Practical 4:

Population Structure in Association Testing and Functional Annotation

Tyler Medina
ID 13101308

This analysis will consist of two separate parts. First, a data set of Northern European ancestry will be analyzed for SNP association, and will then be analyzed for evidence of population structure before proceeding to a second assocation test that excludes outliers if they are found.

The second section will consist of investigating evidence of functional significance for SNPs found to be significant in a previous analysis.

Part I: Population Structure in Assocation Testing

Assocation Testing

The tool SNPTEST2 can be used to test for association of a SNP with the case group in the study:

```
./snptest -data psdata.gen psdata.sam -pheno phen -method score -frequentist 1 -o psdata.snptest.1.out
```

The table below shows an excerpt of the association test, including the SNP ID, position, frequentist p-value, and minor allele frequencies in the case and control..

```r
snptest1 <- read.table("psdata.snptest.1.out", header=T)
snptest1[c(1:10),c(2,4:6,30,31,42)]
```

```
##          rsid position alleleA alleleB cases_maf controls_maf
## 1   rs3934834  1045729       C       T     0.128        0.113
## 2   rs3737728  1061338       A       G     0.275        0.289
## 3   rs6687776  1070488       C       T     0.094        0.085
## 4   rs9651273  1071463       A       G     0.332        0.334
## 5   rs4970405  1088878       A       G     0.048        0.065
## 6  rs12726255  1089873       A       G     0.082        0.076
## 7   rs2298217  1104902       C       T     0.094        0.069
## 8   rs4970357  1116987       C       A     0.092        0.080
## 9   rs4970362  1134661       A       G     0.333        0.326
## 10  rs9660710  1139265       A       C     0.070        0.058
##    frequentist_add_pvalue
## 1               0.2970520
## 2               0.4924070
## 3               0.4912450
## 4               0.9226650
## 5               0.1022420
## 6               0.6183350
## 7               0.0456254
## 8               0.3463970
## 9               0.7399370
## 10              0.2843140
```

The genomic inflation factor is calculated below.

```
gif1 <- median(qchisq(snptest1[,42], df=1, lower.tail=F), na.rm =T)/0.456
gif1
```

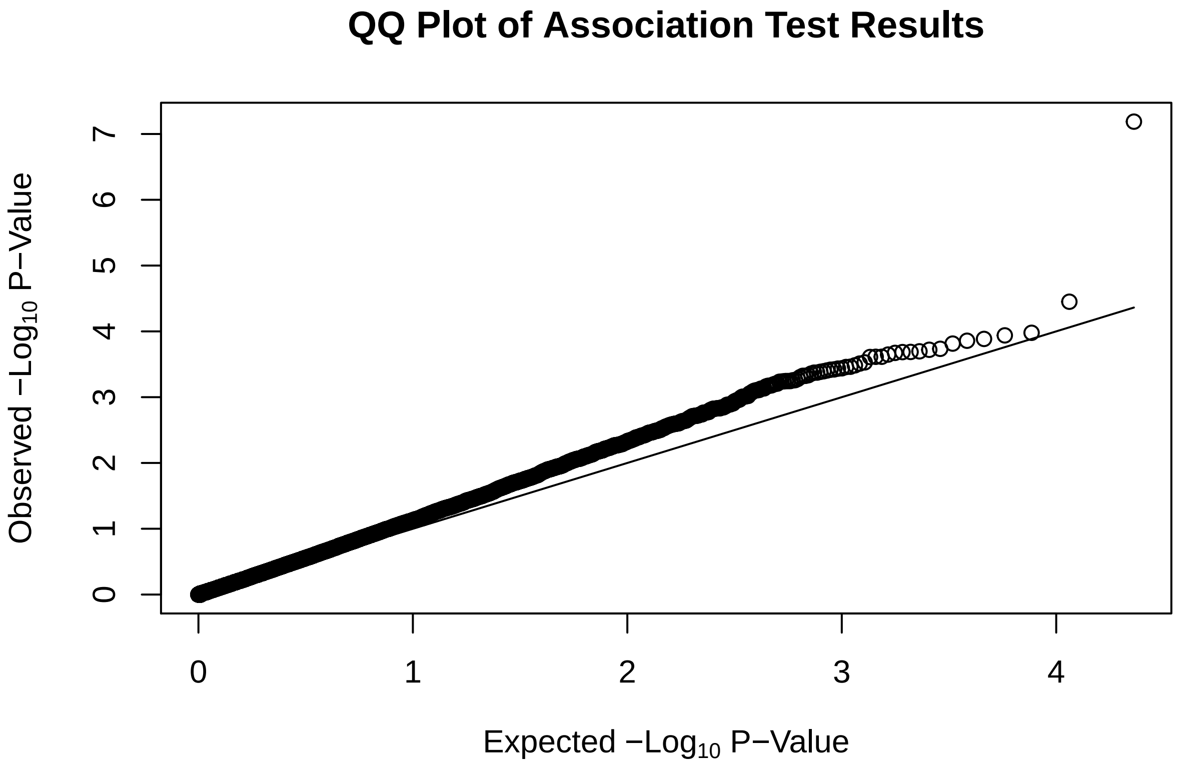
```
## [1] 1.158258
```

The inflation factor value of 1.158 reveals that there may be a small amount of population structure in the the data. In other words, there is some kind of variable shared by a subset of the population that has not been accounted for and is affecting the association data. GIF values closer to 1 indicate no population structure, so because the value here is fairly close, significant amounts of structure are not expected, but some is likely present.

QQ Plot

The QQ (Quantile-Quantile) Plot below compares the quantiles between a standard distribution and the data. If the resultant comparison follows the line x=y, the data distribution shows less bias. If however the plots deviates from this line, as seen in the plot below, there is further evidence for some underlying confounding factor, such as population structure.

```
index1 <- -log10(seq(1,nrow(snptest1))/nrow(snptest1))
logp1 <- -log10(snptest1[,42])
qqplot(index1, logp1, xlab=expression("Expected -Log"[10]*" P-Value"),
       ylab=expression("Observed -Log"[10]*" P-Value"))
title("QQ Plot of Association Test Results")
lines(index1,index1)
```



Significant Results

From the data as it stands, the SNPTEST results shows that the 5 following SNPs show the most significant p-values from the chi-squared tests performed between the case and control distributions:

```
top1 <- snptest1[order(logp1,decreasing=TRUE)[1:5], c(2,42)]
top1
```

```
##             rsid frequentist_add_pvalue
## 1000   rs4908527            6.49417e-08
## 13097   rs401904            3.54195e-05
## 10276  rs1063302            1.05129e-04
## 16173  rs3817586            1.15079e-04
## 17675 rs12095873            1.30307e-04
```

Principle Component Analysis

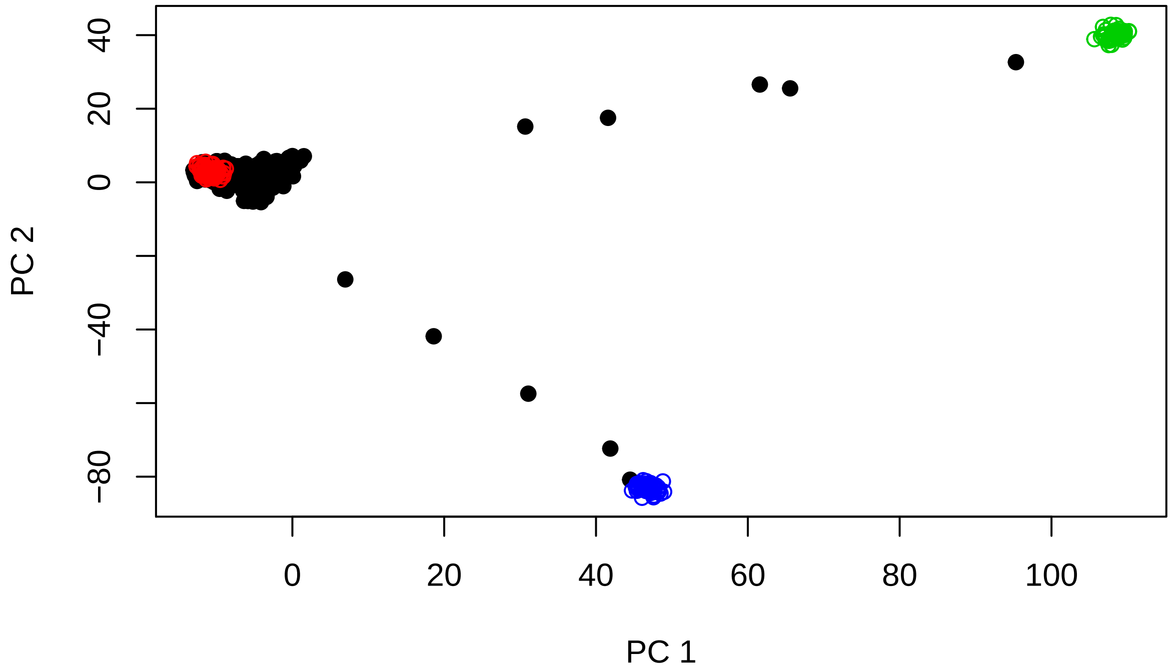
Because there is evidence of some population structure within this Northern European sample, it may be useful to compare the sample data to dissimilar groups. If any individuals in the sample are shown to be more similar to the non-sample data, these may be ruled as outliers and excluded, as they may be introducing an unknown variable into the sample that confounds the results.

To test for similarities between other groups, the sample data is first combined with previous known data. Three groups of data are added, which include groupings of data from European, Chinese/Japanese, and Yoruba individuals.

The combined data set is then analyzed using a principle component analysis of the individuals. This method will allow for the visualization of overall similarities and dissimilarities between individuals. The PCA shows the original European sample in black, the comparison European sample in red, the Yoruba sample in green, and the Chinese/Japanese sample in blue:

```
combo <- read.table("psdata.combined.txt", header=F)
pca <- prcomp(x=combo[,4:ncol(combo)], center=T, scale=T, retx=T)

plot(pca$x[,1], pca$x[,2], xlab="PC 1", ylab="PC 2", type="n")
for(i in 1:nrow(combo)){
  if(combo[i,2]=='CEU') points(pca$x[i,1],pca$x[i,2],col=2)
  if(combo[i,2]=='YRI') points(pca$x[i,1],pca$x[i,2],col=3)
  if(combo[i,2]=='CAJ') points(pca$x[i,1],pca$x[i,2],col=4)
  if(combo[i,2]=='Case') points(pca$x[i,1],pca$x[i,2],col=1,pch=19)
  if(combo[i,2]=='Control') points(pca$x[i,1],pca$x[i,2],col=1,pch=19)
}
```



The results of the PCA reveal that there are indeed individuals with overall similarities with either the Han Chinese population or the Yoruba population, noted as the points falling in lines between the red, green, and blue clusters. These individuals' data will be removed from the association analysis, as this population structure is likely negatively affecting the quality of the data. These individuals can be identified as follows, and results are printed below.

```
outliers <- identify(pca$x[,1], pca$x[,2], labels=combined_data[,1])
outliers_ids <- as.vector(combined_data[outliers,1])
write(outliers_ids, "sample.exclusions")
```

```
##    Excluded Individuals
## 1                ST1991
## 2                ST1992
## 3                ST1993
## 4                ST1994
## 5                ST1995
## 6                ST1996
## 7                ST1997
## 8                ST1998
## 9                ST1999
## 10               ST2000
```

Assocation Test without Outliers

A second association test will be performed with the identified individuals removed from the testing, as noted in the `-exclude_samples` argument below:

```
./snptest -data psdata.gen psdata.sam -pheno phen -method score -frequentist 1 -o psdata.snptest.2.out
-exclude_samples sample.exclusions
```

```
snptest2 <- read.table("psdata.snptest.2.out",header=T)
```

The genomic inflation factor for this second test is now calculated to compare to the original:

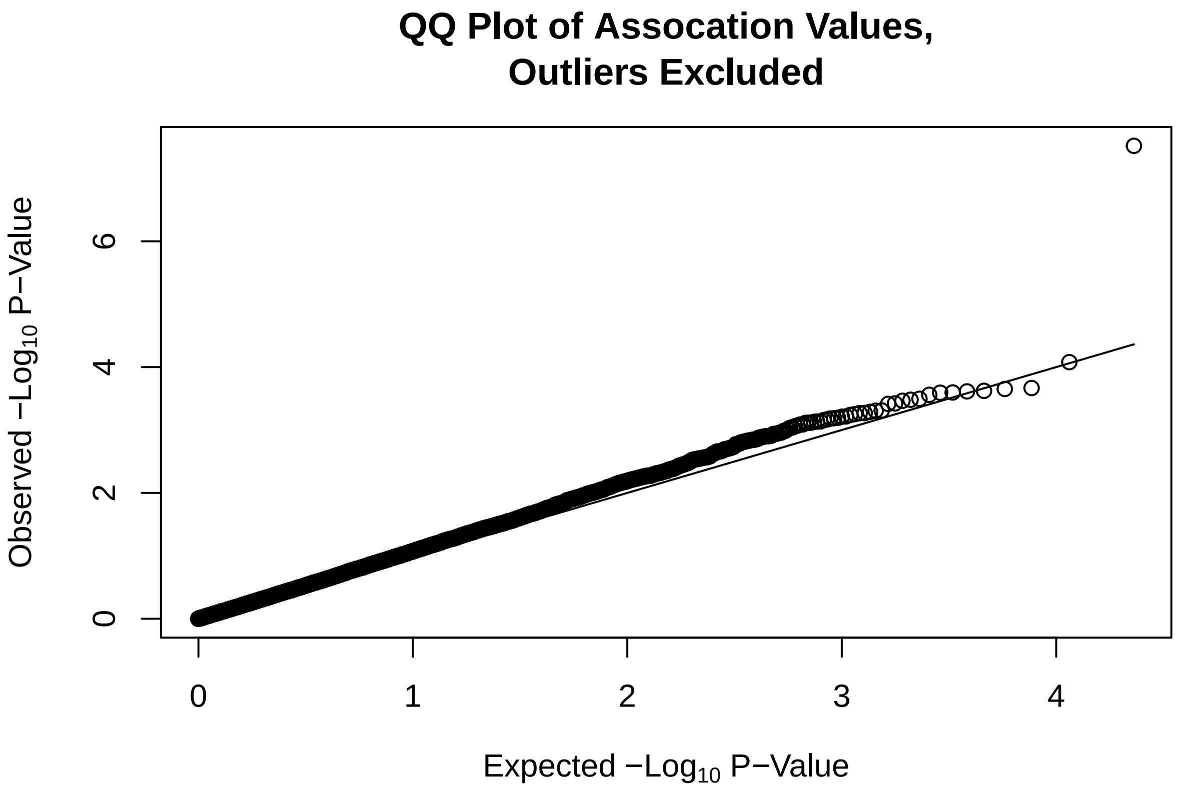
```
gif2 <- median(qchisq(snptest2[,42], df=1, lower.tail=F), na.rm=T)/0.456
gif2
```

```
## [1] 1.082964
```

The new GIF of 1.08 shows that the removal of the outlying individuals has reduced the amount of population structure in the data. This is further noted in the QQ plot below.

QQ Plot without Outliers

```
index2 <- -log10(seq(1,nrow(snptest2))/nrow(snptest2))
logp2 <- -log10(snptest2[,42])
qqplot(index2, logp2, xlab=expression("Expected -Log"[10]*" P-Value"),
       ylab=expression("Observed -Log"[10]*" P-Value"))
title("QQ Plot of Assocation Values,\nOutliers Excluded")
lines(index2,index2)
```



The QQ plot now shows a closer adherence to the x=y line, showing less evidence of bias in the data.

Significance without Outliers

The SNPs with the most significant values after excluding outliers are show below.

```
top2 <- snptest2[order(logp2,decreasing=TRUE)[1:5], c(2,42)]
top2
```

```
##             rsid frequentist_add_pvalue
## 1000   rs4908527            3.04151e-08
## 10276  rs1063302            8.33702e-05
## 17675 rs12095873            2.15150e-04
## 13097   rs401904            2.22508e-04
## 7762  rs11161912            2.38402e-04
```

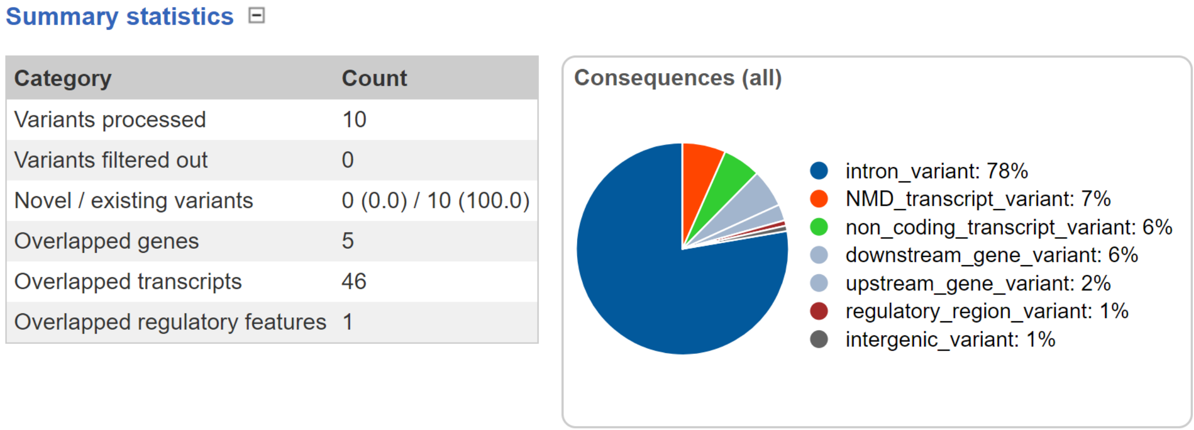
Although similar to the SNPs found before removing the outliers, several SNPs have changed position, and some have dropped out of the top 5 altogether, showing that while the outliers may not have had an enormous effect on the significance testing, they most definitely have a non-zero effect.

Part II: Functional Annotation

9 SNPs with significant adjusted association values were identified in a previous analysis:

```
##   CHR       SNP     UNADJ        GC      BONF      HOLM  SIDAK_SS
## 1   3 rs6802898 2.327e-20 3.442e-20 6.758e-15 6.758e-15       Inf
## 2  10 rs7901695 6.563e-12 8.161e-12 1.906e-06 1.906e-06 1.906e-06
## 3  16 rs8050136 1.006e-08 1.172e-08 2.921e-03 2.920e-03 2.916e-03
## 4  16 rs3751812 1.017e-08 1.185e-08 2.952e-03 2.952e-03 2.948e-03
## 5  10 rs7904519 2.478e-08 2.865e-08 7.197e-03 7.197e-03 7.171e-03
## 6   3 rs7615580 2.789e-08 3.221e-08 8.099e-03 8.099e-03 8.066e-03
## 7  10 rs7903146 3.889e-08 4.478e-08 1.129e-02 1.129e-02 1.123e-02
## 8   3 rs6768587 3.966e-08 4.566e-08 1.152e-02 1.152e-02 1.145e-02
## 9   3 rs2028760 4.220e-08 4.855e-08 1.225e-02 1.225e-02 1.218e-02
##    SIDAK_SD    FDR_BH    FDR_BY
## 1       Inf 6.758e-15 8.891e-14
## 2 1.906e-06 9.530e-07 1.254e-05
## 3 2.916e-03 7.381e-04 9.711e-03
## 4 2.948e-03 7.381e-04 9.711e-03
## 5 7.171e-03 1.350e-03 1.776e-02
## 6 8.066e-03 1.350e-03 1.776e-02
## 7 1.123e-02 1.362e-03 1.791e-02
## 8 1.145e-02 1.362e-03 1.791e-02
## 9 1.218e-02 1.362e-03 1.791e-02
```

The next step in understanding these associations is to investigate the biological properties of these SNPs. The list of SNPs is first run through the Ensembl Variant Effect Predictor (Fig. 1).



Summary statistics ⊟

Category	Count
Variants processed	10
Variants filtered out	0
Novel / existing variants	0 (0.0) / 10 (100.0)
Overlapped genes	5
Overlapped transcripts	46
Overlapped regulatory features	1

Consequences (all)

● intron_variant: 78%
● NMD_transcript_variant: 7%
● non_coding_transcript_variant: 6%
● downstream_gene_variant: 6%
● upstream_gene_variant: 2%
● regulatory_region_variant: 1%
● intergenic_variant: 1%

Figure 1: Ensembl Variant Effect Predictor, Summary Statistics

The results indicate that all 9 variants were previously known (2 records were found for rs7615580, bringing the total to 10 in the image), and the majority are intronic variants.

Next, more specific information about the SNPs can be gathered from SNPedia and GeneCard. The most significant SNP, rs6802898, will be used for example.

Although rs6802898 is an intronic SNP, it is found upstream from a gene. SNPedia shows that is is associated with the PPARG gene, and also displays the relative frequency of the SNP genotypes across populations (Fig. 2).

The CEU (European) population shows low rates of the T allele, particularly when compared to the YRI (Yoruba) population, which conversely shows that a homozygous T genotype predominates. As explored in the previous section, differences like these can cause confounding factors with association analysis, as

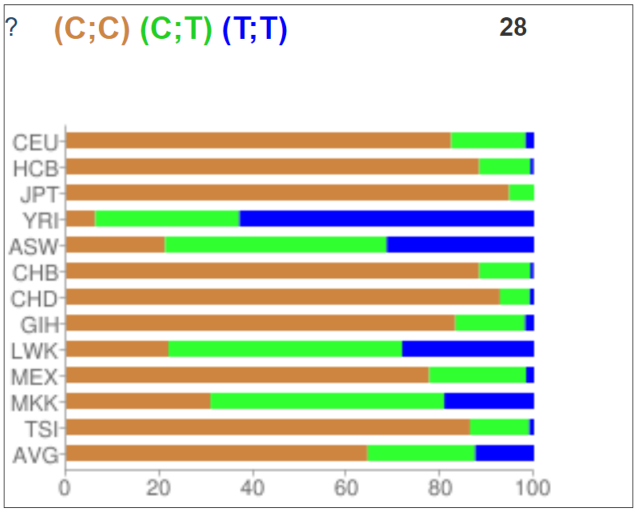


Figure 2: SNPedia Genotype Frequency

population structure can increase the prevelance of an allele in the study, altering the definition of a “minor” allele between individuals in the study.

From GeneCards, the PPARG gene codes for Peroxisome Proliferator-Activated Receptor Gamma. These nuclear receptors are part of pathways that include lipid metabolism and glucose homeostasis, as well as transcriptional regulation and differentiation in adipocytes. Variants in this gene are often involved in type II diabetes, which may have implications in incidence rates between populations such as the Yoruba population, or possibly Americans of African decent, who show elevated incidence of type II diabetes.