Association Mapping with rrBLUP 4

Jeffrey Endelman

June 15, 2013

Prior to version 4.1 of the package, the function for association mapping was called GWA. This function has been completely overhauled and given a new interface, and the new function is called GWAS. To illustrate its use, we will do association mapping with the Arabidopsis dataset published by Atwell et al. (2010). This population consisted of 199 inbred lines genotyped with 250K SNPs and phenotyped for 107 traits.

The genotypes were downloaded from
https://cynin.gmi.oeaw.ac.at/home/resources/atpolydb/250k-snp-data/call_method_32.tar.gz

Upon extracting this archive, the file with the genotypes is "call_method_32.b". The following code reads this file and converts the genotype calls to the {-1,0,1,NA} format required for rrBLUP:

```
> markers <-
read.csv("call_method_32.b",skip=1,header=T,as.is=T,check.names=FALSE)
> convert.snp <- function(x) {
+ #convert to {-1,0,1,NA}
+ alleles <- na.omit(unique(x))
+ y <- rep(NA,length(x))
+ y[which(x==alleles[1])] <- -1
+ y[which(x==alleles[2])] <- 1
+ return(y)
+ }
> M <- apply(markers[,-(1:2)],1,convert.snp)
> dim(M)
[1]    199 216130
```

The last statement shows the dimensions of the marker matrix M: 199 lines x 216,130 markers. The following code sets the rownames of M equal to the genotype identifiers (gid):

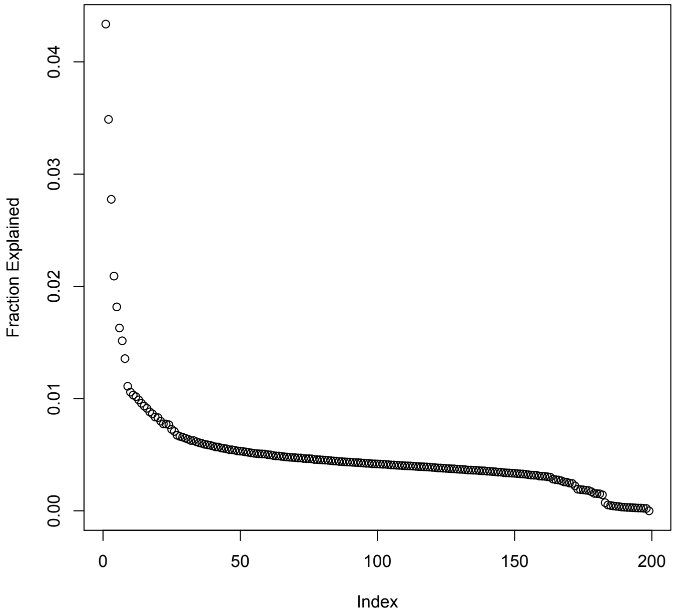
```
> gid <- colnames(markers)[-(1:2)]
> rownames(M) <- gid
> n <- nrow(M)  # number of lines
> m <- ncol(M)  # number of markers
```

Now we will load the rrBLUP package and use its function A.mat to calculate the marker-based relationship matrix:

```
> library(rrBLUP)
> A <- A.mat(M)
> dim(A)
[1] 199 199
```

To look for signs of population structure, we will do a principal components (PC) analysis by eigenvalue decomposition of A:

```
> eig.result <- eigen(A)
> lambda <- eig.result$values
> plot(lambda/sum(lambda),ylab="Fraction Explained")
```



The figure shows that the first PC accounts for less than 5% of the total spectrum. This lack of population stratification means the K mixed model for association mapping should be adequate for controlling population structure: we do not need to add additional PCs as covariates (see the reference manual for instructions on how to include PCs in GWAS).

The GWAS function expects the first three columns of the genotype matrix to be the marker name, chromosome, and position, respectively. The format is illustrated below:

```
> geno <- cbind(1:m,markers[,1:2],t(M))
> colnames(geno) <- c("marker","chrom","pos",gid)
> geno[1:10,1:10]
   marker chrom  pos 6909 6977 100000 6906 8266 6897 6898
1       1     1  657   -1   -1     -1   -1    1   -1    1
2       2     1 3102   -1   -1      1   -1    1   -1    1
3       3     1 4648   -1   -1     -1   -1    1   -1    1
4       4     1 4880   -1   -1     -1   -1   -1   -1   -1
5       5     1 5975   -1   -1     -1   -1   -1   -1   -1
6       6     1 6063   -1    1     -1    1    1    1    1
7       7     1 6449   -1    1     -1   -1   -1   -1   -1
8       8     1 6514   -1   -1     -1   -1    1   -1    1
9       9     1 6603   -1   -1     -1   -1    1   -1    1
10     10     1 6768   -1   -1      1   -1   -1   -1   -1
```

The names of the columns with the marker calls must match the labels in the first column of the phenotype file. The order in the phenotype and genotype matrices does not have to be the same, but the gid labels must match as strings.

The phenotype file for this dataset was downloaded from
https://cynin.gmi.oeaw.ac.at/home/resources/atpolydb/miscellaneous-data/phenotype_published_raw.tsv

The following code reads this file and extracts four phenotypes (FT_LD = flowering time under long days, Dormancy = seed dormancy, avrRpm = disease resistance, FRI = FRI gene expression):

```
> pheno <-
read.table("phenotype_published_raw.tsv",header=T,as.is=T,check.names=FALSE,
sep="\t")
> pheno2 <- pheno[,c(1,3,10,35,43)]
> colnames(pheno2) <- c("ecoid","FT_LD","Dormancy","avrRpm","FRI")
> head(pheno2)
  ecoid    FT_LD Dormancy avrRpm    FRI
1  5837  41.8750       34      1 2.4135
2  6008  26.0000       NA     NA 0.2430
3  6009 200.0000        2      1 1.3987
4  6016  83.0417       NA     NA 1.8900
5  6040 200.0000       NA     NA 1.9010
6  6042  29.1667       NA     NA 0.4250
```

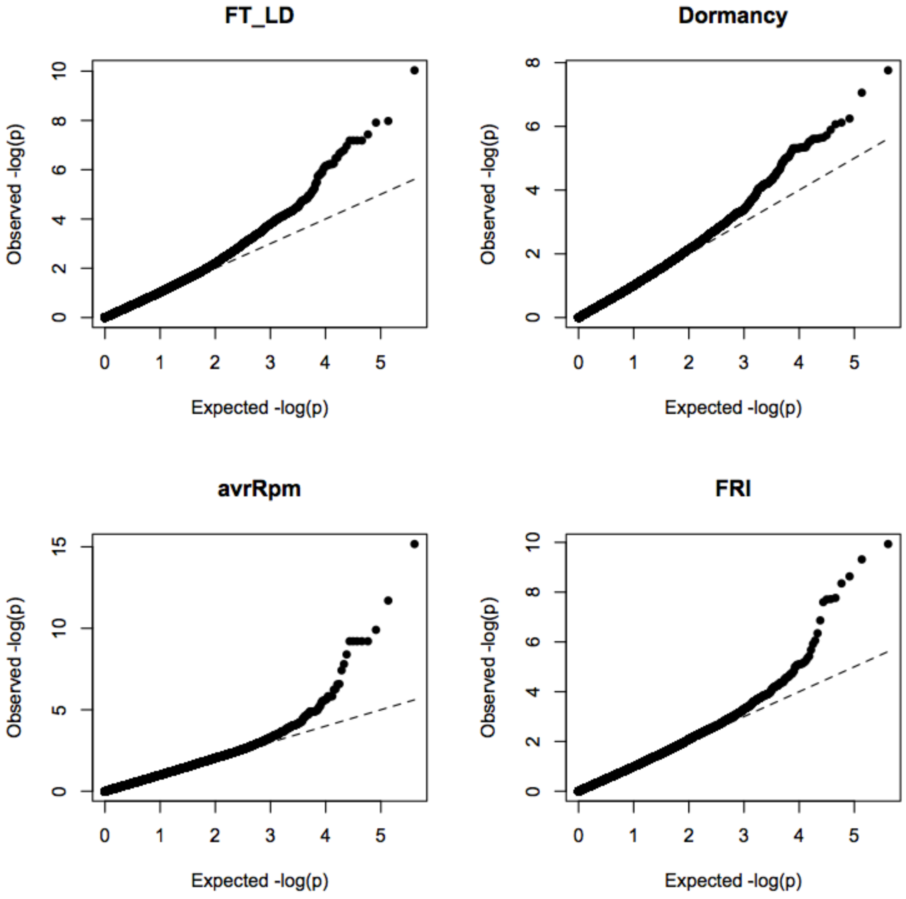
The first column in the phenotype matrix is the gid, and subsequent columns are phenotypes. Notice that for the seed dormancy and avrRpm (disease resistance) traits, not all of the lines have been phenotyped. These missing observations are removed by the GWAS function before analysis.

The GWAS function has the ability to do both the original EMMA (Kang et al. 2008) and the faster but approximate EMMAX/P^{3}D (Kang et al. 2010; Zhang et al. 2010) methods. In the latter, the variance components are estimated only once for the base model without SNPs, which can lead to mild underestimation of significance in some situations (Zhou and Stephens 2012). The following code runs GWAS with the P^{3}D approximation:

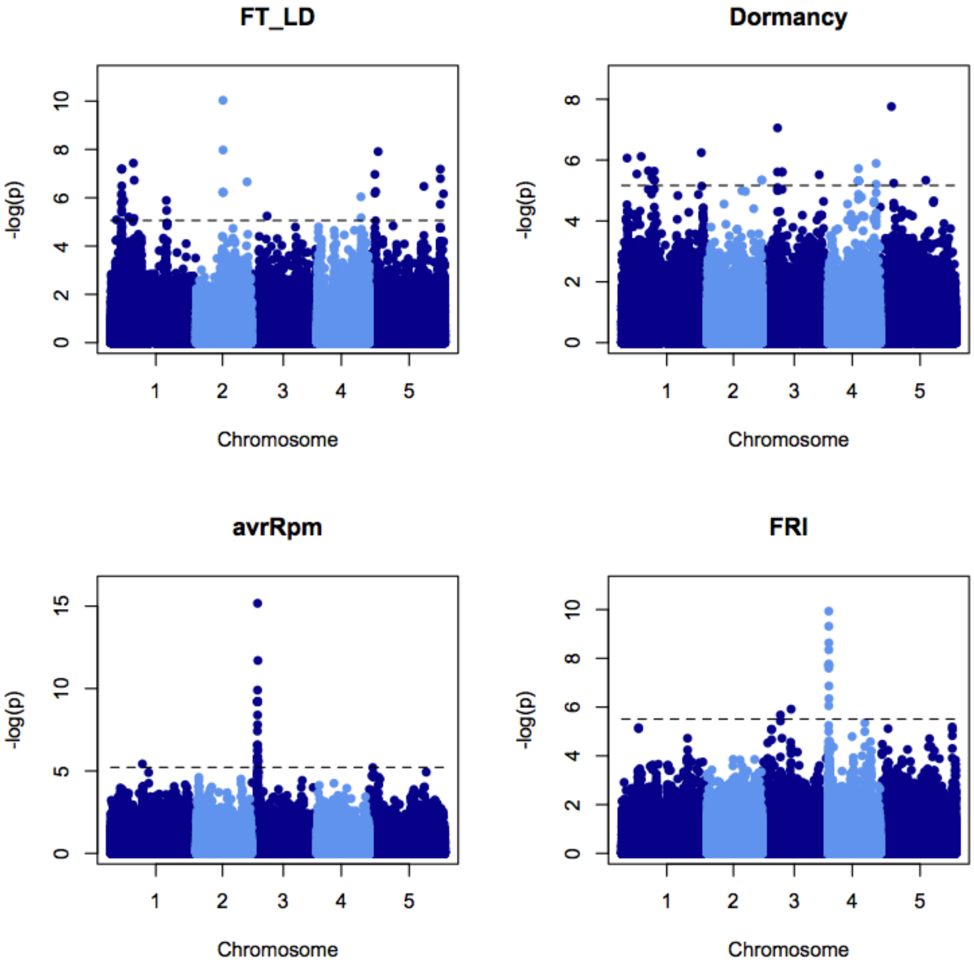
```
> ans.emmax <- GWAS(pheno=pheno2,geno=geno,P3D=TRUE,n.core=16,K=A)
[1] "GWAS for trait: FT_LD"
[1] "Variance components estimated. Testing markers."
[1] "GWAS for trait: Dormancy"
[1] "Variance components estimated. Testing markers."
[1] "GWAS for trait: avrRpm"
[1] "Variance components estimated. Testing markers."
[1] "GWAS for trait: FRI"
[1] "Variance components estimated. Testing markers."
```

By using the option n.core=16, the markers were partitioned into 16 groups and analyzed in parallel on 16 cores in my computer (this requires a UNIX-compliant system, like Linux or MacOS; the default is n.core=1). Because we had already computed the kinship (K = A) matrix for analyzing population structure, I passed this matrix to GWAS to save the time of recomputing it, but this is optional. You can also pass other types of kinship matrices, or you can pass nothing, in which case GWAS uses A.mat to calculate K from the markers.

The function GWAS returns the –$\log_{10}$(p) scores for the traits, and by default it also makes qq and Manhattan plots (turn this off with plot=FALSE):



qq-plots



Manhattan plots

The dashed line in the qq plot shows the expectation under the null hypothesis. When population structure is properly controlled, the low scores should follow this line. In the Manhattan plots, the dashed line shows the p-value corresponding to a false discovery rate (FDR) of 0.05. For more information on the GWAS function, type help("GWAS").

References

Atwell et al. 2010. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465:627–631.

Kang et al. 2008. Efficient control of population structure in model organism association mapping. Genetics 178:1709–1723.

Kang et al. 2010. Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. 42:348–354.

Zhang et al. 2010. Mixed linear model approach adapted for genome-wide association studies. Nat. Genet. 42:355–360.

Zhou, X., and M. Stephens. 2012. Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. 44:821–824.