SM1 - Calibrating and combining results from two statistical models for genome scans for selection

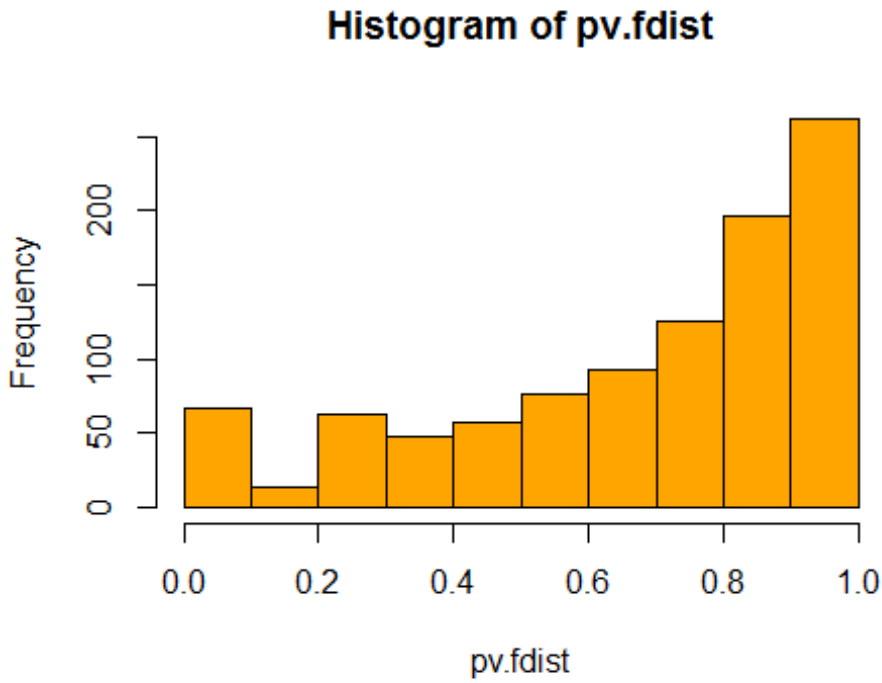
This supplementary material file explains how p-values computed from genome scans for selection (GSS) could be recalibrated to control type I errors and false discovery rates. Then it illustrates how results from two well-calibrated tests could be combined to increase power to reject selective neutrality.

We considered a two-population model in which populations evolved under migration-drift equilibrium (Wright's 2-island model), and for which one hundred diploid organisms were genotyped in each population. The genetic data were simulated from the coalescent program **ms** using the **-s 1** (SNP) option. We used 4Nm = 20 at **900** truly neutral loci and 4Nm = 0.1 at **100** truly adaptive loci. The genotypes for each individual were recorded in a matrix stored in the lfmm format in the file "ex1geno.lfmm".

Running genome scan for selection

We performed GSS using two methods: 1) the computer program **FDIST2**, 2) F_{ST} estimated from an **ANOVA** approach. For FDIST2, the average F_{ST} value over all the loci was equal to F_{ST} = 0.07. A target F_{ST} value of 0.05 was used during the computation of the FDIST test p-values. The p-values for FDIST2 were recorded in the file "FDISTpv.dat" generated by FDIST2 (fourth column). They were read as follows.

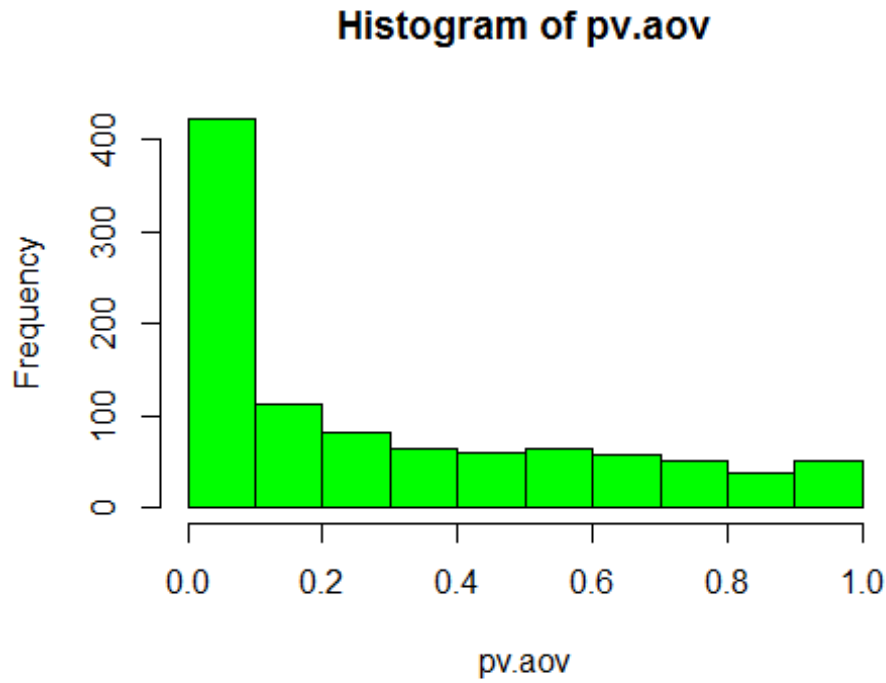
```
pv.fdist = 1 - read.table("FDISTpv.dat")[,4]
hist(pv.fdist, col = "orange")
```



The histogram shows that the FDIST2 tests were overly conservative.

Next, we used R to perform ANOVA tests at each locus, considering the population labels **(pop)** as factors. We used the ANOVA p-values to perform a second GSS for the same data as with FDIST2. This can be reproduced as follows.

```
geno = read.table("ex1geno.lfmm")
z.aov = NULL
pop = rep(0:1, each = 100)
for (l in 1:1000) z.aov[l] = summary(lm(geno[,l]~pop))$coef[2,3]

pv.aov = pchisq(z.aov^2, df = 1, lower = F)
hist(pv.aov, col = "green")
```



The histogram shows that the ANOVA tests were overly liberal.

To summarize these preliminary results, it appears that the FDIST tests were too conservative whereas the ANOVA tests were too liberal. To check this claim, we applied the Benjamini-Hochberg (BH) FDR control algorithm to the sets of p-values computed for each model, and we compared the lists of candidate loci obtained from the BH algorithm. In the BH algorithm, the level of FDR (q.level) was set to 10 percent.

```
L = length(pv.fdist)
q.level = 0.1

# BH algorithm for the fdist and anova pvalues
w = which(sort(pv.fdist) < q.level*(1:L)/L)
candidates.fdist = order(pv.fdist)[w]

w = which(sort(pv.aov) < q.level*(1:L)/L)
candidates.aov = order(pv.aov)[w]
```

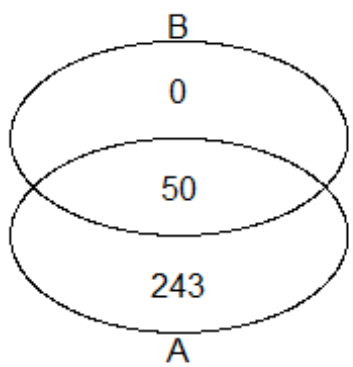
For both methods, the observed FDR and power were computed as follows.

```
FDR.fdist = mean( candidates.fdist < 901)
POW.fdist = sum(candidates.fdist > 900)/100
FDR.aov = mean( candidates.aov < 901)
POW.aov = sum(candidates.aov > 900)/100
print(cbind(FDR.fdist, POW.fdist, FDR.aov, POW.aov))
```

```
##      FDR.fdist POW.fdist   FDR.aov POW.aov
## [1,]         0       0.5   0.70    0.85
```

This result confirms that the FDIST tests were overly conservative (FDR=0) and the ANOVA tests were overly liberal (FDR = 0.70). Let us look at the Venn diagram for the two lists of candidates obtained from the BH algorithms.

```
library(gplots)

venn.1 = list(candidates.aov, candidates.fdist)
#A = ANOVA and B = FDIST
venn(venn.1)
```



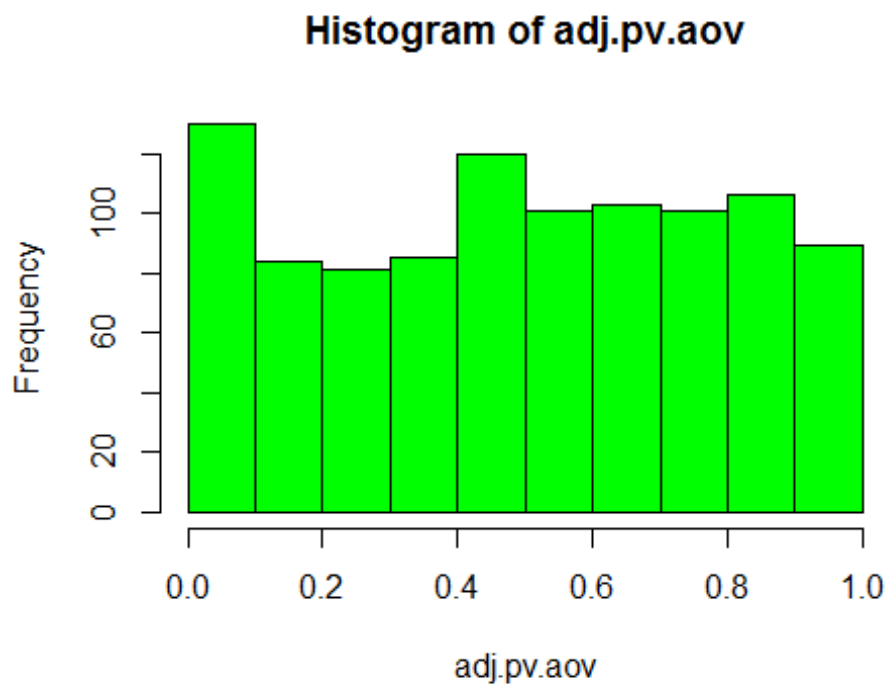
The Venn diagram was highly unbalanced, and suggested that the BH algorithm provided unreliable results (A = ANOVA, B = FDIST).

Recalibrating GSS tests using "empirical null" distributions

Since the two methods gave contrasted results and the tests were poorly calibrated, we applied corrections to the tests in order to recalibrate their null-hypothesis. Our approach followed a general strategy called "empirical-null hypothesis" testing. This approach postulates that enough information about the null-hypothesis is contained in the data to build the null-hypothesis empirically.

We began with recalibrating the ANOVA tests. Using the ANOVA scores (z.aov), the recalibration method worked as follows.

```
#Compute the genomic inflation factor
gif = median(z.aov^2)/qchisq(.5, df = 1)
#re-calibrate p-values
adj.pv.aov = pchisq(z.aov^2/gif, df = 1, lower = F)
hist(adj.pv.aov, col = "green")
```



In this command, we used the **genomic inflation factor** (gif) to correct the tests. The histogram had the desired shape: It was flat and it showed a peak near zero. In other datasets, it might be advisable to set the normalization constant to other values than the gif in order to reach this objective (see examples below).

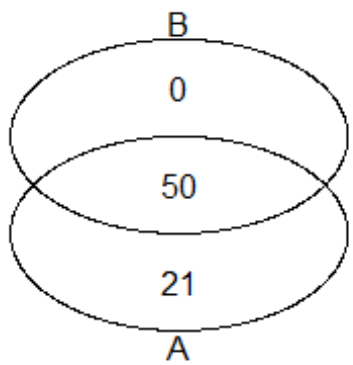
To control the FDR, we used the Benjamini-Hochberg algorithm again, and proposed a new list of candidate loci for the ANOVA GSS (the FDR level was left unchanged) as follows.

```
w = which(sort(adj.pv.aov) < q.level*(1:L)/L)
candidates.aov = order(adj.pv.aov)[w]

FDR.adj.aov = mean(candidates.aov < 901)
POW.adj.aov = sum(candidates.aov > 900)/100
print(cbind(FDR.fdist, POW.fdist, FDR.adj.aov, POW.adj.aov))

##      FDR.fdist POW.fdist FDR.adj.aov POW.adj.aov
## [1,]         0       0.5   0.056        0.67
```

The false positives were almost removed from the list of candidates (observed FDR = 0.056). The observed FDR of 0.056 was less than the expected level of 10 percent, which

was consistent which the expectation of the BH algorithm. The power (0.67) was greater than the value obtained with FDIST (0.50).

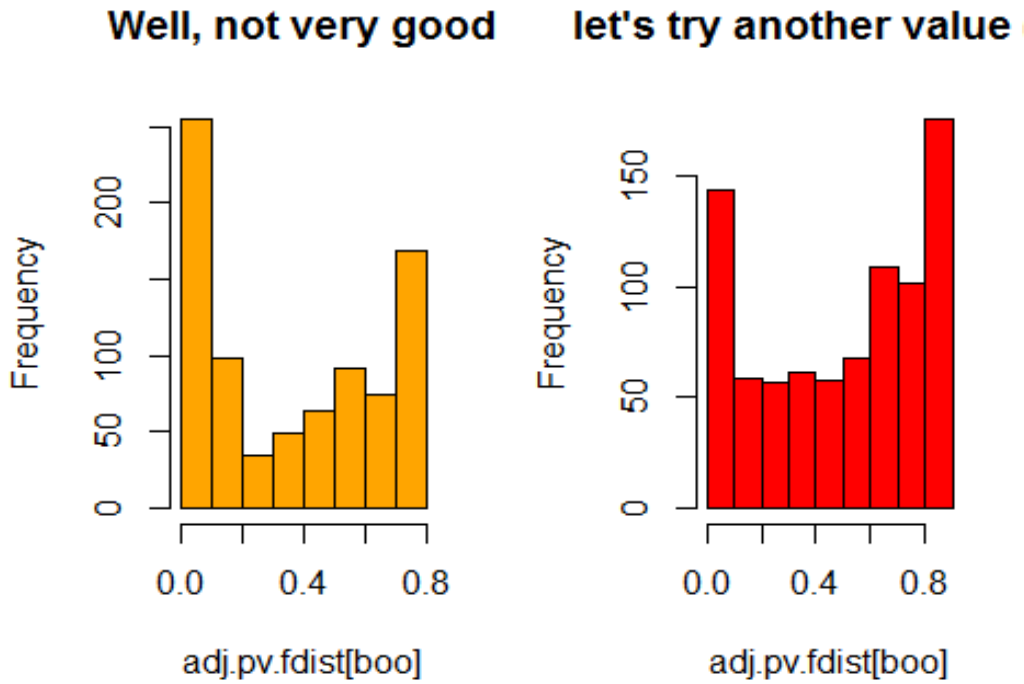
```
venn.2 = list(candidates.aov, candidates.fdist)
#A = ANOVA and B = FDIST
venn(venn.2)
```



The Venn diagram was still unbalanced (A = ANOVA, B = FDIST).

To go one step further, we applied the recalibration approach to the FDIST p-values. Then we could compare the two tests more fairly. The FDIST test recalibration required an additional step, which consisted of back-transforming p-values into squared scores. This was done by using the quantile function (qchisq) as follows.

```
#convert p-values into z^2
z.sq = qchisq(1 - pv.fdist, df = 1, lower = T)
#re-calibrate FDIST p-values
gif = median(z.sq)/qchisq(.5, df = 1)
gif

## [1] 0.1612595

par(mfrow = c(1,2))
adj.pv.fdist = pchisq(z.sq/gif, df = 1, lower = F)
boo = adj.pv.fdist < .8
hist(adj.pv.fdist[boo], col = "orange", main = "Well, not very good")
adj.pv.fdist = pchisq(z.sq/.4, df = 1, lower = F)
hist(adj.pv.fdist[boo], col = "red", main = "let's try another value")
```



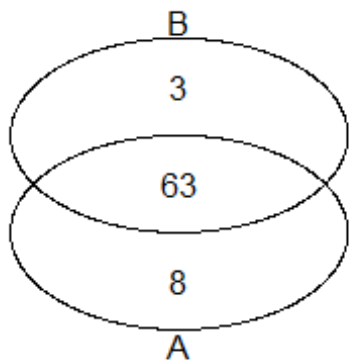
The orange histogram was computed using the genomic inflation factor (gif = 0.16), and it had an undesired shape. So, we estimated the inflation factor using trial and error. Note that large p-values do not influence our list of candidate loci, and they can be ignored safely.

After two trials, we found that a value of lambda = 0.4 led to satisfying results (red histogram). Although an excess of large p-values was still observed, the red histogram had the desired shape. We updated the FDIST list of outlier loci and re-evaluated the observed FDR and power as follows.

```
par(mfrow = c(1,1))
w = which(sort(adj.pv.fdist) < q.level*(1:L)/L)
candidates.fdist = order(adj.pv.fdist)[w]

FDR.adj.fdist = mean(candidates.fdist < 901)
POW.adj.fdist = sum(candidates.fdist > 900)/100
print(cbind(FDR.adj.fdist, POW.adj.fdist, FDR.adj.aov, POW.adj.aov))

##      FDR.adj.fdist POW.adj.fdist FDR.adj.aov POW.adj.aov
## [1,]    0.015           0.65      0.056         0.67
```

The recalibration operations clearly increased the power of FDIST, which reached the value 0.65, greater than the previous value of 0.5. This was achieved at the cost of an increased level of FDR (0.015) which remained lower than the expected value of 10 percent.

```
venn.3 = list(candidates.aov, candidates.fdist)
#A = ANOVA and B = FDIST
venn(venn.3)
```



The Venn diagram exhibited a reasonable balance between the results of the ANOVA (A) and the FDIST (B) GSS.

Combining scores from multiple programs

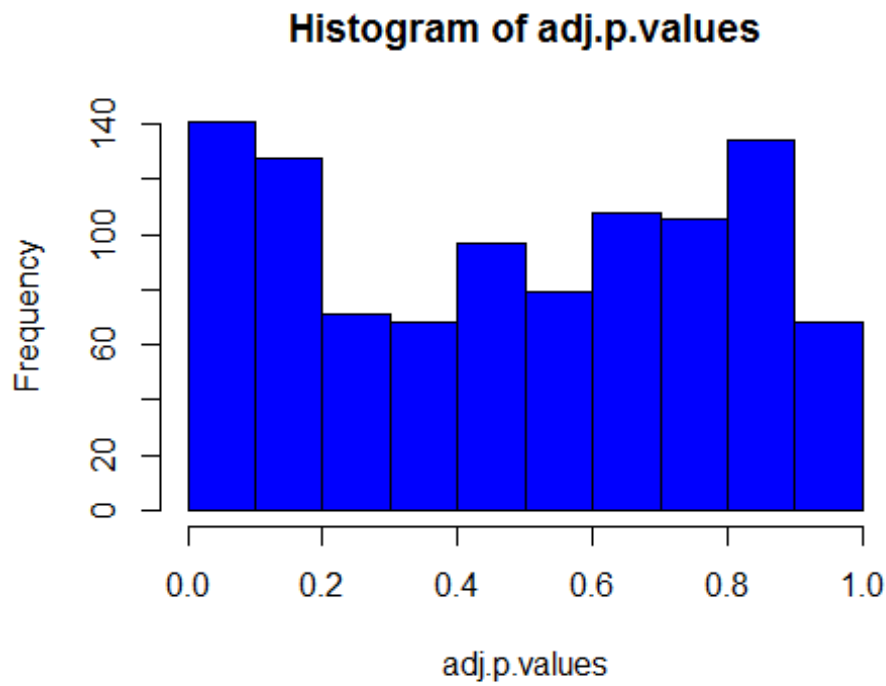
At this point, we obtained two reasonable lists of candidates - each list satisfying the FDR control criterion (FDR less than 10 percent). Next, we asked the question: Can we provide better results by combining the FDIST and ANOVA approaches?

A naive option would merge the 2 lists of candidate loci. A more conservative option would use the intersection of the two lists. We used another technique, based on a meta-analysis procedure to build a single list from combined z-scores. This procedure could be considered as a variant of the Stouffer method for combining Gaussian tests.

The method consisted of taking the median value of adjusted z^2-scores, and computing new p-values from the squared scores.

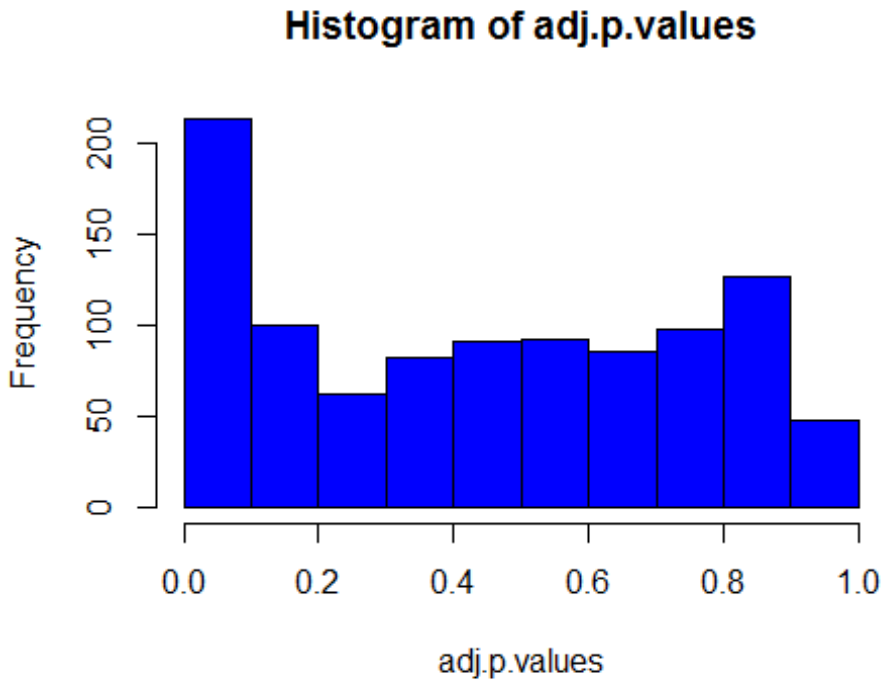
```
#bind the z^2 scores from FDIST and ANOVA
z2 = cbind(z.sq/0.4, z.aov^2/4.39)
#Combine z-scores using the median value
z2.median = apply(z2, MARGIN = 1, median)
#re-adjust for test combination
lambda = median(z2.median)/0.456
lambda

## [1] 0.7847165
```

```r
# compute adjusted p-values from the combined z-scores
adj.p.values = pchisq(z2.median/lambda, df = 1, lower = FALSE)
#histogram of p-values
hist(adj.p.values, col = "blue")
```



The blue histogram of adjusted p-values does not have the desired shape. So, we recalibrated the p-values by using a value of lambda less than the 0.78. We tried lambda = .6, and obtained the following result.

```r
# compute adjusted p-values from the combined z-scores
adj.p.values = pchisq(z2.median/.6 , df = 1, lower = FALSE)

#histogram of p-values
hist(adj.p.values, col = "blue")
```



This was much better! We computed our final list of candidate loci using the BH algorithm.

```
w = which(sort(adj.p.values) < q.level*(1:L)/L)
candidates = order(adj.p.values)[w]

FDR = mean(candidates < 901)
POW = sum(candidates > 900)/100
print(cbind(FDR, POW))

##             FDR  POW
## [1,]       0.13 0.73
```

For this list, the power was equal to 0.73 and this was much higher than our initial value of 0.5. The observed FDR is close to its expected value of 10 percent. We eventually obtained a more powerful testing procedure by combining the two tests.

Conclusions

The main results of this example study could be summarized as follows:

- Uncalibrated tests could not be compared directly.
- A well-calibrated model-free approach (ANOVA) could outperform a poorly calibrated model-based approach (FDIST), although the simulated data were closer to the FDIST model.

- **GSS statistical tests could be recalibrated by using simple procedures**. Here, the use of trial and error sometimes helps to determine a better constant (and the genomic inflation factor is generally overly conservative).
- **Combining well-calibrated tests increased the power to reject neutrality**.