

User Manual for



Association Leveraging Principal  
Component Analysis

By Haixiao Dong and Ryan Oliveira

**Table of contents**

Summary/Getting Started	2
Inputs	3,4,5
Outputs	5
Examples	6,7,8

## Summary

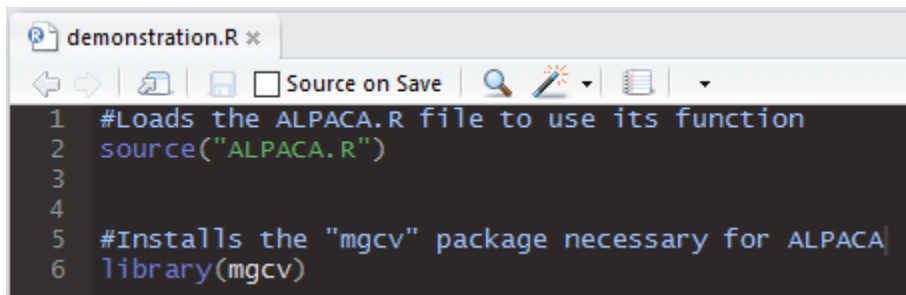
The results of genome-wide association studies (GWAS) can often be confounded by population stratification in the study population<sup>1</sup>. To address this issue, we developed a package for the statistical software R that minimizes this issue with principal components analysis (PCA), named **A**ssociation **L**everaging **P**rin**C**ipal **C**omponent **A**nalysis (ALPACA). ALPACA can better eliminate false positives by picking up on associations within the data using PCA. ALPACA also allows to let the user put in their own covariates and will eliminate its generated principal components if any are found to be linearly dependent on the covariates. Finally, as PCA does not take a great deal of processing power, the average run time is still very quick (under 3 seconds for ~300 individuals and ~3000 SNPs). Three principal components are used by default.

This software uses a general linear model in which covariates and PCA are used as fixed effects. Accordingly, this software is inappropriate for models that use random effects.

## Getting started

As ALPACA is an R package, it requires R to be installed. From the R screen, use the “source” function to load the ALPACA.R file. Note: make sure ALPACA.R is in the working directory or that its directory is specified.

Next, install the “mgcv” package using “library”. This package is necessary for ALPACA to work.



```
demonstration.R x
Source on Save
1 #Loads the ALPACA.R file to use its function
2 source("ALPACA.R")
3
4
5 #Installs the "mgcv" package necessary for ALPACA
6 library(mgcv)
```

You are now ready to enter your data (see next page, “Inputs”).

(1) Price, Alkes L., et al. "Principal components analysis corrects for stratification in genome-wide association studies." *Nature genetics* 38.8 (2006): 904-909.

## Inputs (required)

The following data are required as inputs:

**SNP data:** The SNP data should be a tab-delimited text file where each row corresponds to an individual and each column corresponds to a marker value from a set of markers. The marker values should be designated “0”, “1”, and “2” corresponding to homozygotes of allele A, heterozygotes, and homozygotes of allele B, respectively. Below is an example in which the first column has not been deleted.

	taxa	PZB00859.f	PZA01271.f	PZA03613.2	PZA03613.f	PZA03614.2	PZA03614.f	PZA00258.3
1	33-16	2	0	0	2	2	2	2
2	38-11	2	2	0	2	2	2	0
3	4226	2	0	0	2	2	2	0
4	4722	2	2	0	2	2	2	1
5	A188	0	0	0	2	2	2	0
6	A214N	2	0	2	0	2	0	0
7	A239	0	0	2	2	0	0	0
8	A272	0	0	2	2	0	0	2
9	A441-5	2	0	0	2	2	2	0
10	A554	2	2	2	2	0	2	0

**Phenotypes:** This file should contain the phenotype of interest for each individual. As before, the first column should be removed either in the text file or as the file is loaded in R.

	Taxa	Obs
1	33-16	-1.273103730
2	38-11	0.551527097
3	4226	-0.254927311
4	4722	-6.122964349
5	A188	-2.593982541
6	A214N	-1.851128061
7	A239	2.692522244
8	A272	0.388128049
9	A441-5	-0.401666472
10	A554	-3.307642193

**Chromosome map:** The chromosome map file should contain the SNP ID as the first column, the chromosome number from each SNP as the second column, and the position as the third. *No columns need to be removed from this data.*

	SNP	Chromosome	Position
1	PZB00859.1	1	157104
2	PZA01271.1	1	1947984
3	PZA03613.2	1	2914066
4	PZA03613.1	1	2914171
5	PZA03614.2	1	2915078
6	PZA03614.1	1	2915242
7	PZA00258.3	1	2973508
8	PZA02962.13	1	3205252

## Inputs (optional)

The following inputs are optional:

**Covariates:** The input covariates should be a tab-delimited text file in which the number of rows match that in the original data file. Like before, if the first column is individual IDs, this column must be deleted manually in the text file or removed via R, e.g. via “X=myGD[,-1]”. If no covariates are put in, only the package’s principal components will be used as a fixed effect. An example (where the individual ID column has not been deleted) is below.

	Taxa	FactorA	FactorB
1	33-16	2.53133090	5.5014640
2	38-11	2.63386050	4.6556906
3	4226	1.89069549	6.1368827
4	4722	1.85603497	7.8418581
5	A188	2.55262921	5.4094501
6	A214N	5.43681508	-5.8162307
7	A239	2.02994682	5.1653600
8	A272	6.37522031	6.1192020
9	A441-5	4.92417265	5.5475456
10	A554	2.26038899	5.5215246

**Cutoff:** A cutoff threshold (“cutoff”) can be specified as the exact number in R, e.g. “ $10^{-6}$ ”. Otherwise, a simple Bonferroni correction using the number of SNPs is applied to an alpha of 0.05.

**Set quantitative trait nucleotides:** A vector of positions of known quantitative trait nucleotides (“QTN.position”) can be input for testing purposes.

The phenotypes are read into R as “y” (lower case), SNP data as “X” (upper case), covariates as “C”, the chromosome map as “GM”, the p-value cutoff as “cutoff”, and the QTN position vector as “QTN.position”. In the below example, the appropriate files are loaded with the first column removed using R for phenotypes, SNP data, and the chromosome map.

```

8 #Import data
9 myGD=read.table(file="http://zzlab.net/GAPIT/data/mdp_numeric.txt",head=T)
10 myGM=read.table(file="http://zzlab.net/GAPIT/data/mdp_SNP_information.txt",head=T)
11 myY=read.table(file="http://zzlab.net/GAPIT/data/CROP545_Phenotype.txt", head=T)
12 myCV=read.table(file="http://zzlab.net/GAPIT/data/CROP545_Covariates.txt", head=T)
13
14 myGLM <- ALPACA(
15   y=myY[, -1],
16   X=myGD[, -1],
17   C=myCV[, -1],
18   GM=myGM,
19   cutoff=10^-6,
20   QTN.position=c(687,1060,320,1927,992,698,587,92,204,1306)
21 )

```

## Outputs

The following outputs are given:

**myGLM:** This string, accessible as the variable “myGLM”, contains a list of the obtained p-values, the final cutoff used for significance, a list of significant SNPs as determined by the cutoff, the p-values of the significant SNPs, a list of SNPs ordered by p-value, and a list containing power, false discovery rate, and type I error (if known QTNs are put in).

**Images:** A QQ plot, Manhattan plot, a principal component plot, and (if QTNs are put in) an FDR vs. power plot are all generated as PDFs.

**myGLM.results.csv:** A comma-separated values (CSV) file is stored that contains a list of all tested SNPs, their chromosome number and position, their p-values, and their overall rank (where 1 is the lowest p-value).

These outputs are explored in the examples below.

## Examples

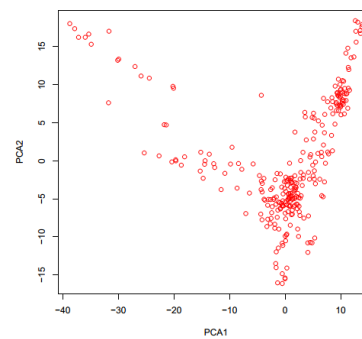
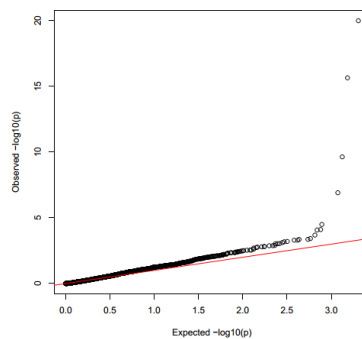
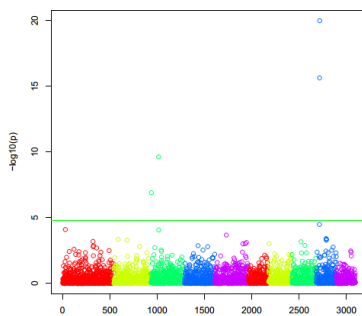
The following examples are contained in the ALPACAExamples.R file and use data files from ZZLab.net.

### Example 1 — GWAS by GLM with PCA as default

The first example uses sample data without covariates. As such, PCA are used as the default fixed effect. The left image is a Manhattan plot where the green line represents the calculated p-value cutoff (approximately  $10^{-5}$  in this example). The middle image is the QQ plot. The right image graphs the first and second principal components used against each other.

```
[1] "Welcome to using ALPACA!"
[1] "PCA[, 1:3] used as default, plots and table generated!"
[1] "Manhattan plot generated!"
[1] "QQ plot generated!"
[1] "ALPACA successfully finished!"
[1] "Use str(myGLM) to check the results, and a list of p-values by SNP is available at myGLM.results.csv"
```

```
> str(myGLM)
List of 6
 $ P.value      : num [1, 1:3093] 0.943 0.432 0.782 0.585 0.961 ...
 $ cutoff.final : num 1.62e-05
 $ sig.SNP      : int [1:4] 2717 2715 1013 937
 $ sig.SNP.P    : num [1:4] 0.00 2.22e-16 2.29e-10 1.23e-07
 $ order.SNP   : int [1:3093] 2717 2715 1013 937 2714 29 1015 1730 2784 2788 ...
 $ power.fdr.type1error: NULL
```



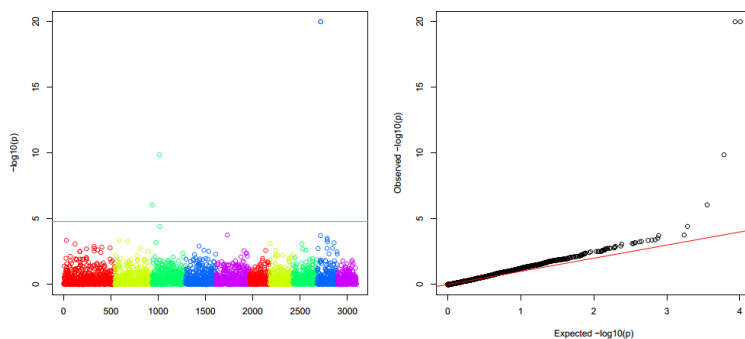
## Example 2 — GWAS by GLM with user provided covariates

The second example uses user-provided covariates. If you run this example, you will notice the program notifies you that the generated principal components are linearly dependent on the provided covariates, and as such they are eliminated.

```
[1] "Welcome to using ALPACA!"
[1] "Linear dependence exists, own co-factors (PCA) eliminated!"
[1] "Manhattan plot generated!"
[1] "QQ plot generated!"
[1] "ALPACA successfully finished!"
[1] "Use str(myGLM) to check the results, and a list of p-values by SNP is available at myGLM.results.csv"
```

As shown in Example 1, the Manhattan plot and QQ plot are successfully generated. No principal components plot is generated as the values are discarded due to being linearly dependent on the covariates.

```
> str(myGLM)
List of 6
 $ P.value          : num [1, 1:3093] 0.929 0.766 0.821 0.926 0.812 ...
 $ cutoff.final     : num 1.62e-05
 $ sig.SNP          : int [1:4] 2715 2717 1013 937
 $ sig.SNP.P        : num [1:4] 0.00 0.00 1.35e-10 8.69e-07
 $ order.SNP        : int [1:3093] 2715 2717 1013 937 1015 1730 2714 2784 2786 587 ...
 $ power.fdr.type1error: NULL
```



## Example 3 — GWAS by GLM for simulation

The third example uses pre-determined QTNs and simulated phenotypes, allowing the “QTN.position” variable to be used. When the function is used for simulations, the pre-determined QTNs are marked specially on the Manhattan plot. In addition, Power, FDR and Type I error were calculated.

```
[1] "Welcome to using ALPACA!"
[1] "PCA[, 1:3] used as default, plots and table generated!"
[1] "Manhattan plot generated!"
[1] "QQ plot generated!"
[1] "ALPACA successfully finished!"
[1] "Use str(myGLM) to check the results, and a list of p-values by SNP is available at myGLM.results.csv"
```

```
> str(myGLM)
List of 6
 $ P.value          : num [1, 1:3093] 0.42 0.565 0.679 0.645 0.541 ...
 $ cutoff.final     : num 1.62e-05
 $ sig.SNP          : int [1:4] 1679 1058 1480 1193
 $ sig.SNP.P        : num [1:4] 0.00 1.84e-08 5.34e-07 1.22e-05
 $ order.SNP        : int [1:3093] 1679 1058 1480 1193 1861 1226 1496 1890 745 2116 ...
 $ power.fdr.type1error:List of 3
  ..$ power        : num [1:3093] 0.1 0.2 0.3 0.3 0.3 0.3 0.3 0.3 0.4 0.4 ...
  ..$ fdr           : num [1:3093] 0 0 0 0.25 0.4 ...
  ..$ type1error    : num [1:3093] 0 0 0 0.000323 0.000647 ...
```

