.P3D</td>
<td>TRUE</td>
<td>FALSE</td>
<td>Logic Variable to Use P3D or Not for Testing SNPs</td>
</tr>
<tr>
<td>SNP.fraction</td>
<td>1</td>
<td>&gt;0 and &lt;1</td>
<td>Fraction of SNPs Sampled to Estimate Kinship and PCs</td>
</tr>
<tr>
<td>SNP.test</td>
<td>TRUE</td>
<td>FALSE</td>
<td>Logic Variable to Test SNPs or Not</td>
</tr>
</tbody>
</table>

### 3.1 Mixed Linear Model (MLM)

MLM includes both fixed and random effects. Including individuals as random effects gives an MLM the ability to incorporate information about relationships among individuals. This information about relationships is conveyed through the kinship (K) matrix, which is used in an MLM as the variance-covariance matrix between the individuals. When a genetic marker-based kinship matrix (K) is used jointly with population structure (commonly called the “Q” matrix, and can be obtained through
STRUCTURE\textsuperscript{11} or conducting a principal component analysis\textsuperscript{12}, the “Q+K” approach improves statistical power compared to “Q” only\textsuperscript{13}. An MLM can be described using Henderson’s matrix notation as follows:

\[
Y = X\beta + Zu + e,
\]

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

\[
\text{Var}
\begin{pmatrix}
  u \\
  e
\end{pmatrix}
= \begin{pmatrix}
  G & 0 \\
  0 & R
\end{pmatrix}
\]

(2)

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

\[
h^2 = \frac{\sigma^2_a}{\sigma^2_a + \sigma^2_e},
\]

(3)

### 3.2 Compressed MLM (CMLM)

As kinship is derived from all the markers, incorporating with the kinship for testing markers in a MLM causes the confounding between the testing markers and the individuals’ genetic effects with variance structure defined by the kinship. To reduce the confounding, individuals are replaced by their corresponding groups in the compressed MLM developed by Zhang et al in 2010\textsuperscript{14}. Cluster analysis is used to assign similar individuals into groups. The elements of the kinship matrix are used as similarity measures in the clustering analysis. Various linkage criteria (e.g., unweighted pair group method with arithmetic mean, UPGMA) can be used to group the lines together. The number of groups is specified by the user. Once the lines are assigned into groups, summary statistics of the kinship between and within groups are used as the elements of a reduced kinship matrix. This procedure is used to create a reduced kinship matrix for each compression level.

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

### 3.3 General Linear Model (GLM)

Regular MLM\textsuperscript{15} is an extreme case of CMLM where each individual is considered as a group. It can be simply performed by setting the number of groups equal to the total number of individuals, e.g. group.from = \(n\) and group.to = \(n\), where \(n\) is total number of individuals shared in both the genotype and phenotype files. Similarly, general linear model (GLM) is another extreme case of CMLM where all individuals are considered as one group. It can be simply performed by setting the number of groups equal to one, i.e. group.from = 1 and group.to = 1. GLM is the working model in PLINK\textsuperscript{16}, a primary software for studies in human genetics.
3.4 **P3D/EMMAx**

In addition to implementing compression, GAPIT uses EMMAx/P3D\textsuperscript{1,17} to reduce computing time for MLM, CMLM, ECMLM, and SUPER. If specified, the additive genetic ($\sigma^2_a$) and residual ($\sigma^2_e$) 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 **SUPER**

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

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

3.6 **Multiple Locus Mixed Linear Model (MLMM)**

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

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

3.7 **FarmCPU**

To solve the problem of false positive control and confounding between testing markers and cofactors simultaneously, an iterative method, named Fixed and random model Circulating Probability Unification (FarmCPU), was developed in 2016\textsuperscript{8}. The associated markers detected from the iterations are fitted as the cofactors to control false positives for testing the rest markers in a fixed effect model. To avoid the over model fitting problem in stepwise regression, a random effect model is used to select the associated markers using maximum likelihood method\textsuperscript{8}.

In the cycle of fixed effect model of iterations, markers are tested against the associated markers, not the confounded kinship used by MLM, CMLM, ECMLM, SUPER, and MLMM. In the cycle of random effect model of iterations, markers are selected among a small number of associated markers using maximum likelihood method to avoid the over model fitting problem in stepwise regression used by MLMM, which select marker among all available markers. Consequently, FarmCPU exhibits higher statistical power than MLMM\textsuperscript{8}. As FarmCPU tests markers in a fixed effect model, it is computational efficient than the methods that test markers in random effect model, such as MLM, CMLM, ECMLM, SUPER, and MLMM\textsuperscript{8}.

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

BLINK method was designed to have both high statistical power and computational efficiency\textsuperscript{10}. It was inspired by FarmCPU method with two major changes to achieve the objectives. One is to eliminate the assumption that causal genes are evenly distributed across genome that required by FarmCPU. As the assumption cause either inclusion of non causal genes, or missing the causal genes that are in the same bin with another causal genes with stronger signal. BLINK works directly on markers instead of bins. Markers that are in linkage disequilibrium (LD) with the most significant marker are excluded. For the second remaining marker, the exclusion is conducted in the same way as the most significant marker, so on and so forth until no marker can be excluded.

The other change is to use Bayesian Information Content (BIC) of a fixed effect model to approximate the maximum likelihood of a random effect model to select the associated markers among the markers remained the exclusion based on LD. As both the models of testing markers and selecting associated markers as cofactors are fixed effect model, the computation complexity reach the maximum. A dataset with one million individuals and one million markers can be solved in hours by using BLINK C version. The BLINK R version can be run as standard alone, or through GAPIT. To run BLINK in GAPIT, simply specify model= "Blink". The performances of the two versions were documented by the BLINK article on GigaScience\textsuperscript{10}.

3.9 Genomic BLUP

Genomic prediction is performed with the method based on genomic best linear unbiased prediction, (gBLUP)\textsuperscript{3}. The method was extended to compressed best linear unbiased prediction (cBLUP) by using the CMLM approach that was proposed for GWAS\textsuperscript{1}. The genetic potential for a group, which is derived from the BLUPs of group effects in the compressed mixed model, is used as a prediction for all individuals in the group.

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

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

\[
\mathbf{k} = \begin{bmatrix}
\mathbf{k}_{RR} & \mathbf{k}_{RI} \\
\mathbf{k}_{IR} & \mathbf{k}_{II}
\end{bmatrix},
\]

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.

\[
y_R = \mathbf{X}_R \beta + \mathbf{Z}_R \mathbf{u}_R + \mathbf{e}_R,
\]

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:

\[
\mathbf{u}_I = \mathbf{K}_{IR} \left( \mathbf{K}_{RR}^{-1} \right) \mathbf{u}_R
\]
where \( k_{IR}, k_{RR}, \) and \( u_R \) are as previously defined, and \( u_I \) is the predicted genomic values of the individuals in the inference group.

The reliability of genomic prediction is calculated as follows:

\[
\text{Reliability} = 1 - \frac{\text{PEV}}{\sigma^2_a}
\]

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

### 3.10 Compressed gBLUP

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

### 3.11 SUPER gBLUP

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

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

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

4.2 Marker density
Marker density is critical to establish Linkage Disequilibrium (LD) between markers and causal mutations. Comparison between the marker density and the LD decade over distance provides the indication if markers are dense enough to have good coverage of LD.
Figure 4.2.1 Frequency and accumulative frequency of marker density. Distribution of marker density is displayed as a histogram and an accumulative distribution.

4.3 Linkage Disequilibrium Decade
Linkage disequilibrium are measured as R square for pair wise markers and plotted against their distance. The moving average of adjacent markers were calculated by using a sliding windows with ten markers.
4.3.1 Linkage disequilibrium (LD) decade over distance. LDs were calculated on sliding windows with 100 adjacent genetic markers. Each dot represents a pair of distances between two markers on the window and their squared correlation coefficient. The red line is the moving average of the 10 adjacent markers.

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

4.5 Principal Component (PC) plot
For each PC included in the GWAS and GPS models, the observed PC values are plotted.
Figure 4.5.1 Pair-wise plots and 3D plots of principal component (PC).

4.6 Kinship plot
The kinship matrix used in GWAS and GPS is visualized through a heat map. To reduce computational burden, this graph is not made when the sample size exceeds 1,000.
Figure 4.6.1 Kinship plot. A heat map of the values in the values in the kinship matrix is created.
4.7 Neighbor-Joining (NJ)-tree
Following the compress group, we classify individuals to explain population structure. We also can plot group PCA plot with previous group.

Figure 4.7.1 Q Neighbor-Joining (NJ)-tree The dotted lines show the 95% confidence interval for the QQ-plot under the null hypothesis of no association between the SNP and the trait.

4.8 QQ-plot
The quantile-quantile (QQ) plot is a useful tool for assessing how well the model used in GWAS accounts for population structure and familial relatedness. In this plot, the negative logarithms of the $P$-values from the models fitted in GWAS are plotted against their expected value under the null hypothesis of no association with the trait. Because most of the SNPs tested are probably not associated with the trait, the majority of the points in the QQ-plot should lie on the diagonal line. Deviations from this line suggest the presence of spurious associations due to population structure and familial relatedness, and that the GWAS model does not sufficiently account for these spurious associations. It is expected that the SNPs on the upper right section of the graph deviate from the diagonal. These SNPs are most likely associated with the trait under study. By default, the QQ-plots in GAPIT show only a subset of the larger $P$-values (i.e., less significant $P$-values) to reduce the file size of the graph.
Figure 4.8.1 Quantile-quantile (QQ) –plot of $p$-values. The Y-axis is the observed negative base 10 logarithm of the $p$-values, and the X-axis is the expected observed negative base 10 logarithm of the $p$-values under the assumption that the $p$-values follow a uniform[0,1] distribution. The dotted lines show the 95% confidence interval for the QQ-plot under the null hypothesis of no association between the SNP and the trait.
4.9 Manhattan Plot

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

**Figure 4.9.1** Manhattan plot. The X-axis is the genomic position of the SNPs in the genome, and the Y-axis is the negative log base 10 of the $P$-values. Each chromosome is colored differently. SNPs with stronger associations with the trait will have a larger Y-coordinate value.
4.10 Association Table
The GWAS result table provides a detailed summary of appropriate GWAS results. The rows display the results for each SNP above the user-specified minor allele frequency threshold. The SNPs sorted by their P values (from smallest to largest).

Table 4.10.1 GWAS results for all SNPs that were analyzed.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Chromosome</th>
<th>Position (bp)</th>
<th>P-value</th>
<th>maf</th>
<th>nobs</th>
<th>(R^2) Model without SNP</th>
<th>(R^2) Model with SNP</th>
<th>FDR Adjusted P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>PZB0123.3</td>
<td>3</td>
<td>92305332</td>
<td>3.8E-05</td>
<td>0.346975086</td>
<td>281</td>
<td>0.079004463</td>
<td>0.137105003</td>
<td>0.059980668</td>
</tr>
<tr>
<td>PZB03748.1</td>
<td>2</td>
<td>7481079</td>
<td>0.000196769</td>
<td>0.195729537</td>
<td>281</td>
<td>0.079004463</td>
<td>0.132657611</td>
<td>0.193907122</td>
</tr>
<tr>
<td>PZB03734.11</td>
<td>2</td>
<td>569832711</td>
<td>0.000025959</td>
<td>0.145907473</td>
<td>281</td>
<td>0.079004468</td>
<td>0.124583793</td>
<td>0.193907122</td>
</tr>
<tr>
<td>PZB00219.6</td>
<td>3</td>
<td>219309117</td>
<td>0.000313961</td>
<td>0.379003558</td>
<td>281</td>
<td>0.079004468</td>
<td>0.123302869</td>
<td>0.193907122</td>
</tr>
</tbody>
</table>

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\(^{18}\).

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

Table 4.11.1 Information of associated SNPs.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Chromosome</th>
<th>Position</th>
<th>DF</th>
<th>t Value</th>
<th>std Error</th>
<th>effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>PZB00859.1</td>
<td>1</td>
<td>157104</td>
<td>276</td>
<td>0.255638527</td>
<td>0.501027645</td>
<td>0.128081969</td>
</tr>
<tr>
<td>PZB01271.1</td>
<td>1</td>
<td>1947984</td>
<td>276</td>
<td>-0.205390736</td>
<td>0.451124818</td>
<td>-0.092656858</td>
</tr>
<tr>
<td>PZB03613.2</td>
<td>1</td>
<td>2914066</td>
<td>276</td>
<td>-1.143780776</td>
<td>0.4887135</td>
<td>-0.558981106</td>
</tr>
<tr>
<td>PZB03613.1</td>
<td>1</td>
<td>2914171</td>
<td>276</td>
<td>1.194922452</td>
<td>0.540019499</td>
<td>0.645281424</td>
</tr>
<tr>
<td>PZB03614.2</td>
<td>1</td>
<td>2915078</td>
<td>276</td>
<td>0.277681398</td>
<td>0.489962989</td>
<td>0.136053608</td>
</tr>
</tbody>
</table>

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

4.12 Compression Profile
There are seven algorithms available to cluster individuals into groups for the compressed mixed linear model. There are also four summary statistics available for calculating the group kinship matrix. When only one group number (i.e., one dimension for the group kinship matrix) is specified, a column chart is created to illustrate the compression profile for 2*\(\log\) likelihood function (the smaller the better), genetic variance, residual variance and the estimated heritability.
Figure 4.12.1. Compression profile with single group. The X-axis on all graphs display the summary statistic method used to obtain the group kinship matrix. The rectangles with different colors indicate the clustering algorithm used to group individuals.

Note: This graph is not created when multiple groups are specified.
When a range of groups (i.e., a range of dimensions for the group kinship matrix) is specified, a different series of graphs are created. In this situation, the X-axis displays the group number. Lines with different style and colors are used to present the combinations between clustering algorithm and the algorithm to calculate kinship among groups.

**Figure 4.12.2.** Compression profile over multiple groups. The X-axis on each graph is the number of groups considered, and the Y-axes on the graphs are the -2*log likelihood function, the estimated genetic variance components, the estimated residual variance component, the estimated total variance, and the heritability estimate. Each clustering method and group kinship type is represented as a line on each graph.

*Notice:* This graph is not created when only one group is specified.
4.13 The Optimum Compression

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

Figure 4.13.1 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 (i.e., the dimension of the group kinship matrix) is 251, the value of $-2\times \log$ likelihood function is 1800.11, and the heritability is 41.6%.
4.14 Model Selection Results

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

<table>
<thead>
<tr>
<th>Number of PCs/Covariates</th>
<th>BIC (larger is better) - Schwarz 1978</th>
<th>log Likelihood Function Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-816.3884646</td>
<td>-807.9578633</td>
</tr>
<tr>
<td>1</td>
<td>-810.406992</td>
<td>-799.1998974</td>
</tr>
<tr>
<td>2</td>
<td>-796.269878</td>
<td>-782.2188759</td>
</tr>
<tr>
<td>3</td>
<td>-798.153656</td>
<td>-781.292763</td>
</tr>
</tbody>
</table>

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.

4.15 Multiple traits or methods

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

Figure 4.15.1 Multiple traits or methods Manhattan and QQ plot.
### 4.16 Genomic Prediction

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

**Table 4.16.1 Genomic Breeding values and prediction error variance.**

<table>
<thead>
<tr>
<th>Taxa</th>
<th>Group</th>
<th>RefInf</th>
<th>ID</th>
<th>BLUP</th>
<th>PEV</th>
<th>BLUE</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>33-16</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4.61889306</td>
<td>9.86652431</td>
<td>-9.2195121</td>
<td>-4.600619</td>
</tr>
<tr>
<td>38-11</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>4.81979176</td>
<td>10.1420079</td>
<td>-8.8888307</td>
<td>-4.069039</td>
</tr>
<tr>
<td>4226</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>-0.8609631</td>
<td>11.3324656</td>
<td>-8.8692075</td>
<td>-9.7301705</td>
</tr>
<tr>
<td>4722</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>5.68756264</td>
<td>9.29840109</td>
<td>-9.1363798</td>
<td>-3.4488171</td>
</tr>
<tr>
<td>A188</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>-4.404809</td>
<td>11.3810847</td>
<td>-9.1053928</td>
<td>-13.510202</td>
</tr>
<tr>
<td>A214N</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>0.15659848</td>
<td>17.6333752</td>
<td>-8.1102155</td>
<td>-7.9536171</td>
</tr>
<tr>
<td>A239</td>
<td>7</td>
<td>1</td>
<td>7</td>
<td>5.12523349</td>
<td>10.8215374</td>
<td>-8.9566662</td>
<td>-3.8314327</td>
</tr>
<tr>
<td>A272</td>
<td>8</td>
<td>1</td>
<td>8</td>
<td>-0.4981601</td>
<td>11.08823</td>
<td>-9.6011763</td>
<td>-10.099336</td>
</tr>
<tr>
<td>A441-5</td>
<td>9</td>
<td>1</td>
<td>9</td>
<td>-2.1785473</td>
<td>11.2196816</td>
<td>-9.7514142</td>
<td>-11.929961</td>
</tr>
<tr>
<td>A556</td>
<td>11</td>
<td>1</td>
<td>11</td>
<td>3.39391482</td>
<td>11.1632936</td>
<td>-9.0252992</td>
<td>-5.6313843</td>
</tr>
<tr>
<td>A6</td>
<td>12</td>
<td>1</td>
<td>12</td>
<td>3.32945838</td>
<td>13.176845</td>
<td>-10.578872</td>
<td>-7.2494132</td>
</tr>
<tr>
<td>A619</td>
<td>13</td>
<td>1</td>
<td>13</td>
<td>-1.224313</td>
<td>8.85099809</td>
<td>-8.6909708</td>
<td>-9.9152838</td>
</tr>
<tr>
<td>A632</td>
<td>14</td>
<td>1</td>
<td>14</td>
<td>-2.4776767</td>
<td>21.1823568</td>
<td>-7.5293175</td>
<td>-10.066994</td>
</tr>
</tbody>
</table>

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

**Figure 4.16.1** Interactive GS plot for gBLUP, cBLUP and sBLUP.
4.17 Distribution of BLUPs and their PEV.

A graph is provided to show the joint distribution of GBV and PEV. The correlation between them is an indicator of selection among the sampled individuals\(^{19}\).

![Graph showing joint distribution of genomic breeding value and prediction error variance.](image)

**Figure 4.17.1** Joint distribution of genomic breeding value and prediction error variance.
4.18 Interactive plot

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

Figure 4.18.1 Interactive Manhattan and QQ plot.
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\textsuperscript{20}. 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 \textit{et al.}\textsuperscript{21} and the fixed effects used to account for population structure were obtained from STRUCTURE \textsuperscript{22}.

\textit{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 method\textsuperscript{23}, 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:

```r
#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). The default is kinship.cluster="average", and kinship.group="Mean". Their expansion, the Enriched CMLM\textsuperscript{4} improves statistical power. Additionally, a specific range group numbers (i.e., dimension of the kinship matrix) can be specified. This range is controlled by the “group.from”, “group.to”, and “group.by” parameters. The analysis can be performed by typing these command lines:
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,
  KI=myKI,
  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, 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:

```r
#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:

```r
#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 "file.fragment" parameter in GAPIT. Here is an example of using "file.fragment =128".

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

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

This scenario is the same as previous scenario except changing “file.fragment” from default (512) to 128. As SNPs (minimum of two) are sampled within 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)
```

5.11 SUPER
GAPIT also implements the SUPER GWAS method\(^6\), which extracts a small subset of SNPs and uses them in FaST-LMM. This method not only retains the computational advantage of FaST-LMM, but also increases statistical power.
#Step 1: Set data directory and import files
myCV <- read.table("Copy of Q_First_Three_Principal_Components.txt", head = TRUE)
myY <- read.table("mdp_traits.txt", head = TRUE)
myG <- read.table("mdp_genotype_test.hmp.txt", head = FALSE)

#Step 2: Run GAPIT
myGAPIT_SUPER <- GAPIT(
  Y = myY[,c(1,2)],
  G = myG,
  #KI = myKI,
  CV = myCV,
  #PCA.total = 3,
  sangwich.top = "MLM", #options are GLM, MLM, CMLM, FaST and SUPER
  sangwich.bottom = "SUPER", #options are GLM, MLM, CMLM, FaST and SUPER
  LD = 0.1,
)

5.12 MLMM

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

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

5.13 Farm-CPU

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

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

5.14 Multiple model
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=c("GLM","MLM","MLMM","FarmCPU","Blink")
)
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 objects from GAPIT (please see the gallery below). Using these objects require knowledge of R. This is designed for users who are interested in: 1) developing new statistical approaches or software package using the results of GAPIT as a starting point; 2) comparing GAPIT with other new or existing statistical methods or software packages; 3) studying a specific result from GAPIT.

6.1 Preparation

#Install packages (Do this section only for new installation of R)

# References Download URL: http://cran.rstudio.com/, http://ftp.ctex.org/mirrors/CRAN/

#Step 0: Import library and GAPIT functions run this section each time to start R)

library(multtest)
library("gplots")
library("LDheatmap")
library("genetics")
library(MASS)
library("compiler")
#library(RColorBrewer)
library("scatterplot3d") #Please Download and Install
#The downloaded link at: http://cran.r-project.org/package=scatterplot3d

source("http://www.zzlab.net/GAPIT/emma.txt")
source("http://www.zzlab.net/GAPIT/gapit_functions.txt")

#source("/Users/Zhiwu/Dropbox/Current/revolution/gapit/gapit_functions.txt")

#download tutorial data and save them in myGAPIT directory under C drive and run tutorials

setwd("/myGAPIT")

6.2 Gallery of GAPIT Graphic Output

Statistical distributions of phenotype. The distributions are presented in four formats: Scatter plots, histogram, boxplot and accumulative distribution.

#Tutorial Step: Using ECMLM by Li and et. al. (BMC Biology, 2014)
(b) Plot of the first three principal components (PC1, PC2 and PC3);

(c) Manhattan plot. -log P-values are plotted against physical map position of SNPs. Chromosomes have different colors, and simulated QTNs are marked with grey horizontal dotted lines;

(d) Quantile-quantile (QQ) plot determines how GWAS results compare to the expected results under the null hypothesis of no association;

(e) Minor allele frequency (MAF) plot. -log P-values from each marker are plotted against their MAF;

(f) Type 1 error plot, which depicts the tradeoff between type I error rate (x-axis) and power (y-axis) for quantitative trait nucleotides (QTNs) with different effect size;
# Tutorial Step: Basic Scenario of Compressed MLM by Zhang and et. al. (Nature Genetics, 2010) 

---

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

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

(g-1) Power-false discovery rate (FDR) plot, which displays the power (y-axis) of GLM, MLM,CMLM,ECMLM and SUPER model in different FDR (x-axis);

(g-2) Power-false discovery rate (FDR) plot, which displays the power (y-axis) of GLM, MLM,CMLM,ECMLM,SUPER and P-LINK model in different FDR (x-axis);
# Tutorial Step(g-1): Compare to Power against FDR for GLM, MLM, CMLM, ECMLM, SUPER

# Hint: Program running time is more than 24 hours for repetition 100 times.

```r
# Step 1: Set data directory and import files
myGD <- read.table("mdp_numeric.txt", head = TRUE)
myGM <- read.table("mdp_SNP_information.txt", head = TRUE)
myKI <- read.table("KSN.txt", head = FALSE)
myG <- read.table("mdp_genotype_test.hmp.txt", head = FALSE)

# Step 2: Run GAPIT
GAPIT.Power.compare(
    myG=myG,
    myGD=myGD,
    myGM=myGM,
    myKI=myKI,
    rel=100,
    h2=0.9,
    NQTN=5
)
```

# Object: Compare to Power-FDR with different model method for the same data
# Input: myG - genotype data
# Input: myGD - The numeric SNP file (content format: 0,1 and 2)
# Input: myGM - SNP information file (content: SNP ID, chromosome, position)
# Input: myKI - Kinship data
# Input: rel - Program running time - repetition 100 times
# Input: h2 - heritability
# Input: NQTN - number of QTNs
#Tutorial Step (g-2): Compare to Power against FDR for GLM, MLM, CMLM, ECMLM, SUPER, PLINK

#Hint: Please download P-LINK Tool for different operating system, and copy it to R work path #(set setwd("/myGAPIT") location)

# References Download URL: http://pngu.mgh.harvard.edu/~purcell/plink/download.shtml

#P-LINK running have different forms for different operating systems, such windows: plink --ped #file1.ped --map file1.map --assoc

# Program running time is more than 24 hours for repetition 100 times.

Step 1: Set data directory and import files
myGD <- read.table("mdp_numeric.txt", head = TRUE)
myGM <- read.table("mdp_SNP_information.txt", head = TRUE)
myKI <- read.table("KSN.txt", head = FALSE)
myG <- read.table("mdp_genotype_test.hmp.txt", head = FALSE)

Step 2: Run GAPIT
GAPIT.Power.compare.plink(
  myG = myG,
  myGD = myGD,
  myGM = myGM,
  myKI = myKI,
  rel = 100,
  h2 = 0.9,
  NQTN = 5
)

(h) Genomic predictions vs. observed for samples of reference;

(i) Genomic predictions vs. observed for samples of inference;
# Tutorial Step: Genetic Prediction one time by cross validation
#-------------------------------------------------------------------------------------------

(j) Cross validation plot, which depicts the accuracy values of genomic prediction by splitting total samples into different number of folds.

Result Demo:

![Cross validation plot](image)

# Tutorial Step: Compare accuracy to different folds by replicate times
#-------------------------------------------------------------------------------------------

} Step 1: Set data directory and import files
y<-read.table("mdp_traits.txt", head = TRUE)
myK<-read.table("KSN.txt", head = FALSE)

# Step 2: Run GAPIT
OnePre<-GAPIT.Prediction(
  myK=myK,
y<-y[,c(1,3)],
  num=5)

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

# Step 2: Run GAPIT
myCross<-GAPIT.cross_validation.compare(
  myGD=myGD,
y=y,
  rel=100,
tc=c(2,5,10,20) ##input compare to folds num)
### 6.3 Gallery of GAPIT File Output

<table>
<thead>
<tr>
<th>Object</th>
<th>Dimension</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GWAS</td>
<td>mX7</td>
<td>GWAS results (SNPID,CHR,BP,P,MAF,DF,Effect)</td>
</tr>
<tr>
<td>Pred</td>
<td>nX9</td>
<td>GPS results (Taxa,Group,ref/inf,ID,BLUP,PEV,BLUE,Prediction,Pred_Heritable)</td>
</tr>
<tr>
<td>compression</td>
<td>kX7</td>
<td>Compression result(VA,VE,REML,h2,group,cluster,method)</td>
</tr>
<tr>
<td>kinship</td>
<td>nXn</td>
<td>kinship among individuals</td>
</tr>
<tr>
<td>kinship.optimum</td>
<td>sXs</td>
<td>The optimum group kinship</td>
</tr>
<tr>
<td>PCA</td>
<td>NA</td>
<td>Principal components</td>
</tr>
</tbody>
</table>

### 6.4 Cross validation with replacement

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

First we randomly set 25% of original phenotype (Y) as missing (NA) and generate a genomic prediction model by using their kinship. Then we record the correlation between the predicted phenotypic values and the original phenotype. We repeat this process for 1000 times. The average correlation is used as the criteria of genome prediction accuracy. The corresponding R code is displayed in the following box. The accuracy (correlation coefficient) over the 100 replicates (for demonstration purposes only; we suggest using 1000 replications) were 0.9203 and 0.6749 in the reference and inference (cross validation), respectively. The standard deviations were 0.078 and 0.0054 in the reference and inference, respectively.
R code for cross validation with replacement

```r
# 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
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$Pred

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

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

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

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)
```
6.5 Cross validation without replacement

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

```r
#Initial
nj= 200 # number of Jack Knifes, nj>0, nj!=1
Y.raw=myY[,c(1,3)] #choos a trait
Y.raw=Y.raw[!is.na(Y.raw[,2]),] #Remove missing data
n=nrow(Y.raw)
if(nj>=1){nLoop=nj}
else{
  nLoop=1/nj
}
assignment=ceiling((1:n)/(n/nLoop))
randomization=sample(1:n,n)
assignment=assignment[randomization]
NLoop=ceiling(nLoop)

#Loop on replicates
for(rep in 1:nLoop){
  #Set missing data
  if(nj>=1){Y0=Y.raw[assignment!=rep,]
  }else{
    Y0=Y.raw[assignment==rep,]
  }
  #Prediction
  myGAPIT <- GAPIT( 
    Y=Y0, 
    KI=myKI, 
    CV=myCV, 
    group.from=254, 
    group.to=254, 
    group.by=10, 
    kinship.cluster=c("average"), 
    kinship.group=c("Mean")
  )
  prediction=myGAPIT$Pred
  #Separate reference (with phenotype) and inference (without phenotype)
  if(rep==1){
    prediction.inf=prediction[prediction[,3]==2,]
  }else{
    prediction.inf=rbind(prediction.inf,prediction[prediction[,3]==2,])
  }
  #End of for (rep in 1:t)
  #Merge prediction with original Y
  YP.inf <- merge(Y.raw, prediction.inf, by.x = "Taxa", by.y = "Taxa")
  #Calculate correlation and store them
  r.inf=cor(as.numeric(as.vector(YP.inf[,2])),as.numeric(as.vector(YP.inf[,6])))
  write.table(YP.inf, "GAPIT.Jack.Knife.txt", quote = FALSE, sep = "\t", row.names = TRUE,col.names = NA)
  print(r.inf)
```

```
6.6 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.
```r
myG <- read.table("mdp_genotype_test.hmp.txt", head = FALSE)
myGAPIT <- GAPIT(G=myG, output.numerical=TRUE)
myGD = myGAPIT$GD
myGM = myGAPIT$GM
```

6.7 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://zzlab.net/GAPIT/emma.txt")
source("http://zzlab.net/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\"
```

```
nname <- 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 allelic effect estimates onto one folder
GWAS.Results  <- rbind(GWAS.Results.1, GWAS.Results.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.MultipleCorrection.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(GLMP = GWAS.Results.with.FDR[,2:4], name.of.trait = name[i], DPP=50000, plot.type = "Genomewise",cutoff=0.00)
GAPIT.Manhattan(GLMP = 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 Allelic effect estimates
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))
6.8 Genomic Prediction

We develop two new GP methods: Compress Best Linear Unbiased Prediction (CBLUP) and SUPER BLUP (SBLUP).

```r
rm(list=ls())
library('MASS')  # required for ginv
library(multtest)
library(gplots)
library(compiler)  # required for cmpfun
library("scatterplot3d")
library("EMMREML")
library(ape)

# preparing GAPIT function and data
source("http://www.zzlab.net/GAPIT/emma.txt")
source("http://www.zzlab.net/GAPIT/gapit_functions.txt")
myY=read.table(file="http://zzlab.net/GAPIT/data/mdp_traits.txt", head = TRUE)
myGD=read.table("http://zzlab.net/GAPIT/data/mdp_numeric.txt",head=T)
myGM=read.table("http://zzlab.net/GAPIT/data/mdp_SNP_information.txt",head=T)
setwd("/Users/Jiabo/Documents/gapit/test/Jiabo")
e=1  # select order of trait
m_o=2  # selet GS method, 1 is GBLUP, 2 is CBLUP, and 3 is SBLUP.
sed.seed(198521)
model=c("GBLUP","CBLUP","SBLUP")
mm=model[m_o]
y0=myY[!is.na(myY[,1+e]),c(1,1+e)]
inclusion.from=100000
  bin.from=10000
  bin.by=10000
  inclosure.to=10000
  inclosure.by=100
  sangwich.top=NULL
  sangwich.bottom=NULL
SUPER_GS=T
if(m_o==1)
  {  group.from=10000
     group.to=10000
     group.by=50}
if(m_o==2)
  {  group.from=40
     group.to=10000
     group.by=40}
if(m_o==3)
  {  group.from=10000
     group.to=10000
     inclosure.from=100
     inclosure.to=400
     group.by=100
     sangwich.top="MLM"
     sangwich.bottom="SUPER"
     SUPER_GS=T}
myGAPIT=GAPIT(
  Y=y0,
  GD=myGD,
  GM=myGM,
  group.from=group.from,group.to=group.to,group.by=group.by,
  PCA.total=3,
  SNP.test=FALSE,
  inclusion.from=inclusion.from,
  bin.from=bin.from,bin.to=bin.to,bin.by=bin.by,
  inclusion.to=inclusion.to,
  inclusion.by=inclusion.by,
  sangwich.top=sangwich.top,
  sangwich.bottom=sangwich.bottom,
  SUPER_GS=SUPER_GS,
  file.output=FALSE)
my_pred=myGAPIT$Pred
head(my_pred)
```
7 Appendix

7.1 Tutorial Data sets
The data set contains 26 files and can be downloaded at: http://zzlab.net/GAPIT/GAPIT_Tutorial_Data.zip

Table 7.1. Properties of tutorial files

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

These files 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 contains 1094 SNPs.

7.2 Typical ways of reading data

```r
# 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 GAPIT output files

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

1. How to cite GAPIT?
A: Citation may vary based on the usage of versions (the first version\(^2\), or the second version\(^5\)), and methods involved, such as the regular MLM\(^{13}\) methods, CMLM\(^1\), ECMLM\(^4\), SUPER\(^6\), and P3D\(^1\). Here is an example: “The GWAS was conducted by using the SUPER algorithm\(^6\) implemented in GAPIT R package (2.0)\(^5\)”.

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

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

4. There are many methods implemented in GAPIT, which one should I use?
A: Literature demonstrated the order of statistical power: BLINK > FarmCPU > MLMM > SUPER > ECMLM > CMLM > MLM > GLM.

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

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

7. How do I report an error?
A: In order to fix the problem, please copy and paste the error message from the R environment and attach your R source code and the dataset that allow us to repeat the error.
8. What should I do with “Error in file (file, "rt") : cannot open the connection”?  
A: In most cases this error is caused by incorrect file name or number of file specific is more than exist.

9. What should I do with “Error in GAPIT (... : unused argument(s) ...”?
A: In most cases this error is caused by incorrect spelling of GAPIT key word such as upper or lower case, or missing a comma.

10. How deal with “Error in solve.default(crossprod(X, X)) : system is computationally singular”?
A: Check covariate variables and remove the ones that are linear dependent with others.

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

12. Should I remove SNPs with MAF below 5%?
A: The answers are Yes/No. Rare SNPs with low Minor Allele Frequency (MAF) usually cause false positives, especially for small samples and traits that do not have normal distribution. However, many causal genetic variants are rare. A recommended practice is to not remove them, but interpret them with caution.

13. My trait was measured in multiple environments, how do I use them simultaneously?
A: They can be averaged across environments, and use means by GAPIT. The genetic and environmental interaction was implemented in a separated software package: GEMT (http://zzlab.net/GEMT).

14. Is it OK to analyze binary traits (case-control) with GAPIT?
A: Yes, there are many applications.

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

16. Should I use PCs or Q matrix?
A: Keyan Zhao and et al (PLoS Genetics, 2007) compared the two methods and demonstrated that they had similar statistical power.
### 7.5 GAPIT Biography

<table>
<thead>
<tr>
<th>Date</th>
<th>Version</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>11-May-11</td>
<td>1.2</td>
<td>First public release with following method implemented:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Principal component method to encounter population structure (Price).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Unified mixed model to encounter both population structure and kinship.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• EMMMA method to improve the speed to estimate variance components (ratio).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Compressed mixed model to improve statistical power and speed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• P3D or EMMAx to improve speed by estimating population parameters (e.g. variances and grouping) only once.</td>
</tr>
<tr>
<td>13-Jun-11</td>
<td>1.22</td>
<td>Genotype in numerical format in addition to hapmap format</td>
</tr>
<tr>
<td>2-Sep-11</td>
<td>1.31</td>
<td>Reading fragment within single genotype file to save memory</td>
</tr>
<tr>
<td>17-Sep-11</td>
<td>1.36</td>
<td>Interface change for prototyping</td>
</tr>
<tr>
<td>24-Oct-11</td>
<td>1.41</td>
<td>Option to impute missing genotypes as middle, major, minor, present/absent.</td>
</tr>
<tr>
<td>1-Nov-11</td>
<td>1.42</td>
<td>Matrix partitioning to improve speed (5-10 fold faster)</td>
</tr>
<tr>
<td>7-Dec-11</td>
<td>2.01</td>
<td>FaST-LMM method implemented</td>
</tr>
<tr>
<td>19-May-12</td>
<td>2.19</td>
<td>Splitting big genotype file into small ones</td>
</tr>
<tr>
<td>8-Jul-12</td>
<td>NA</td>
<td>GAPIT Bioinformatics paper accepted for publication</td>
</tr>
<tr>
<td>8-Nov-12</td>
<td>2.2</td>
<td>Automatic sorting taxa in kinship for regular MLM</td>
</tr>
<tr>
<td>14-Feb-13</td>
<td>2.25</td>
<td>Confidence interval on QQ plot</td>
</tr>
<tr>
<td>26-Apr-13</td>
<td>2.26</td>
<td>Labeling QTNs on Manhattan plot</td>
</tr>
<tr>
<td>15-Jul-14</td>
<td>2.27</td>
<td>SUPER implemented</td>
</tr>
<tr>
<td>20-Aug-14</td>
<td>2.28</td>
<td>ECMLM implemented</td>
</tr>
<tr>
<td>25-Oct-14</td>
<td>2.29</td>
<td>VanRaden kinship algorithm with centralization</td>
</tr>
<tr>
<td>28-Feb-15</td>
<td>2.3.41</td>
<td>Enrichment on output figures and tables</td>
</tr>
<tr>
<td>5-Sep-15</td>
<td>by date</td>
<td>Uniform output across models (Version 3 initiation)</td>
</tr>
<tr>
<td>12-Oct-15</td>
<td></td>
<td>Simulation of category phenotype</td>
</tr>
<tr>
<td>22-Oct-15</td>
<td></td>
<td>Compression BLUP (cBLUP) for genomic prediction</td>
</tr>
<tr>
<td>1-Apr-16</td>
<td></td>
<td>GAPIT version 2 paper published by Plant Genome</td>
</tr>
<tr>
<td>4-Apr-16</td>
<td></td>
<td>Pedigree-like marker based kinship</td>
</tr>
<tr>
<td>4-Apr-16</td>
<td></td>
<td>SUPER BLUP (sBLUP) for genomic prediction</td>
</tr>
<tr>
<td>31-Mar-17</td>
<td></td>
<td>Indicating LD to the strongest associated marker above threshold</td>
</tr>
<tr>
<td>2-May-18</td>
<td></td>
<td>Add cBLUP and sBLUP demo script into user manual</td>
</tr>
<tr>
<td>May 15,2018</td>
<td></td>
<td>Fixed the error with different SNP.fraction</td>
</tr>
<tr>
<td>23-May-18</td>
<td></td>
<td>Fixed the error with SNP.MAF</td>
</tr>
<tr>
<td>29-May,2018</td>
<td></td>
<td>Fixed errors in converting hmp into numeric</td>
</tr>
<tr>
<td>Jul 17,2019</td>
<td></td>
<td>Add command for loading GAPIT from GitHub</td>
</tr>
</tbody>
</table>
REFERENCES