Lab 3 - Trying out GWAS

Alencar Xavier, Gota Morota

October 26, 2018

Prelude: Data & Structure

Getting some data

Example dataset from the package. We are querying two of the forty biparental families with a shared parental IA3023, grown in 18 environment.

```
Data = SoyNAM::BLUP(trait = 'yield', family = 2:3)
```

```
## solving BLUE of checks
## solving BLUP of phenotypes
## No redundant SNPs found
## There are 312 markers with MAF below the threshold
## Removing markers with more than 50% missing values
## Imputing with expectation (based on transition prob)
## removing repeated genotypes
## solving identity matrix
## indiviual 1 had 37 duplicate(s)
## indiviual 169 had 1 duplicate(s)
## indiviual 182 had 1 duplicate(s)
```

Genomic relationship matrix

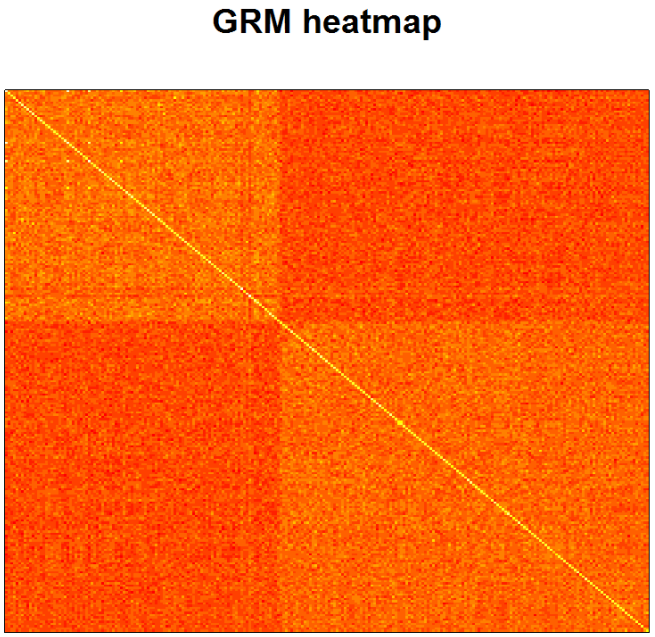
```
y = Data$Phen
M = Data$Gen
#
Z = apply(M,2,function(snp) snp-mean(snp))
ZZ = tcrossprod(Z)
Sum2pq = sum(apply(M,2,function(snp){p=mean(snp)/2; return(2*p*(1-p))}))
G = ZZ/Sum2pq
```

Kernel commonly deployed, referred in VanRaden (2008)

$$G = \frac{(M-P)(M-p)'}{2\sum_{j=1}^{J} p_j(1-p_j)}$$

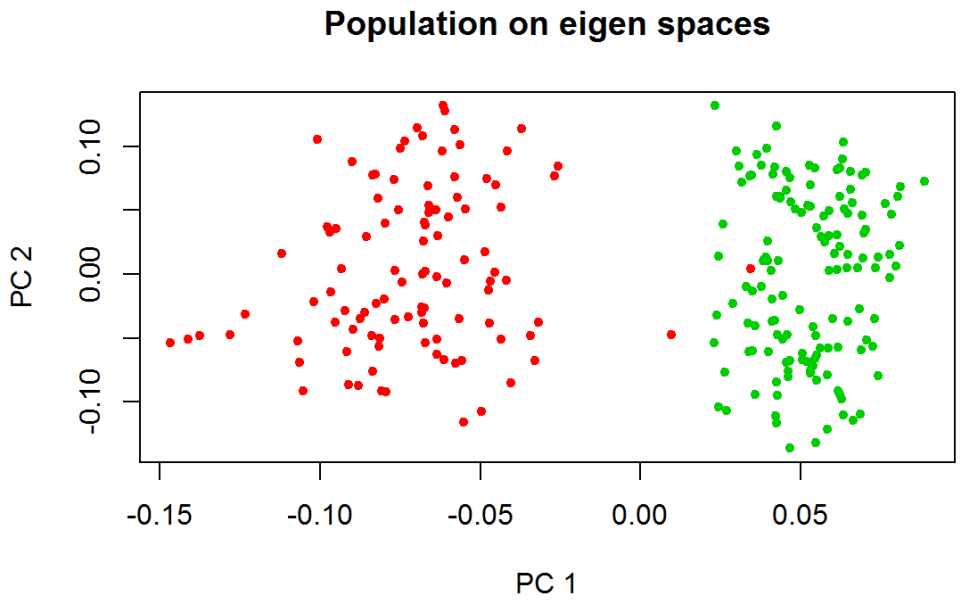
Genomic relationship matrix

```
image(G[,241:1], main='GRM heatmap',xaxt='n',yaxt='n')
```



Structure parameters (1) PCs

```
Spectral = eigen(G,symmetric = TRUE)
PCs = Spectral$vectors[,1:5]
plot(PCs,xlab='PC 1',ylab='PC 2',main='Population on eigen spaces',col=Data$Fam,pch=20)
```

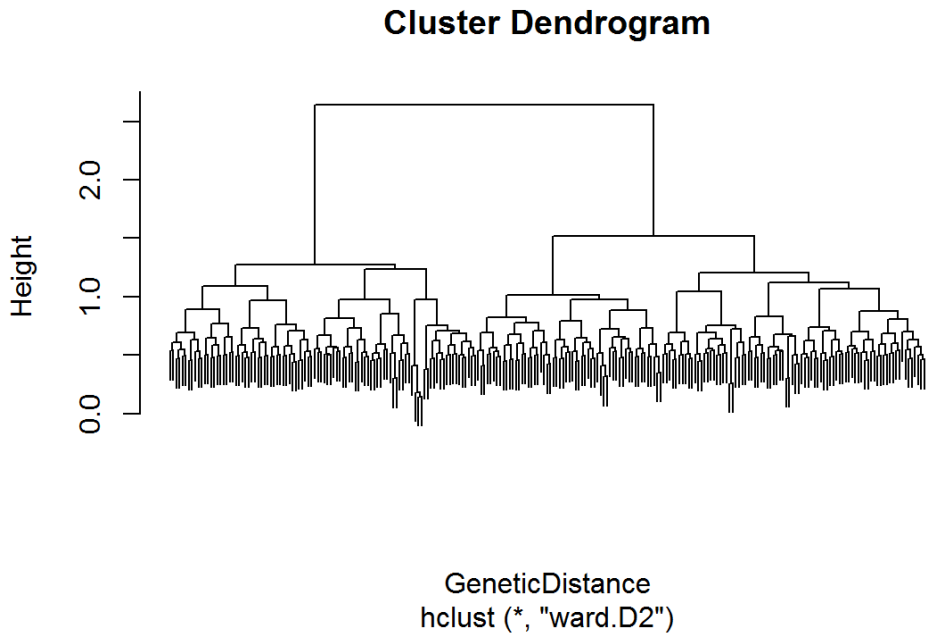


Structure parameters (2) Clusters

```
GeneticDistance = Gdist(M,method=6)


## Modified Rogers' distance


Tree = hclust(GeneticDistance,method = 'ward.D2')
plot(Tree,labels = FALSE)
```




```
Clst = factor(cutree(Tree,k=2))
```

Single marker analysis

GLM (1) - No structure

```
Marker = M[,117]
#
fit = lm( y ~ Marker )
anova( fit )
```

```
## Analysis of Variance Table
##
## Response: y
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## Marker      1  476321  476321  20.172 1.102e-05 ***
## Residuals 239 5643504   23613
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
-log(anova(fit)$`Pr(>F)`[1],base = 10)
```

```
## [1] 4.957736
```

GLM (2) - Principal Components

```
reduced_model = lm( y ~ PCs )
full_model = lm( y ~ PCs + Marker )
anova( reduced_model, full_model )
```

```
## Analysis of Variance Table
##
## Model 1: y ~ PCs
## Model 2: y ~ PCs + Marker
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    235 4060362
## 2    234 3562067  1    498295 32.734 3.215e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
-log((anova( reduced_model, full_model ))$`Pr(>F)`[2],base = 10)
```

```
## [1] 7.492813
```

GLM (3) - Population Clusters

```
reduced_model = lm( y ~ Clst )
full_model = lm( y ~ Clst + Marker )
anova( reduced_model, full_model )
```

```
## Analysis of Variance Table
##
## Model 1: y ~ Clst
## Model 2: y ~ Clst + Marker
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1    239 4275698
## 2    238 3652041  1    623657 40.643 9.4e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
-log( anova(reduced_model,full_model)$`Pr(>F)`[2],base = 10)
```

```
## [1] 9.026884
```

MLM - K+Q model

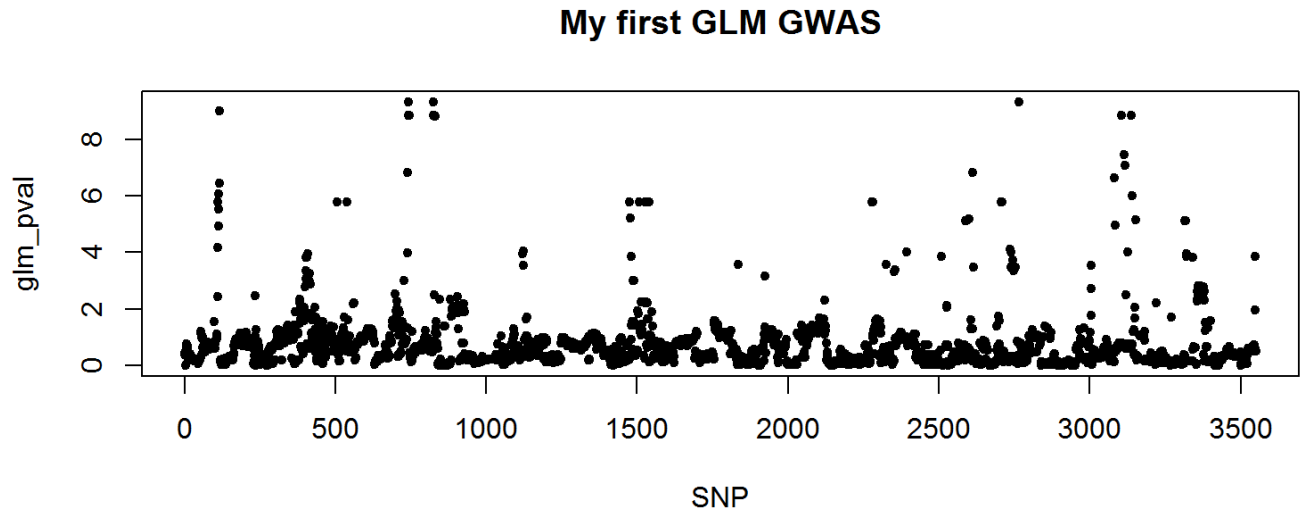
```
Q = model.matrix(~Clst)
reduced_model = reml( y=y, X=Q, K=G)
full_model = reml( y=y, X=cbind(Q, Marker), K=G)
LRT = -2*(full_model$loglik - reduced_model$loglik)
-log(pchisq(LRT,1,lower.tail=FALSE),base=10)
```

```
## [1] 10.80903
```

Whole genome screening

DYI (example with GLM)

```
reduced_model = lm( y ~ Clst )
glm_pval = apply(M,2,function(Marker){
  pval = anova(reduced_model, lm(y~Clst+Marker) )$`Pr(>F)`[2]
  return(-log(pval,base = 10))})
plot(glm_pval,pch=20,xlab='SNP',main='My first GLM GWAS')
```



Using CRAN implementations

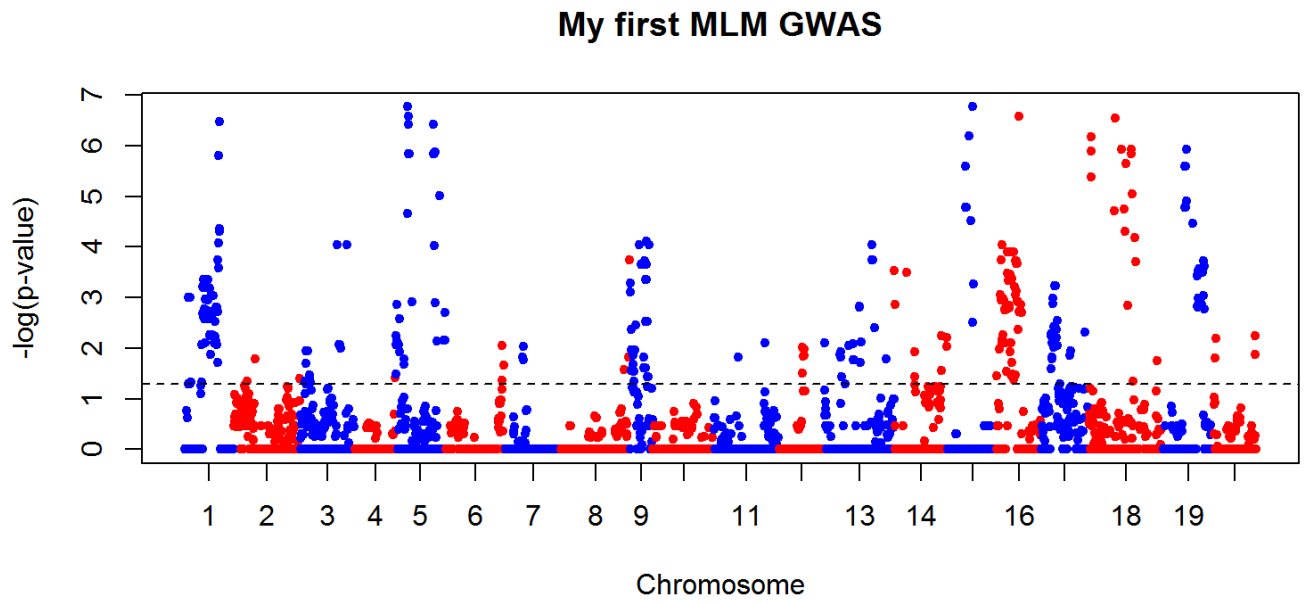
NAM random model: $y = \mu + Marker \times Pop + Zu + e$

```
fit_gwa = gwas3(y=y, gen=M, fam=c(Clst), chr=Data$Chrom)
```

```
## Calculating G matrix
## Solving polygenic model
## Starting Eigendecomposition
## Starting Marker Analysis
##
  |
  |                                                                      |   0%
  |
  |                                                                      |   1%
  |
  |=                                                                     |   1%
  |
  |=                                                                     |   2%
  |
  |==                                                                    |   2%
  |
  |==                                                                    |   3%
  |
  |==                                                                    |   4%
```

Manhattan plot

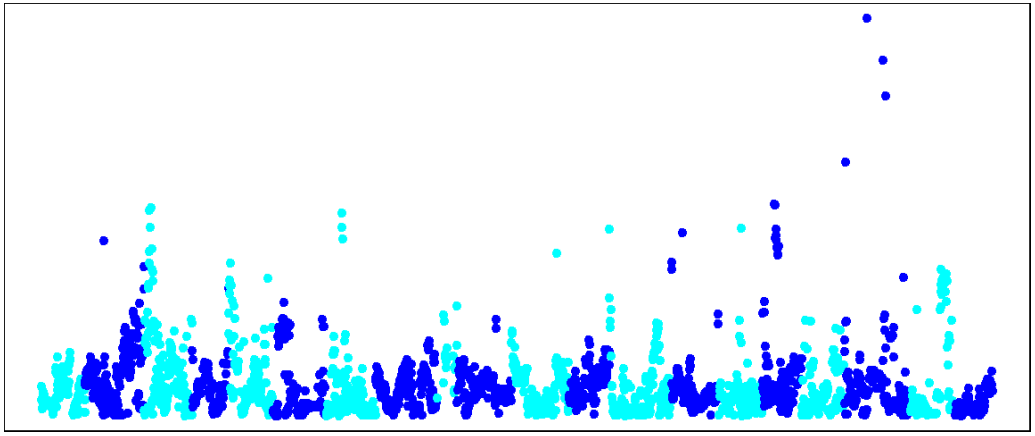
```
plot(fit_gwa, pch=20, main = "My first MLM GWAS")
```



Yet another R implementations

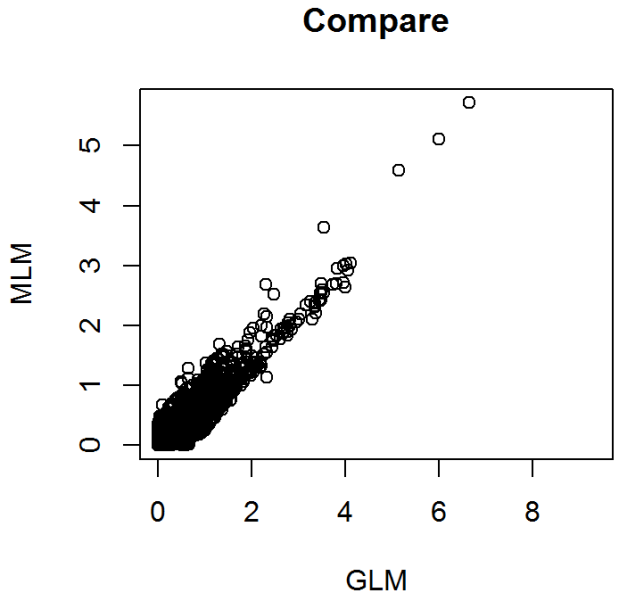
```
require(rrBLUP,quietly = TRUE); COL = fit_gwa$MAP[,1]%%2+1 # Color chromosomes
geno=data.frame(colnames(M),fit_gwa$MAP[,1:2],t(M-1),row.names=NULL)
pheno=data.frame(line=colnames(geno)[-c(1:3)],Pheno=y,Clst,row.names=NULL)
fit_another_gwa=GWAS(pheno,geno,fixed='Clst',plot=FALSE)
```

```
## [1] "GWAS for trait: Pheno"
## [1] "Variance components estimated. Testing markers."
```



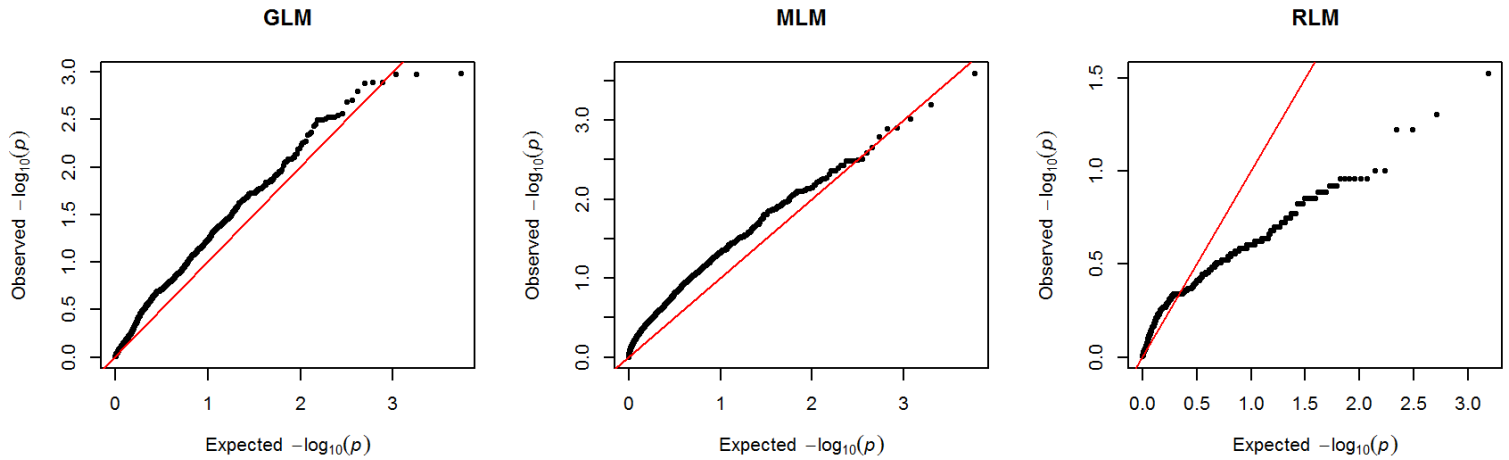
Comparing results

```
mlm_pval=fit_another_gwa$Pheno; mlm_pval[mlm_pval==0]=NA
plot(glm_pval,mlm_pval,xlab='GLM',ylab='MLM',main='Compare')
```



Power analysis - QQ plot

```
nam_pval = fit_gwa$PolyTest$pval
par(mfrow=c(1,3))
qqman::qq(glm_pval,main='GLM')
qqman::qq(mlm_pval,main='MLM')
qqman::qq(nam_pval,main='RLM')
```



Multiple testing

Multiple testing

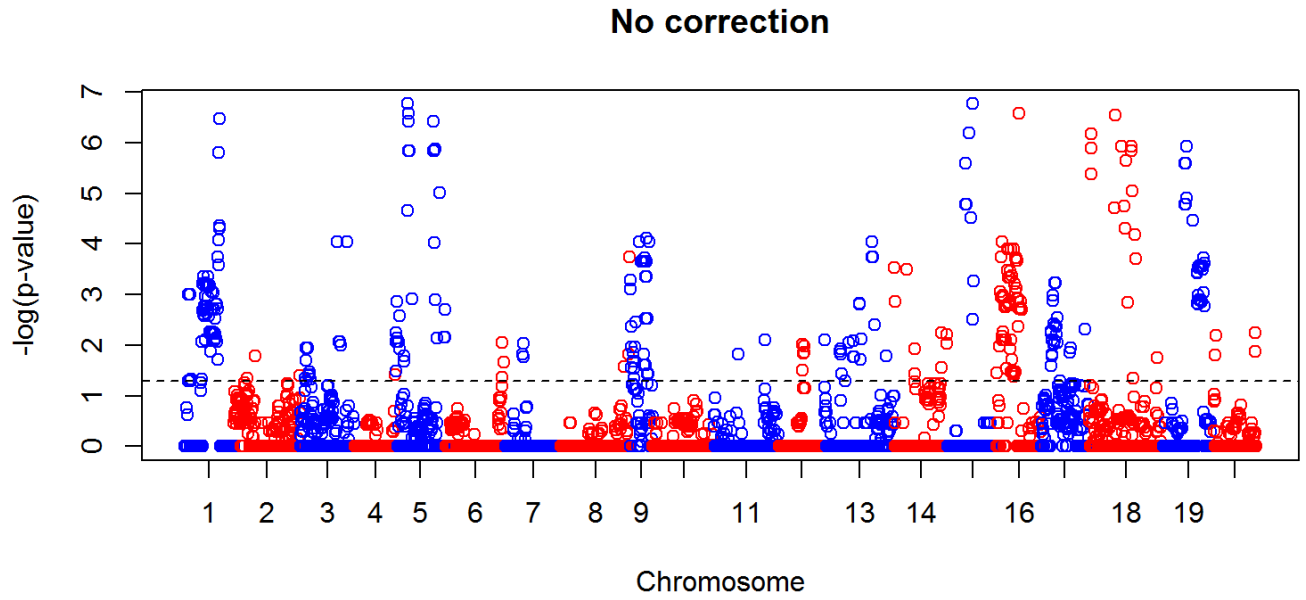
In statistics, the **multiple comparisons**, **multiplicity** or **multiple testing** problem occurs when one considers a set of statistical inferences simultaneously or infers a subset of parameters selected based on the observed values.In certain fields it is known as the look-elsewhere effect:

Several statistical techniques have been developed to prevent this from happening, allowing significance levels for single and multiple comparisons to be directly compared. These techniques generally require a **stricter significance threshold** for individual comparisons, so as to compensate for the number of inferences being made.

Baseline - No correction

Base significance threshold: $\alpha = 0.05/m$

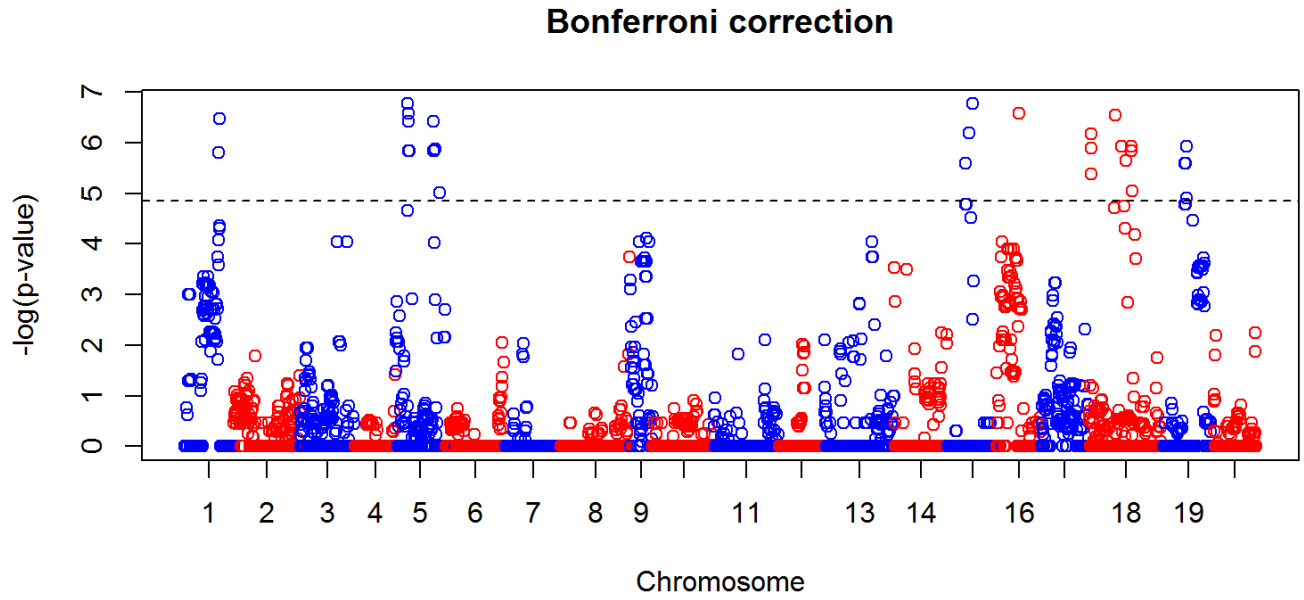
```
plot(fit_gwa, alpha=0.05, main = "No correction")
```



Multiple testing correction

Bonferroni: $\alpha = 0.05/m$

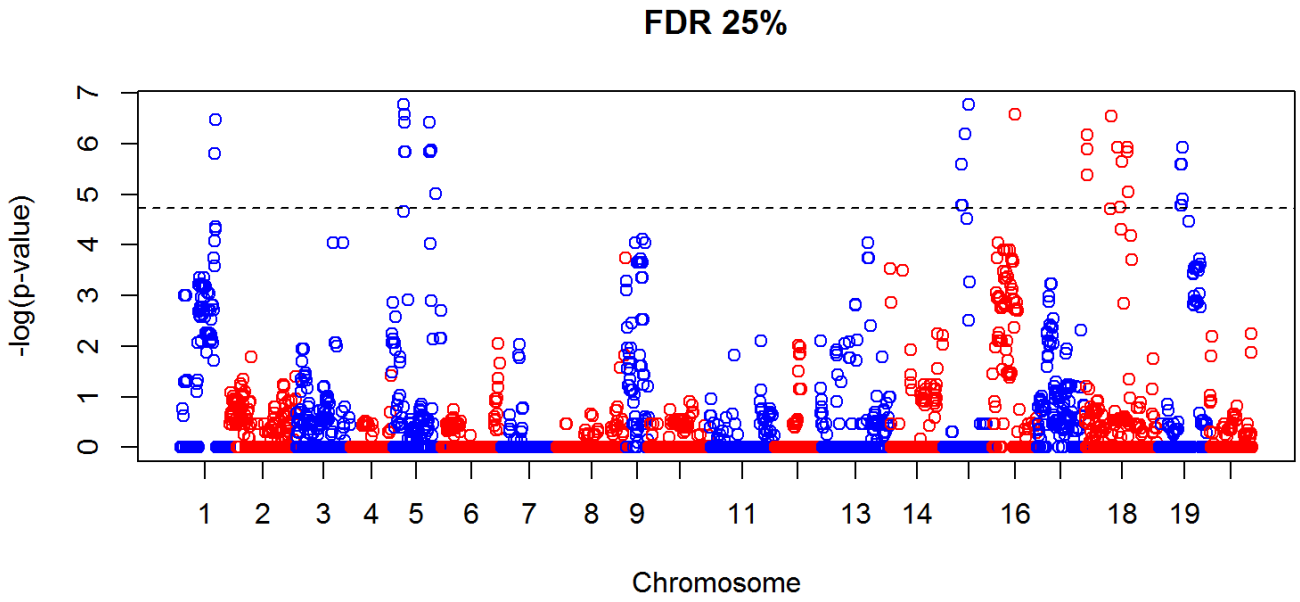
```
plot(fit_gwa, alpha=0.05/ncol(M), main = "Bonferroni correction")
```



False-Discovery Rate

Benjamini-Hochberg FDR: $\alpha = \frac{0.05}{m \times (1 - FDR)}$

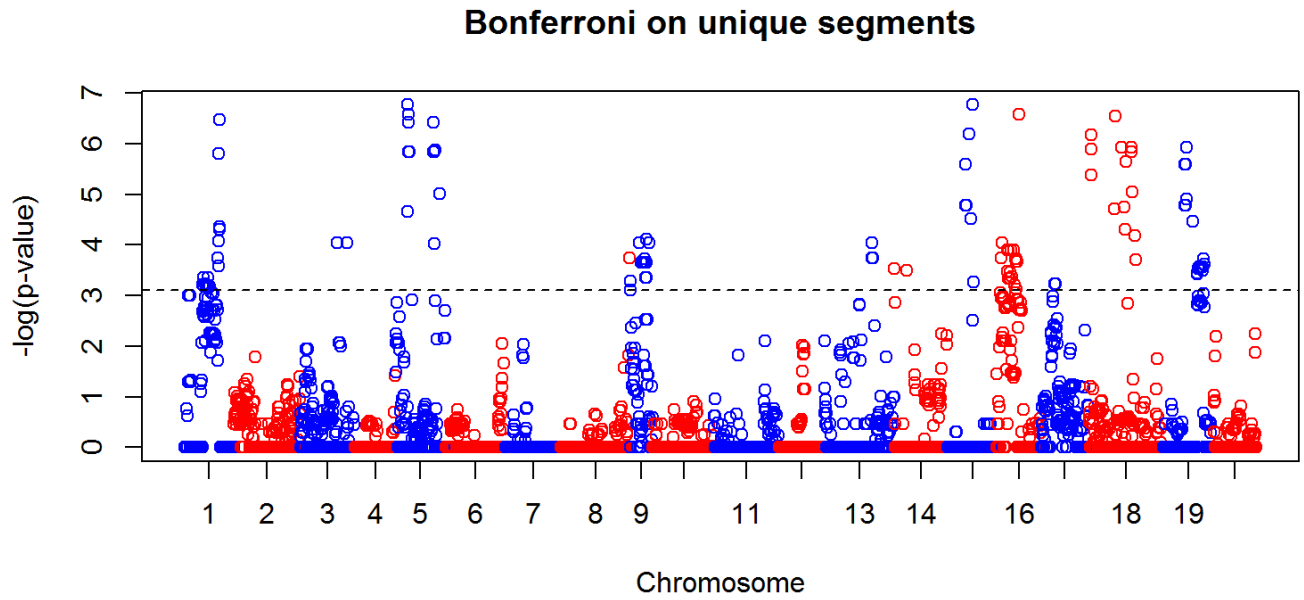
```
plot(fit_gwa, alpha=0.05/(ncol(M)*.75), main = "FDR 25%")
```



False-Discovery Rate

Unique segments based on Eigenvalues: $m^* = D > 1$

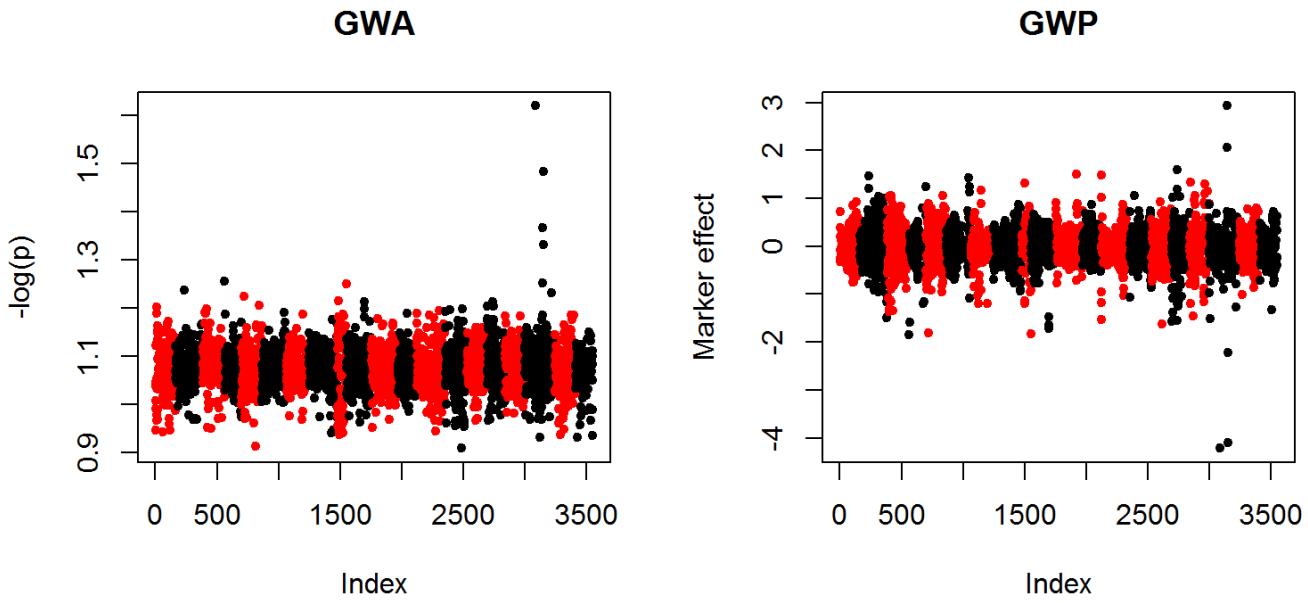
```
m_star = sum(Spectral$values>1)
plot(fit_gwa, alpha=0.05/m_star, main="Bonferroni on unique segments")
```



Multi-loci analysis

Whole genome regression

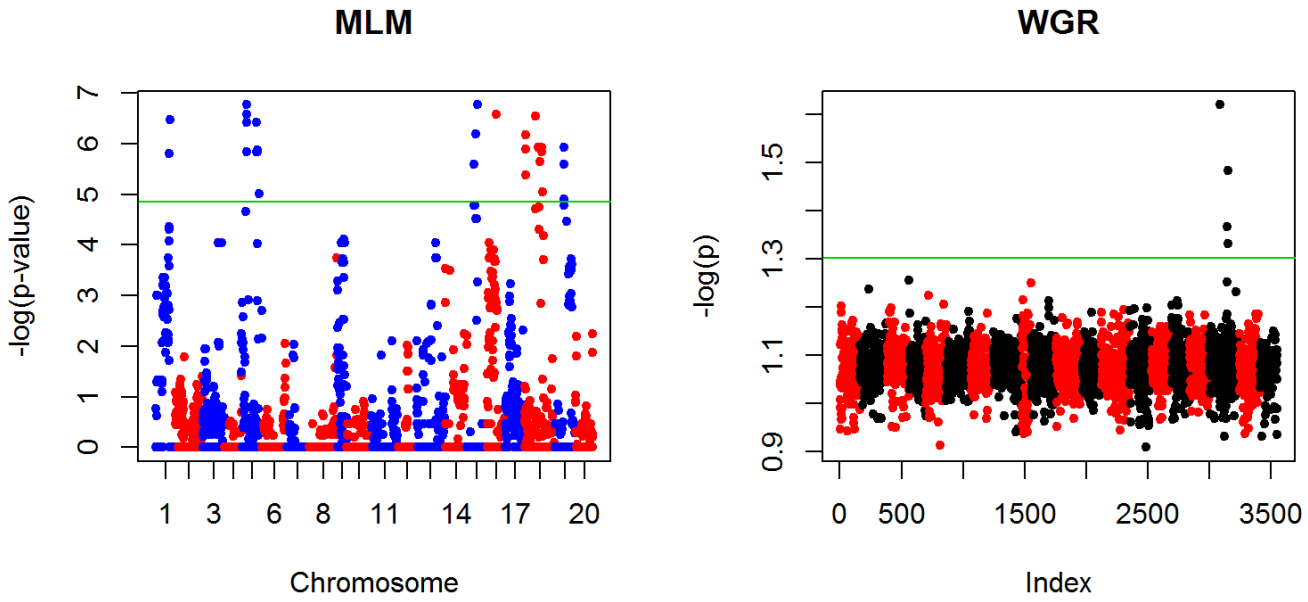
```
fit_wgr = bWGR::BayesDpi(y=y,X=M,it=3000); par(mfrow=c(1,2));
plot(fit_wgr$PVAL,col=COL,pch=20,ylab='-log(p)',main='GWA')
plot(fit_wgr$b,col=COL,pch=20,ylab='Marker effect',main='GWP')
```



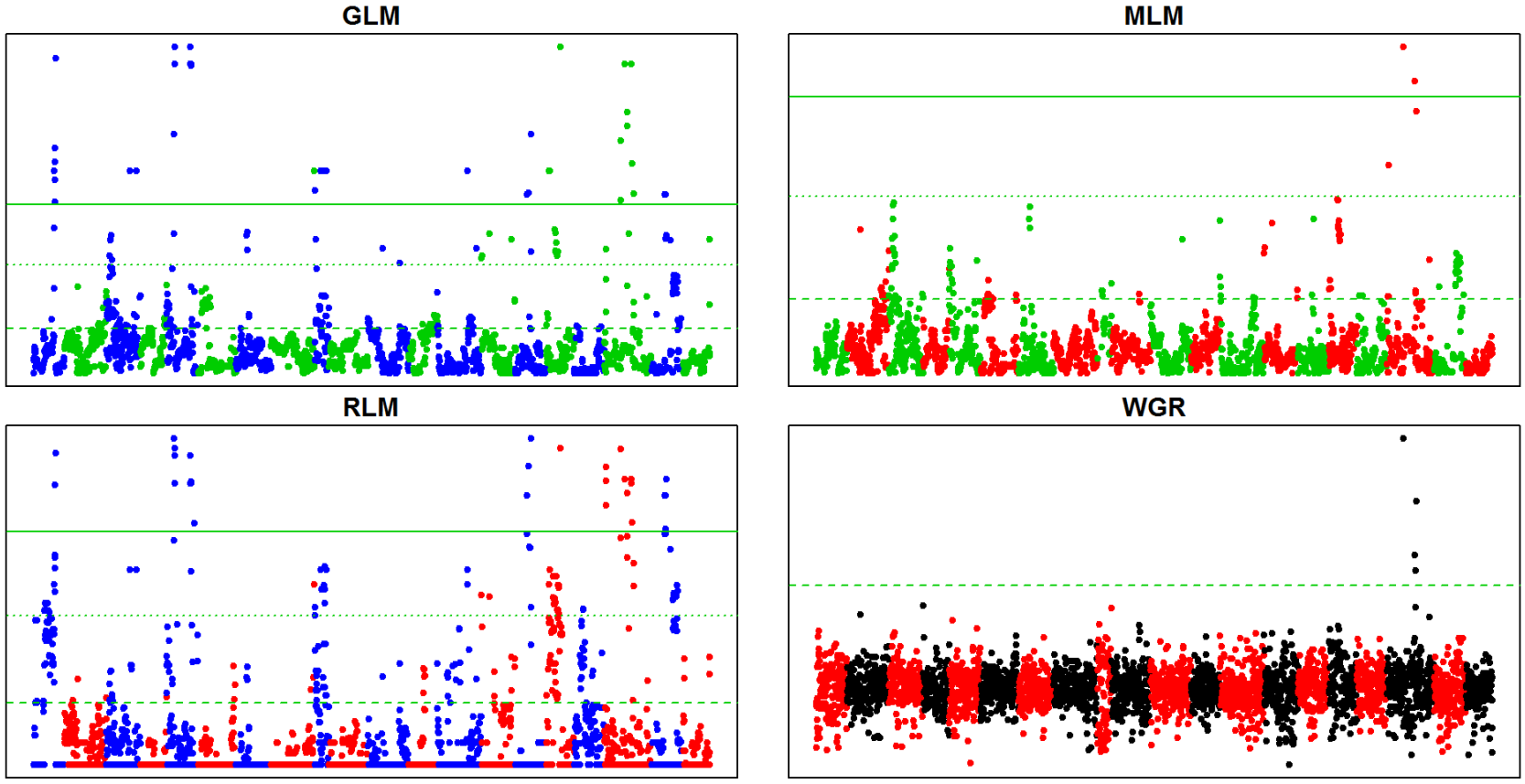

```
plot(fit_wgr$hat,y,pch=20)
```

WGR - No need for multiple testing

```
thr_none = -log(pchisq(qchisq(1-0.05/ncol(M),1),1,lower.tail=FALSE),base=10)
thr_bonf = -log(pchisq(qchisq(1-0.05,1),1,lower.tail=FALSE),base=10)
par(mfrow=c(1,2)); plot(fit_gwa,alpha=NULL,main="MLM",pch=20); abline(h=thr_none,col=3)
plot(fit_wgr$PVAL,col=COL,ylab='-log(p)',main="WGR",pch=20); abline(h=thr_bonf,col=3)
```

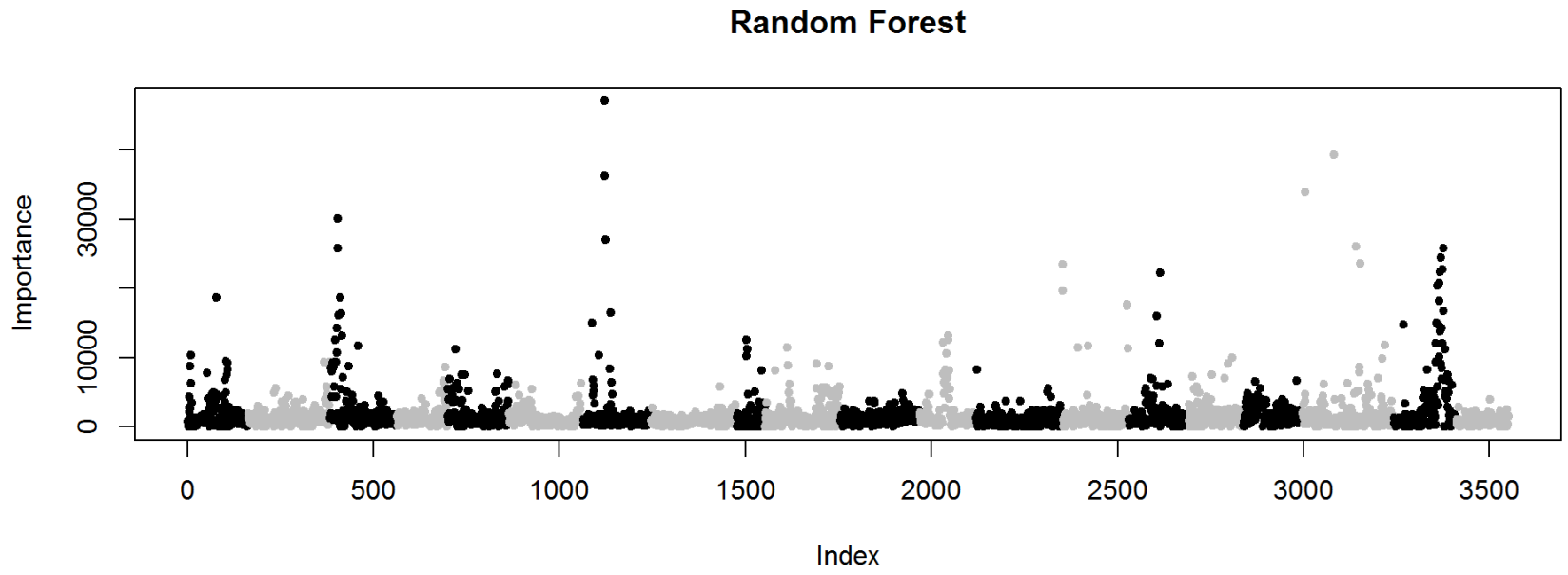


Approaches are complementary



Random forest

```
fit_rf = ranger::ranger(y~.,data= data.frame(y=y,M),importance='impurity')
plot(fit_rf$variable.importance,ylab='Importance',main='Random Forest',col=COL+7,pch=20)
```



Break