Population Genetic Analysis III: Biallelic SNP Data
BZ/MIP 577 Fall 2015

Kelly Pierce

Part II: Evaluating drift vs. selection as mechanisms producing observed structure

F_{ST} outliers may indicate selection

We saw in Part I that there is some evidence for population structure in American dog ticks found at two different collection sites. We now want to determine whether that structure arose as the result of genetic drift or natural selection. We have used F_{ST} to describe broad patterns of differentiation within populations by aggregating allele frequency data from multiple loci, but F_{ST} can also be used to detect evidence of selection at individual loci.

Evolutionary rates vary throughout the genome, and consequently F_{ST} can vary for different loci under consideration. Just as some loci are highly conserved while others are highly divergent, F_{ST} can be very high for some loci and very low for others. We can generate a distribution of F_{ST} values for the population by calculating the F_{ST} for each locus individually. For small datasets, such as the prairie dog dataset with seven loci, there really is not enough information to develop a distribution of F_{ST} values. However, next generation sequencing can provide data on thousands of loci. This is sufficient to properly describe a distribution of F_{ST} values. If we state our null hypothesis to be that all variation in F_{ST} values across the genome is the result of genetic drift, we can then infer that any outliers in that distribution are the result of natural selection acting to change allele frequencies.

Objectives

1. Describe allele frequences at multiple loci.
2. Understand the importance of correcting for multiple comparisons in statistical analysis.
3. Identify F_{ST} outliers from a large SNP dataset and interpret results.

The Bayescan Model

Bayescan (Foll and Gaggiotti 2008) is a software package that estimates F_{ST} for each locus in a multilocus SNP dataset. As you know, there are many formulae for estimating F_{ST}, and the appropriate choice depends on the characteristics of your data (e.g., ploidy, number of loci, number of alleles per locus). The Bayescan model describes allele frequences with *locus-level* and *population-level* parameters. The frequency of allele p at locus i in subpopulation j follows a multinomal Dirichlet distribution,

(1) $p_{ij} \sim \mathrm{Dir}(\theta_{ij}p_{i1}, \ldots, \theta_{ij}p_{iK_i})$,

where p_{iK} is the frequency of allele K at locus i and θ_{ij}=$1/F_{ST}^{ij}$-1.

The *locus-level* and *population-level* parameters are α_i and β_j, respectively. These parameters are related to θ_{ij} by the formula

(2) $\log(1/\theta_{ij})$=α_i+β_j.

The parameters α_i and β_j are estimated for each SNP in the dataset using a Markov Chain Monte Carlo (MCMC). Briefly, MCMC procedures try different values of parameters and calculate a likelihood score for each proposed parameter set. By trying a number of different α_i and β_j parameters in the model and determining the likelihood of each set over many successive iterations, eventuall the estiamtes of α_i and β_j convege on the most likely values.

Our null hypothesis is that genetic drift is the primary determinant of allele frequencies. When this is the case, the locus-level parameter $\alpha_{ij} = 0$. If the locus-level parameter for a SNP is significantly different than 0 we have evidence consistent with natural selection acting on allele frequencies. Non-zero locus-level parameters correspond with higher F_{ST} values. Bayescan reports both the calculated F_{ST} and a p-value indicating whether the corresponding locus-level parameter was significantly different from the null expectation of genetic drift.

Data

Input SNP data

Once again you will be working with the American dog tick (*Dermacentor variabilis*) SNP dataset. We have 2387 tick SNPs from a reduced-representation *de novo* sequencing project. These ticks were collected from two sites in Texas, and you saw in the previous lab that there is some evidence for spatial structure in the tick population. We will use Bayescan to test the hypothesis that this structure is the result of selection against a null hypothesis that drift drives structure.

As you may have noticed by now, every different software package we use requires data in a new and different format. Bayescan is, of course, no different. The Bayescan input format has the following structure:

```
[loci]=2387

[populations]=2

 [pop]=1
 1  116  2  0 116
 2  122 2   117 5
 3  124 2   12 112
 .
 .
 .

[pop]=2
 1  58  2   1 57
 2  48  2   44 4
 3  54  2   5 49
 .
 .
 .
```

There is one row for each locus i in each subpopulation j. The rows for each population have the following columns (in order from left to right):

- locus number
- number of observations of that locus in the population (not all loci will be sequenced in all individuals, but Bayescan can handle missing data)
- number of alleles present (Bayescan can handle microsatellite data with more than two alleles, but in our case we have only biallelic SNPs)

- frequency of allele A
- frequency of allele B

R cannot handle the Bayescan input file well because of the header rows in the file. We are providing you instead with a slightly reformatted table of raw allele frequencies, "wide_format_Bayescan.txt", to use for exploratory analysis. That dataset has the following columns from left to right:

- locus number
- count of allele A in population 1
- count of allele B in population 1
- frequency of allele A in populaiton 1
- frequency of allele B in population 1
- count of allele A in population 2
- count of allele B in population 2
- frequency of allele A in populaiton 2
- frequency of allele B in population 2

Output F_{ST} data

Like many programs designed for analyzing large NGS datasets, Bayescan takes a long time to run. Rather than have you install Bayescan on your computers and wait for the approximately 1 hour the program would need to analyze the allele frequency data, we have provided you with the output containing the estimated F_{ST} values calculated from the fitted α_i and β_j parameters. Each row contains information on one locus, with the columns "prob" (p-value for locus), "log10.PO" (negative log-likelihood score), "qval" (multiple-comparison corrected p-value; more on this below), "alpha" (estimated locus-specific parameter), and "fst" (as expected, the F_{ST} value).

Exercises

Instructions

Complete the following steps in R. Your end result should be a script that loads the requisite packages and walks through the analyses described. Write answers to specific questions in the comments of your code.

Install and load the required packages.

You will need to obtain the *qvalue* library from BioConductor.

```
source("https://bioconductor.org/biocLite.R")
biocLite("qvalue")
library(qvalue)
```

Download and load the data

Allele frequency table containing the same information as in the Bayescan input file:

```
# load the raw allele frequency data
setwd('~/Desktop/CSU_PopGen_Labs/Data/')
allele.freqs<-read.table('wide_format_Bayescan.txt', header=T)

# inspect the data format
head(allele.freqs)
```

```
##   Locus_1 A_1 B_1        p_1       q_1 A_2 B_2       p_2       q_2
## 1 locus_1 126  30 0.80769231 0.1923077  26   6 0.8125000 0.1875000
## 2 locus_2  30 126 0.19230769 0.8076923   6  26 0.1875000 0.8125000
## 3 locus_3  22 138 0.13750000 0.8625000   8  22 0.2666667 0.7333333
## 4 locus_4  13 145 0.08227848 0.9177215   2  30 0.0625000 0.9375000
## 5 locus_5 145  15 0.90625000 0.0937500  28   4 0.8750000 0.1250000
## 6 locus_6  39 115 0.25324675 0.7467532   8  24 0.2500000 0.7500000
```

Bayescan output file:

```
# load in the Bayescan output data
setwd('~/Desktop/CSU_PopGen_Labs/Data/')
bayescan.out<-read.table('Final_Merged_Both_Sites_Bayescan_Input_fst.txt')

# inspect the data format
head(bayescan.out)
```

```
##        prob log10.PO.    qval      alpha       fst
## 1 0.0938188 -0.984925 0.89356 -0.0129100 0.0049603
## 2 0.0908180 -1.000500 0.90035 -0.0054275 0.0050024
## 3 0.0902180 -1.003600 0.90144  0.0138280 0.0052013
## 4 0.0938190 -0.984930 0.89356 -0.0117640 0.0049650
## 5 0.0808160 -1.055900 0.90774 -0.0056248 0.0049612
## 6 0.0864170 -1.024100 0.90559 -0.0095594 0.0049918
```

Preliminary analysis

Before launching into the F_{ST} outlier analysis, it is often useful to visualze the raw data. Looking at the allele frequencies may provide clues as to which loci (if any) are under selection. Loci that have different allele frequences in the two subpopulations are possible contenders for being under selection and having outlier F_{ST} values.

We will start by looking for loci with an allele frequencies with large differences between the site 1 and site 2 subpopulations. We will use the *abs()* function to calculate the absolute value of the difference in allele frequencies between the two populations, and search for alleles where that difference is $> 25\%$. We will also plot the results as a barplot.

```
# extract the loci that have more than 25% difference
# in frequency between the two populations
disparate.freq <- allele.freqs[which(abs(allele.freqs$p_1-allele.freqs$p_2)>0.25),]

# how many were there?
length(disparate.freq[,1])

par(mar=c(7, 4, 3, 2))
```

```
barplot(rbind(disparate.freq$p_1, disparate.freq$p_2),
        beside=T, col=c('red', 'blue'),
        names=disparate.freq$Locus_1,
        ylab='Frequency, Allele A',
        las=2, ylim=c(0,1))
legend(x=-1, y=1.06, legend=c('Site 1', 'Site 2'), fill=c('red', 'blue'),
       bty='n') # put the legend exactly where you want it by specifying x and y
```

Question 1. How many SNPs with disparate allele frequencies in the two subpopulations did you find?

Question 2. Do you expect any of these will have high and/or statistically significant F_{ST} values? How high should F_{ST} be before you believe that locus might be under selection?

Multiple Comparisions in Genetic Inference

NGS studies, including reduced-representation genome sequencing studies, can identify 1,000s of SNPs. When we calculate F_{ST} or any other descriptive statistic for each SNP, we also want to determine if those values are *statistically significant*; that is, do they differ from our null hypothesis?

Recall from basic statistics that we define a *p-value* as the probability of observing our data if the null hypothesis is true. It is possible to obtain a data sample that is consistent with our null hypothesis but highly unlikely to have been observed. When we reject our null hypothesis even when our data are consistent with it, we are making a Type I error. To control for this possiblity of making the wrong conclusion about our data's consistency with the null hypothesis, we infer statistical significance when our p-values are lower than our threshold for making a Type I error. This threshold is called the α-level (not to be confused with the α_i parameter in the Bayescan model!). The α level is typically 0.05.

Multiple comparisons increase the probability of a Type I error

Every time we make a comparison of a sample to a given null hypothesis with α=0.05, we have a 5% chance of rejecting our null hypothesis even when it is true. Over numerous comparisons to the same null hypothesis, that 5% chance of error compounds. If you make enough comparisons it is almost guaranteed that you will make a Type I error.

This is particularly problematic for analysis of individual loci in NGS studies where 1,000s of loci have been identified. We will make 2387 comparisons in conducting the F_{ST} outlier analysis.

Question 3. If we perform 2387 statistical tests – one for each SNP – and expect that 5% of our p-values would be statistically significant by chance alone, roughly how many Type I errors do you predict we would make?

Question 4. What might be the consequence of incorrectly inferring that some American dog tick SNPs have significantly high F_{ST} values when they really do not?

Simulated data show the consequence of multiple comparisons

We can draw 2387 samples from a Student's T distribution to represent possible test statistics obtained in statistical analysis. If all 2387 test statistics come from the same distribution they will all be consistent with the same null hypothesis and none should be statistically significant. But this is the equivalent of doing 2387 t-tests – will some yield *p-values* less than α? (Note: we do not do t-tests to analyze SNP data, but this simple example will highlight the dangers of multiple comparisons in even the simplest statistical context.)

```
# Draw sample data points (test statistics) from a simulated Student's T distribution.
# Degrees of freedom (df) is arbitrarily set.
simulated.values<-rt(length(bayescan.out[,1]), df=98)

# Determine the probability of making each observation randomly from
# the given distribution using the function dt().
# dt() gives the area under the curve for values
# equal to or more extreme than the observed value,
# otherwise known as the p-value.

simulated.p<-dt(simulated.values, df=1)
```

Question 5. In the case of the Student's T distribution we will state our null hypothesis as mean = 0. All of our simpulated p-values were calculated from data drawn from the same Student's T distribution with mean = 0 – all of our data should be consistent with the null hypothesis and none should be statistically significant. How many observations in the simulated data were made with less than 5% probability from our original distribution?

```
par(mfrow=c(1,2), mar=c(6,2,2,2))

# panel 1 shows the original values from the Student's T distribution
hist(simulated.values, main='', xlab='Test-statistic Distribution',
     las=1, freq=F, axes=F)
axis(side=1)

# panel 2 shows the distribution of p-values
# the gray area of the histogram represents the
# false-positive p-values < 0.05, which we would like to remove.
hist(simulated.p, main='', xlab='', ylab='', las=1,
     breaks=seq(0,0.4, 0.05), freq=F, col=c('gray', rep('white', 8)),
     axes=F)
axis(side=1)
mtext(side=1, line=3.5, text='p-value Distribution \nP(data|True Null Hypothesis)')

# length(simulated.p[which(simulated.p<0.05)]) counts
# the number of p-values less than alpha = 0.05
text(x=0.021, y=1.2, labels=length(simulated.p[which(simulated.p<0.05)]), cex=0.7)
```

Correcting multiple comparisons with *qvalues*.

We can detect and remove Type I errors in many ways, but the standard for genetic data is to perform a False Discovery Rate (FDR) correction by calculating a *q-value*. The *q-value* for a SNP will always be higher than the p-value, but if the allele frequencies for the SNP are truly significantly different from our null model of drift then the *q-value* should still be less than α.

We can illustrate the process of calculating *q-values* with our simulated data:

```
# perform the FDR correction with the qvalue() function:
simulated.q<-qvalue(simulated.p, lambda=0.2)

# plot the comparison
par(mfrow=c(1,2), mar=c(4,2,2,2))
```

```
# panel 1 shows the distribution of p-values
# the gray area of the histogram represents the
# false-positive p-values < 0.05, which we would like to remove.
hist(simulated.p, main='', xlab='', ylab='', las=1,
     breaks=seq(0,0.4,0.05), freq=F, col=c('gray', rep('white', 8)),
     axes=F)
axis(side=1)
mtext(side=1, line=2.5, text='p-value Distribution')
text(x=0.021, y=1.2, labels=length(simulated.p[which(simulated.p<0.05)]), cex=0.7)

# panel 2 shows the q-values (corrected p-values)
hist(simulated.q$qvalues, main='', ylab='', xlab='', las=1,
     breaks=seq(0,0.4,0.05), axes=F)
axis(side=1)
mtext(side=1, line=2.5, text='q-value Distribution')

# length(simulated.q$qvalues[which(simulated.q$qvalues<0.05)]) counts
# the number of q-values less than alpha = 0.05
text(x=0.021, y=100, labels=length(simulated.q$qvalues[which(simulated.q$qvalues<0.05)]), cex=0.7)
```

Question 6. How many simulated q-values are less than 0.05?

Bayescan Analysis & Results

Now that we understand how to properly control for the multiple comparisons that Bayescan will perform, it is time to look at the Bayescan results. Performing the F_{ST} calculation on the allele counts using Bayescan requires about one hour of supercomputer time. Instead of having all of you simultaneously access a supercomputer to run the same job, you are provided with the output of the Bayescan analysis on these data.

Bayescan also calculates the *q-value* for each F_{ST} estimate, so you will not need to calculate it manually (though now you know how should the need arise in the future).

Question 7. Plot the F_{ST} values from the Bayescan output against their corresponding *q-values* using the code below.

```
par(mar=c(5,6,4,2))
plot(bayescan.out$qval, bayescan.out$fst, pch=16, cex=1.5, cex.axis=1.5, xlim=c(0,1),
     ylim=c(0,0.15), bty='n', las=1, cex.lab=1.5, xlab='q-value', ylab='')
mtext(side=2, line=4, text=expression('F'['ST']), cex=1.5)
abline(v=0.1, col='red')
```

Question 8. How many SNPs have significant F_{ST} at the $\alpha = 0.1$ level? What are the corresponding F_{ST} values?

```
# Count the number significant at alpha = 0.1
sig<-bayescan.out$qval[bayescan.out$qval < 0.1]
length(sig)

# What is the corresponding FST value? (round to 3 digits)
fst<-bayescan.out$fst[bayescan.out$qval < 0.1]
round(fst, 3)
```

Question 9. Modify the code from Question 8 to count the number of SNPs with significant F_{ST} values at the $\alpha = 0.05$ level.

Question 10. Did your expectation about the number of SNPs with significant F_{ST} values based on the raw allele frequencies differ from the Bayescan results?

Question 11. In the introduction we described the Bayescan model as estimating parameters α_i and β_j to describe F_{ST}^{ij}. However, we have talked almost exclusively about the F_{ST}^{ij} estimates themselves. As you may have noticed, the Bayescan output also contains the α_{ij} estimate for each locus. Plot F_{ST}^{ij} vs α_{ij} and describe how these parameters are related. Is the relationship what you would expect?

```
par(mfrow=c(1,1), mar=c(5,5,2,2))
plot(bayescan.out$fst, bayescan.out$alpha, pch=16,
     xlab=expression('F'['ST']^'ij'), ylab=expression(alpha['ij']))
```

Question 12. Why do you think Bayescan does not report β_j?

Question 13. What do these results imply about the structuring of the two American dog tick subpopulations? Is it driven by genetic drift or natural selection?

References

Foll, Matthieu, and Oscar Gaggiotti. 2008. "A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: a Bayesian Perspective." *Genetics* 180 (2): 977–93. doi:10.1534/genetics.108.092221. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2567396/&tool=pmcentrez/&rendertype=abstract.