GWAS_Project

Zhi Li

10/11/2017

Introduction

Genome-wide association study (GWAS) is an examination of genome-wide set of genetic variants in different individuals to find the signal of associations between genetic variants and traits in samples from population (Visscher et al., 2017). With the dramatically decreasing genotyping and sequencing costs, GWAS are becoming the major method for studying the genetics of natural variation. GWAS has been widely used in human genetics and has made a great progress on identifying new disease loci that were previously unknown (Zhu et al., 2008). GWAS also has been applied to a range of model organisms including Arabidopsis and mouse and to non-model systems including crops and cattle (Korte and Farlow, 2013).

Nowadays, GWAS use high-throughput genotyping technologies to genotype single-nucleotide polymorphisms (SNPs) and relate them with quantitative traits. GWAS analysis depends on the linkage disequilibrium (LD), which can measure the degree of non-random association between alleles at different loci. A squared correlation (r2) has been commonly used to quantify LD. The typical GWAS has four parts (Chandra et al., 2014): (1) selection of a large number of individuals with the disease or trait of interest and, in a binary condition, a suitable comparison group; (2) DNA isolation, genotyping, and quality control; (3) statistical tests for associations between the SNPs and the disease/trait; (4) replication of identified associations in independent population samples or experimental examination of functional implications.Several factors can influence the success of GWAS analysis, including the number of loci affecting the trait segregate in a population, sample size, the panel of genome wide variants, and the joint distribution of effect size and allele frequency at those loci and the LD between observed genotyped DNA variants and the unknown causal variants (Visscher et al., 2017). Generally, GWAS with high density SNP coverage, large sample size, and minimum population structure, can have high chance in dissection of a complex trait.

Both GWAS and Linkage mapping rely on the same principle: coinheritance of functional polymorphisms and adjacent DNA variants. Compared to linkage mapping, GWAS have several advantages. For example, it can reduce research time by finding genetic association without the costly and time-consuming construction of large experiment populations. It can have the potentially high genetic resolution provided by the many meiotic events that occurred during past generations. In addition, it has the possibility of surveying many functionally diverse alleles (Zhu et al., 2008). However, GWAS also introduce several other drawbacks as a trade-off, such as population stratification, false positive association, and gene-environmental interaction. For example, genetic structure can cause false positive associations. Population structure must either be eliminated through between study design or controlled in the analysis. Mixed model is a power way to handle population structure by accounting for the amount of phenotypic covariance that is due to genetic relatedness (Vilhjálmsson and Nordborg, 2013). Several methods have been developed to account for population structure, such as Principal components analysis(PCA), Genomic control (GC) and Structured association (SA).

GWAS Analysis

Material used. Describe the population and markers used. Add a discussion considering the differences compared to our last project (QTL Mapping). Include comments about the power and resolution of the analysis. (2 pts)

```
library(BGLR)
data(wheat)
dim(wheat.X) ## 599 individuals and 1279 markers
```

```
## [1]  599 1279
```

```
head(wheat.Y) ## phentypic data(individual X environments)
```

```
##               1           2           4          5
## 775   1.6716295 -1.72746986 -1.89028479  0.0509159
## 2166 -0.2527028  0.40952243  0.30938553 -1.7387588
## 2167  0.3418151 -0.64862633 -0.79955921 -1.0535691
## 2465  0.7854395  0.09394919  0.57046773  0.5517574
## 3881  0.9983176 -0.28248062  1.61868192 -0.1142848
## 3889  2.3360969  0.62647587  0.07353311  0.7195856
```

```
wheat.X[1:5,1:5] ## genotypic data (individual X markers)
```

```
##      wPt.0538 wPt.8463 wPt.6348 wPt.9992 wPt.2838
## [1,]        0        1        1        1        1
## [2,]        1        1        1        1        1
## [3,]        1        1        1        1        1
## [4,]        0        1        1        1        1
## [5,]        0        1        1        1        1
```

```
wheat.A[1:5,1:5] ## pedigree (individual X individual)
```

```
##         775   2166   2167   2465   3881
## 775  1.9698 0.5742 0.5742 0.7236 1.0578
## 2166 0.5742 1.9930 1.9930 0.6724 0.4116
## 2167 0.5742 1.9930 1.9966 0.6724 0.4116
## 2465 0.7236 0.6724 0.6724 1.9970 0.5090
## 3881 1.0578 0.4116 0.4116 0.5090 1.9928
```

The material we used from a collection of 599 historical CIMMYT wheat lines with 1279 markers. The phenotypic trait evaluated here were the average grain yiled of the 599 wheat lines evaluted in each of these four mega- enviroments. In this project,GWAS analysis was tested under environment 1.

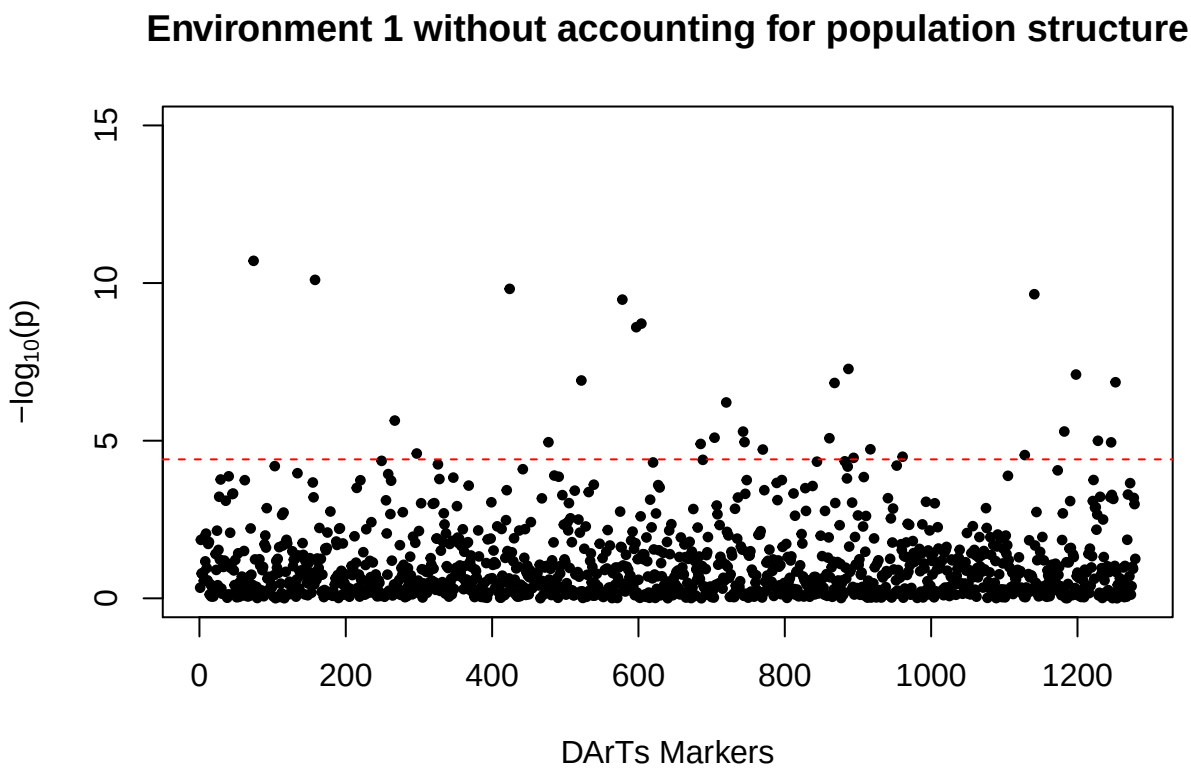
There are only a few recombination event occur within families and pedigrees in QTL mapping and fewer alleles were sampled (2-4 max.), resulting a relative low mapping resolution (Zhu et al., 2008). For instance, linkage analysis generally can localize QTL to 10-20 cM intervals in plants(Price, 2006). In addition, QTL mapping need to generate and evaluate a large number of line, which is time consuming. While GWAS overcome the two main limitations of QTL analysis mentioned above. It can increase mapping resolution by using historical recombination and exploiting natural genetic diversity. It can also reduce research time and sample great number of allele by using the panel of lines/ wild mapping population. In addition, QTL mapping used bi-parental mapping population, which can only provide pertinent information about traits that tends to be specific to same or genetically related populations, while results from GWAS are more applicable to a much wide germplasms base (Zhu et al., 2008). In QTL mapping, few traits segregate and few markers required. While in GWAS, all traits segregate and many markers required. Therefore, GWAS can provide a potential good QTL-to- gene resolution with low dection power while QTL analysis can provide poor (no) QTL-to-gene resolution with high detection power.

Perform the GWAS analysis considering single marker analysis with and without accounting for population structure. Use graphics to represent your results. (2 pts)

```
library(BGLR)
data(wheat)

## Single Marker regression -Environemt 1
X<-scale(wheat.X) ## centering data with mean 0 and variation 1
Y=wheat.Y[,1]## use the 1st environment data
p.value=vector()
```

```
for(i in 1:ncol(X)){
  regression=(lm(Y~X[,i]))
  p.value[i]=summary(regression)$coef[2,4]
}
## Bonferroni correction
bonferroni=0.05/ncol(X)
## Manhattan plot without accounting for population structure.
plot(-log(p.value,base=10),pch=19,cex=0.6,
     ylab=expression(paste("-log"[10],"(p)", sep="")),
     xlab="DArTs Markers",
     main="Environment 1 without accounting for population structure",
     ylim=c(0,15))
abline(h=-log(bonferroni,base=10),col="red",lty=2)
```

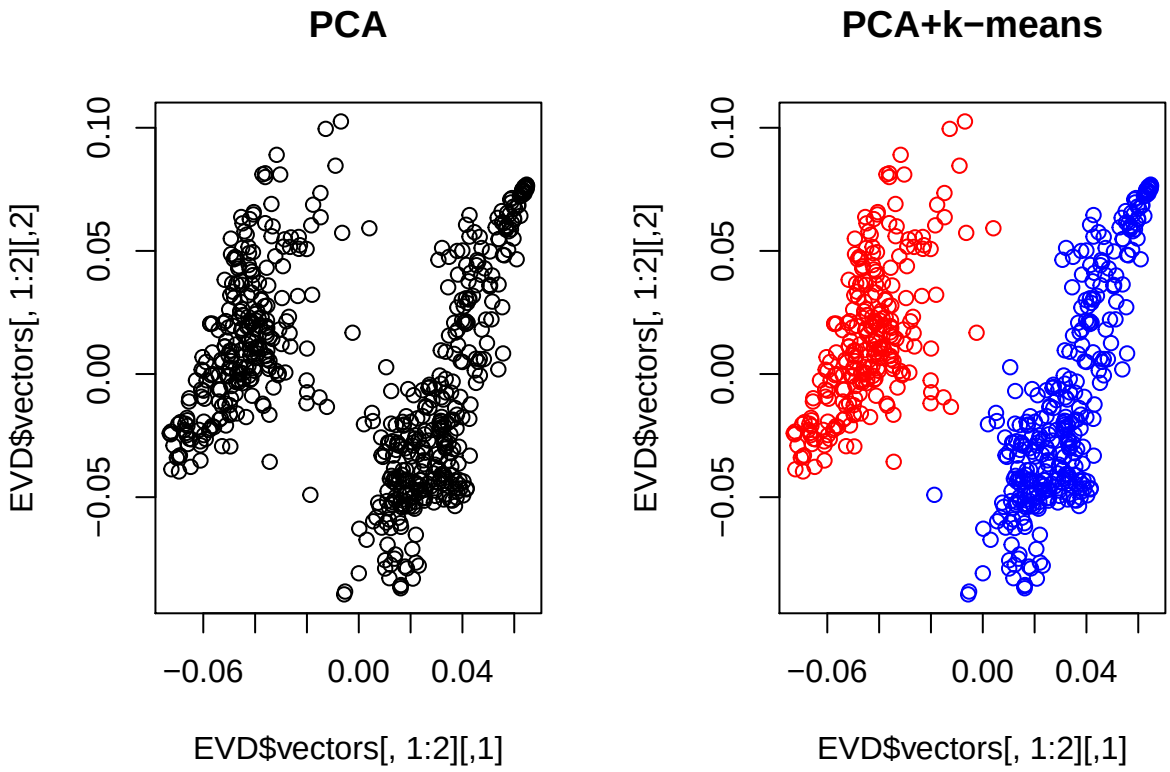



```
which(p.value<bonferroni)
```

```
##  [1]   74  158  267  297  424  477  522  578  597  604  685  704  720  743
## [15]  745  770  861  868  887  894  917  961 1128 1141 1182 1198 1228 1246
## [29] 1252
```

```
# Original data set
X<-scale(wheat.X)
G<-tcrossprod(X)
# PCA
par(mfrow=c(1,2))
EVD<-eigen(G)
plot(EVD$vectors[,1:2], main="PCA")
#k-means(2 groups)
set.seed(1)
new_df=cbind(EVD$vectors[,1], EVD$vectors[,2])
```

```
fit=kmeans(new_df,2)
color=ifelse(fit$cluster==1, "blue","red")
plot(EVD$vectors[,1:2],col=color,main="PCA+k-means")
```

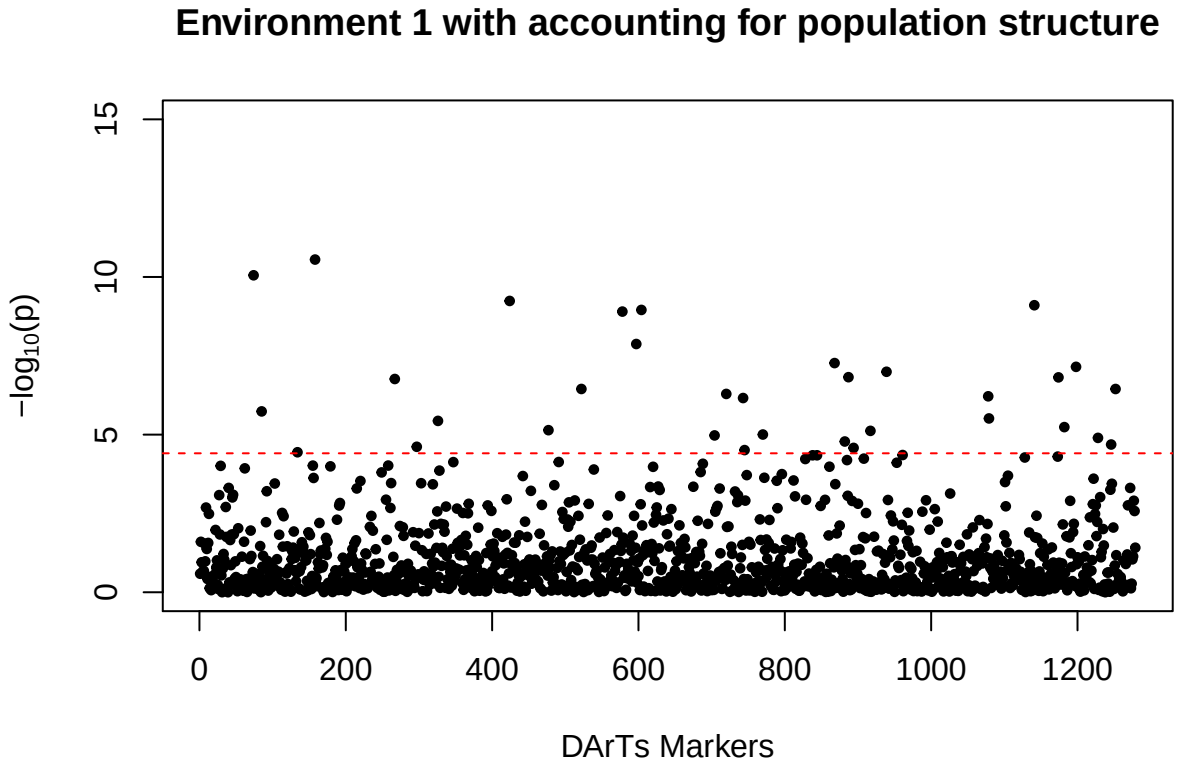


The PCA analysis and the k-means clustering showed a stratification of two groups in the population. This information was added as fixed effect in the previous single marker regresssion model ## correction for population structure (PCA)

```
pop.cofactor=factor(fit$cluster)
p.value=vector()
for(i in 1:ncol(X)){
  regression=lm(Y~pop.cofactor+X[,i])
  p.value[i]=summary(regression)$coef[3,4]
}

##Bonferroni correction
bonferroni=0.05/ncol(X)

## Manhattan plot with accounting for population structure.
plot(-log(p.value,base=10),pch=19,cex=0.6,
     ylab=expression(paste("-log"[10],"(p)", sep="")),
     xlab="DArTs Markers",
     main="Environment 1 with accounting for population structure",
     ylim=c(0,15))
abline(h=-log(bonferroni,base=10),col="red",lty=2)
```

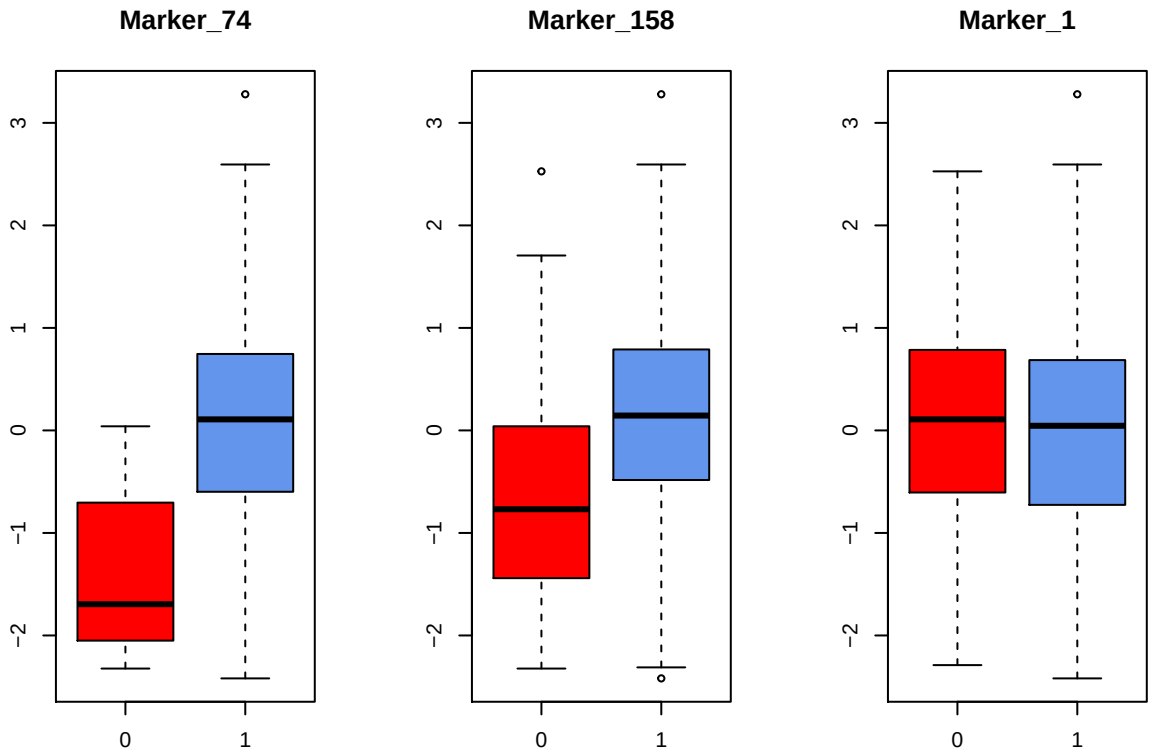



```
which(p.value<bonferroni)
```

```
##  [1]   74   85  134  158  267  297  326  424  477  522  578  597  604  704
## [15]  720  743  745  770  868  882  887  894  917  939 1078 1079 1141 1174
## [29] 1182 1198 1228 1246 1252
```

We can access the siginificant markers (p.value<-bonferroni threshold) and visuaalize the differences at the phenotypic level, as suggested in the QTL analysis.Markers 74 and 158 showed significant p-values, whereas marker 1 did not not. The boxplot can be used to explain these differences in terms of average grain yield.

```
par(mfrow=c(1,3))
boxplot(Y~ wheat.X[,74], main="Marker_74", col= c("red","cornflowerblue", ylab="average grain yield"))
boxplot(Y~ wheat.X[,158], main="Marker_158", col= c("red","cornflowerblue", ylab="average grain yield"))
boxplot(Y~ wheat.X[,1], main="Marker_1", col= c("red","cornflowerblue", ylab="average grain yield"))
```



Discuss the importance to consider population structure in GWAS analysis. Based on the literature, indicate different methods used to account for population structure.

Importance of considering population structure in GWAS:

Population structure represent genetic relatedness between samples at different scales, and it can create spurious associations. Without considering the population structure, it can decrease power and increase the false positive rate of tests of association (Price et al., 2010).

There are several methods can be used to account for population strucure.

1. Principal components analysis (PCA). The inferred principal components capturing the genetic ancestry of each individual are often included as fixed effects in a regression-based of association in order to account for population structure (Price et al., 2010).
2. Genomic control (GC). Estimating the degree of overdispersion generated by structure (Devlin et al., 2001).
3. Structured association (SA), assumes that the sampled population while heterogeneous, is composed of subpopulations that are themselves homogeneous. By using multiple polymorphisms throughout the genome, the SA method probabilistically assigns sampled individuals to latent subpopulations. (Pritchard et al., 2000).

Conclusion

Interpretation about how GWAS and these results could be used in plant breeding scenario. You can use this section to expand your knowledge, discuss advantages and disadvantages performing single marker regression, methods to improve the analysis, etc.

Generally, GWAS without consideration of genetic structure can create a lot of false positive association. Although we found a stratification of two groups in the population, the analysis with consideration of genetic structure did not decrease the number of association in this project.

As for breeding, GWAS can identify superior alleles and support introgression of these allele into elite breeding germplasm or facilitate marker-assisted selection (Zhu et al., 2008). In this project, we found several superior loci. We can use introgression one or more of these loci into an elite breeding germplasm to increase yield.

Single marker regression has some advantages such as simplicity and have the ability to handle a large number of markers. However, it can cause the type I error rate inflated. This is because single-marker regression only fits one SNP at a time, so a separate hypothesis test is performed for each marker locus. The resulting p-value do not take the multiple tests into account, so it may increase the rate of type I error.

Multiple test correction for significant threshold can improve this problem. For example, empirical thresholds calculated with permutation are the gold standard in QTL mapping studies. Another method is Bonferroni correction which compensates for that increase by testing each individual hypothesis at a significance level of a/m, where a is the desired overall significance level and m is the number of hypothesis. However, the typical Bonferroni correction is too conservative when the number of markers is extreme large, which may fail to detect many important loci(Wang et al., 2016).

Other Methods like a random-SNP-effect mixed linear model (RMLM) can improve the analysis. The RMLM treats the SNP-effect as random, but it allows a modified Bonferroni correction to be used to calculate the threshold p value for significance tests(Wang et al., 2016).

References

Chandra, A., Mitry, D., Wright, A., Campbell, H., and Charteris, D. (2014). Genome-wide association studies: applications and insights gained in Ophthalmology. Eye 28, 1066.

Devlin, B., Roeder, K., and Wasserman, L. (2001). Genomic control, a new approach to genetic-based association studies. Theoretical population biology 60, 155-166.

Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. Plant methods 9, 29.

Price, A.L., Zaitlen, N.A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. Nature reviews. Genetics 11, 459.

Pritchard, J.K., Stephens, M., Rosenberg, N.A., and Donnelly, P. (2000). Association mapping in structured populations. The American Journal of Human Genetics 67, 170-181.

Vilhjálmsson, B.J., and Nordborg, M. (2013). The nature of confounding in genome-wide association studies. Nature Reviews. Genetics 14, 1.

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation. The American Journal of Human Genetics 101, 5-22.

Wang, S.-B., Feng, J.-Y., Ren, W.-L., Huang, B., Zhou, L., Wen, Y.-J., Zhang, J., Dunwell, J.M., Xu, S., and Zhang, Y.-M. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. Scientific reports 6.

Zhu, C., Gore, M., Buckler, E.S., and Yu, J. (2008). Status and prospects of association mapping in plants. The plant genome 1, 5-20.