Hardy-Weinberg Equilibrium Activity: Teacher Version

Norah Saarman

9/5/2021

Learning goals

Understand:

- what is the Hardy-Weinberg (HW) principle
- assumptions and utility in population genetics studies
- how population structure, admixture, finite populations, and data source can impact interpretation of tests for HW equilibrium

Be able to:

- test for HW equilibrium
- estimate locus-by-locus FIS and mean FIS
- identify potential causes of deviations from HW equilibrium
- distinguish the Wahlund effect from inbreeding (the mating of individuals that are closely related through common ancestry)

Learning self-assessment questions (before):

Describe in your own words, what is the Hardy-Weinberg principle?

How might this principle be useful to you in understanding the health of a population of conservation concern?

Can you think of a scenario where it might be difficult to interpret what is the level of inbreeding in a population sample you have genotyped based on estimates of heterozygosity and FIS alone?

Background

The Hardy-Weinberg (HW) principle

At Hardy-Weinberg equilibrium, (i) allele frequencies in a population will remain constant indefinitely, and (ii) genotypic proportions occur at Hardy-Weinberg proportions in the population as determined by the "square law".

What is the "square law"? Think of the Punnett Square you learned about in introductory genetics. Consider a single locus with two alleles A1 and A2.

Let:
p = frequency of A1 allele
q = frequency of A2 allele

Three genotypes are thus possible: A1A1, A1A2, and A2A2.

Let:
P = frequency of A1A1 homozygote
H = frequency of A1A2 heterozygote
Q = frequency of A2A2 homozygote

From the frequencies, we can estimate allele frequencies:
$p = P + \frac{1}{2} H$
$q = Q + \frac{1}{2} H$

These frequencies will sum to 1, since there are only 2 alleles present:
$p + q = 1$

If mating occurs at random in the population, what will be the frequencies of A1 and A2 in the next generation? It depends on the frequencies of each genotype in the parents:

	Female genotypes		
Male genotypes	A_1A_1 (P)	A_1A_2 (H)	A_2A_2 (Q)
A_1A_1 (P)	P^2	PH	PQ
A_1A_2 (H)	PH	H^2	HQ
A_2A_2 (Q)	PQ	HQ	Q^2

The progeny produced by this set of matings would be:

		Progeny		
Mating	Total Frequency	A_1A_1 (P)	A_1A_2 (H)	A_2A_2 (Q)
$A_1A_1 \times A_1A_1$	P^2	P^2		
$A_1A_1 \times A_1A_2$	2PH	PH	PH	
$A_1A_1 \times A_2A_2$	2PQ		2PQ	
$A_1A_2 \times A_1A_2$	H^2	$H^2/4$	$H^2/2$	$H^2/4$
$A_1A_2 \times A_2A_2$	HQ		HQ	HQ
$A_2A_2 \times A_2A_2$	Q^2			Q^2
	$= (P + H + Q)^2$	$= (P + H/2)^2$	$= 2(P + H/2)^* (Q + H/2)$	$= (Q + H/2)^2$
	$= 1$	$= p^2$	$= 2pq$	$= q^2$

The frequencies of the alleles have not changed, and the genotypic proportions are determined by the "square law". For two alleles, genotypic proportions are given by expanding the term (p+q)^2.

Assumptions of the HW principle

The reason the Hardy-Weinberg equilibrium is so important is that for evolutionary change to occur in a population, it is necessary for one or more specific assumptions to be violated. We can use information about

the way the population deviates from HW expectations to understand which assumptions have been violated (and thus the relative importance of different forces of evolutionary change). What are these assumptions?

1) Generations are discrete (i.e. non-overlapping)
2) The species is diploid
3) Reproduction is sexual
4) The gene being considered has 2 alleles
5) Allele frequencies are the same in males and females
6) Mating is random
7) The population size is infinite (i.e. no genetic drift)
8) There is no migration (gene flow)
9) There is no mutation
10) There is no selection

The fixation index (FIS) and interpretation

Since the Hardy-Weinberg principle predicts that no evolution will occur unless one of the above assumptions is violated, it is often useful to test if a population is in HW equilibrium and use information about the way the population deviates from HW expectations to understand which assumptions have been violated. In other words, deviations from HW expectations can help to determine the relative importance of random drift, migration, mutation, and natural selection in affecting the frequency of genetic polymorphism in natural populations.

FIS (Nei, 1987) provides a simple way of summarizing in what direction the frequency of genetic polymorphism in natural populations deviate from HW equilibrium. FIS is based on a comparison of observed heterozygosity (Hobs) and the HW expected heterozygosity given the allele frequencies in the population:

FIS = 1 - (Hobs/Hexp)

Negative FIS indicates a homozygote deficit and heterozygote excess. Some of many potential causes of heterozygote excess include:

- Small population size, this is because allele frequencies are likely to differ between sexes just due to chance.
- Negative assortative mating when reproduction occurs between individuals bearing phenotypes more dissimilar than by chance.
- Heterozygote advantage, something that sometimes occurs in hybrid zones
- Selection, this can occur in cases of balancing selection, but usually occurs in only a small proportion of the genome.
- See the list of assumptions and let your mind run!

Positive FIS indicates a homozygote excess and heterozygote deficit. Some of many potential causes of heterozygote deficit include:

- Inbreeding, this is because matings between close relatives are more likely to result in pairing even rare alleles in homozygote form.
- Population structure, this is because of the "Wahlund effect", where two or more subpopulations are in Hardy-Weinberg equilibrium but have different allele frequencies such that the overall heterozygosity is reduced compared to if the whole population was in equilibrium.
- Selection, this is because alleles that have a selective advantage are more likely to be in homozygous than heterozygous form. Note that these alleles are also more likely to go to fixation unless there is clinal variation, frequency-dependence, or other processes that maintain both alleles.
- Technical issues, for example miss-scoring of heterozygotes as homozygotes because of low next-gen sequencing read depth.
- See the list of assumptions and think through the logical consequences!

In-Class Activity

Part 1: Four Scenarios.

You will be split into 4 working groups (breakout rooms) A-D. Each group will be blindly assigned one of four datasets, and it is your goal to perform several analyses on these datasets and identify which dataset your group received.

Four Scenarios: 1) Marten dataset from the admixture zone in Idaho 2) Marten dataset from a healthy population north of the admixture zone 3) Bull trout SNP dataset with very small Ne 4) Rainbow trout SNP dataset from a genome-wide association study

Within your breakout group, determine which of the scenarios you have, use the R package "hierfstat" following the code provided in Part 1 of HW_student.Rmd to estimate and plot basic statistics including FIS, and answer the following question (also listed in the Rmd file, feel free to type into the Rmd save it for your records):

```
# import data in genepop format as "myData"
myData <- read.genepop("HW_FourScenarios.gen", ncode = 2 , quiet = TRUE)

# fill in "pop" slot of genind object with proper dataset A-D
pop_list <- as.factor(c(rep("A",25),rep("B",25),rep("C",25),rep("D",25)))
myData@pop <- pop_list

# use hierfstat to get basic stats and FIS per locus, "E" is for example, change to "A", "B", "C", or "
statsA <- basic.stats(myData[myData@pop == "A"])
statsB <- basic.stats(myData[myData@pop == "B"])
statsC <- basic.stats(myData[myData@pop == "C"])
statsD <- basic.stats(myData[myData@pop == "D"])
```

Students will want to take a look at the basic stats output.

Question 3: When you look at the per-locus FIS, does anything stand out to you? Are there any loci that appear to be outliers?

Question 4: If so, what are some possible interpretations of what may have caused this deviation for expected levels of heterozygosity?

Question 5: Do you think there is an overall heterozygote excess or defecit (or neither) in this dataset? What are some possible interpretations of this result?

Question 6: Is there an obvious alternative interpretation of the pattern of FIS you observed that you are left unable to distinguish with the available information? What might you do to test this alternative hypothesis?

When you have completed these questions, as a group prepare a few sentences that describe which scenario you think your group was assigned, and why you think this. Then return to the main room to share with the other groups.

Part 2: The Wahlund effect.

How can we distinguish population structure from inbreeding (high overall FIS)? Together, we will produce two different simulated datasets, then you will again break off into groups to complete some analysis and answer some questions to allow you to distinguish population structure from inbreeding in an idealized situation. This activity should also give you some strategies to consider in the real world when you encounter patterns of deviation from Hardy-Weinberg equilibrium.

Together, we will look at the PCA and clustering plots from two different simulated datasets: 1) high overall FIS because of inbreeding 2) high overall FIS because of population structure

```
# simulated inbred dataset
iSim <- sim.genot(size=100,nbal=8,nbloc=15,nbpop=1,N=1000,mut=0.001,f=0.2) # simulate
colnames(iSim) <- NULL # replace column names with null to make adegenet happy
iData <- df2genind(iSim[-1], ncode=1) # convert to adegenet genind object
iData@pop <- as.factor(rep("inbred",100)) # fill in "pop" slot of genind object to make hierfstat happy

# simulated population structure dataset
sSim <- sim.genot.metapop.t(size=50,nbal=8,nbloc=15,nbpop=2,N=50,mig=matrix(c(1,0,0,1),nrow=2,byrow=TRU
colnames(sSim) <- NULL # replace column names with null to make adegenet happy
sData <- df2genind(sSim[-1], ncode=1) # convert to adegenet genind object
sData@pop <- as.factor(rep("structured",100)) # fill in "pop" slot of genind object to make hierfstat h
```

Question 7: What are the major parameter choices for the simulation of both datasets? Is there anything you would change with less limited computation time?

Within your breakout group, use the R package "hierfstat" following the code beelow to estimate and plot basic statistics including FIS, and visualize the genetic structure present in the simulated datasets. Then answer the following question (also listed in the Rmd file, feel free to type into the Rmd save it for your records):

```
# basic stats
iStats <- basic.stats(iData)
iStats
```

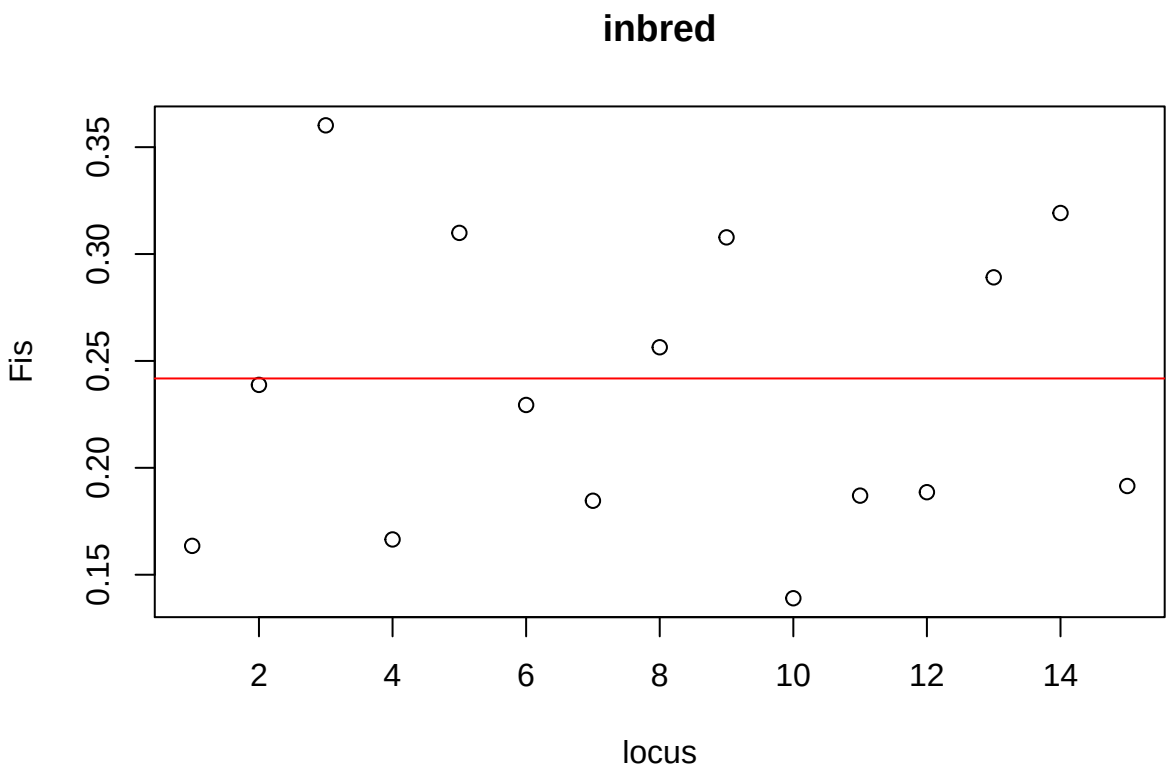
```
## $perloc
##         Ho     Hs     Ht Dst Htp Dstp Fst Fstp    Fis Dest
## loc01 0.37 0.4423 0.4423   0 NaN  NaN   0  NaN 0.1635  NaN
## loc02 0.23 0.3022 0.3022   0 NaN  NaN   0  NaN 0.2388  NaN
## loc03 0.52 0.8127 0.8127   0 NaN  NaN   0  NaN 0.3602  NaN
## loc04 0.51 0.6119 0.6119   0 NaN  NaN   0  NaN 0.1665  NaN
## loc05 0.41 0.5941 0.5941   0  NA   NA   0  NaN 0.3099   NA
## loc06 0.44 0.5710 0.5710   0 NaN  NaN   0  NaN 0.2294  NaN
## loc07 0.33 0.4047 0.4047   0 NaN  NaN   0  NaN 0.1846  NaN
## loc08 0.26 0.3496 0.3496   0  NA   NA   0  NaN 0.2564   NA
## loc09 0.48 0.6935 0.6935   0 NaN  NaN   0  NaN 0.3078  NaN
## loc10 0.42 0.4878 0.4878   0 NaN  NaN   0  NaN 0.1390  NaN
## loc11 0.54 0.6642 0.6642   0  NA   NA   0  NaN 0.1870   NA
## loc12 0.60 0.7394 0.7394   0 NaN  NaN   0  NaN 0.1886  NaN
## loc13 0.43 0.6049 0.6049   0 NaN  NaN   0  NaN 0.2891  NaN
## loc14 0.48 0.7051 0.7051   0 NaN  NaN   0  NaN 0.3192  NaN
## loc15 0.53 0.6556 0.6556   0  NA   NA   0  NaN 0.1915   NA
```

```
##
## $overall
##     Ho     Hs     Ht    Dst    Htp   Dstp    Fst   Fstp    Fis   Dest
## 0.4367 0.5759 0.5759 0.0000    NaN    NaN 0.0000    NaN 0.2418    NaN
```

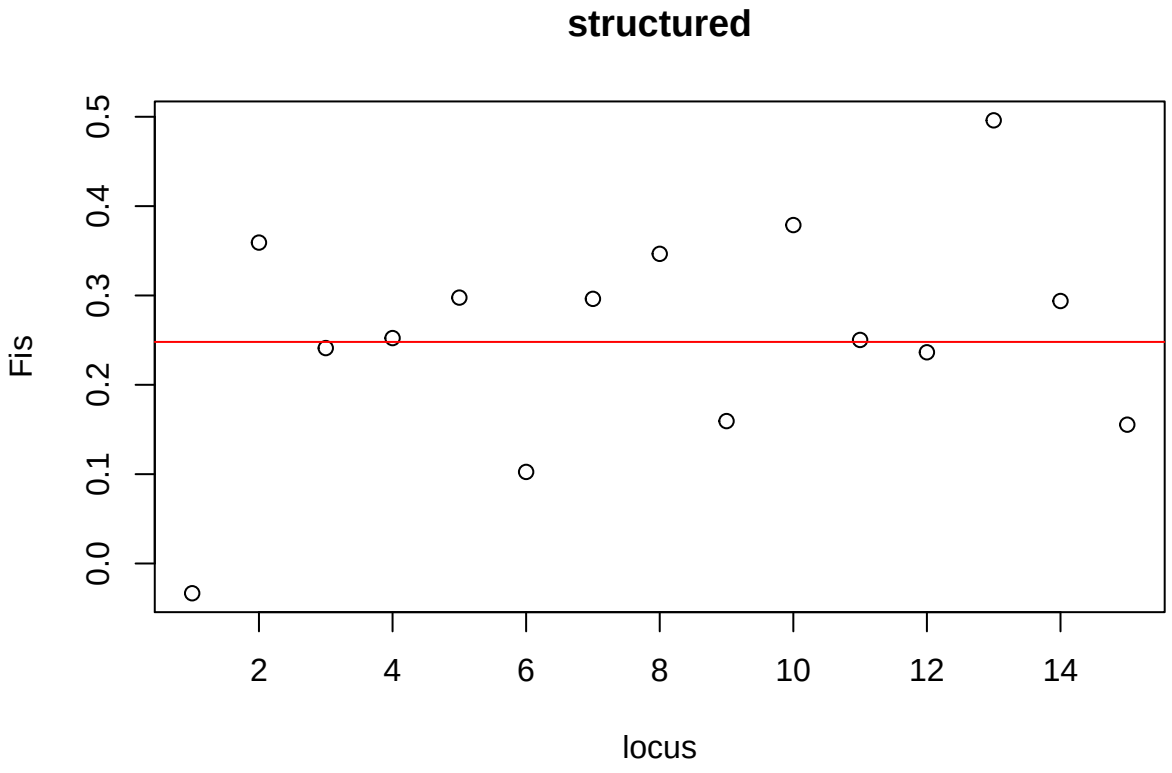
```
sStats <- basic.stats(sData)
sStats
```

```
## $perloc
##         Ho     Hs     Ht Dst Htp Dstp Fst Fstp     Fis Dest
## loc01 0.78 0.7549 0.7549   0  NA   NA   0  NaN -0.0333   NA
## loc02 0.45 0.7023 0.7023   0  NA   NA   0  NaN  0.3592   NA
## loc03 0.53 0.6983 0.6983   0 NaN  NaN   0  NaN  0.2411  NaN
## loc04 0.58 0.7757 0.7757   0  NA   NA   0  NaN  0.2523   NA
## loc05 0.54 0.7688 0.7688   0 NaN  NaN   0  NaN  0.2976  NaN
## loc06 0.72 0.8022 0.8022   0 NaN  NaN   0  NaN  0.1025  NaN
## loc07 0.49 0.6962 0.6962   0 NaN  NaN   0  NaN  0.2962  NaN
## loc08 0.50 0.7653 0.7653   0  NA   NA   0  NaN  0.3466   NA
## loc09 0.69 0.8208 0.8208   0  NA   NA   0  NaN  0.1594   NA
## loc10 0.31 0.4990 0.4990   0 NaN  NaN   0  NaN  0.3788  NaN
## loc11 0.49 0.6535 0.6535   0 NaN  NaN   0  NaN  0.2502  NaN
## loc12 0.58 0.7596 0.7596   0 NaN  NaN   0  NaN  0.2364  NaN
## loc13 0.33 0.6548 0.6548   0 NaN  NaN   0  NaN  0.4960  NaN
## loc14 0.50 0.7080 0.7080   0 NaN  NaN   0  NaN  0.2938  NaN
## loc15 0.68 0.8052 0.8052   0  NA   NA   0  NaN  0.1554   NA
##
## $overall
##     Ho     Hs     Ht    Dst    Htp   Dstp    Fst   Fstp    Fis   Dest
## 0.5447 0.7243 0.7243 0.0000    NaN    NaN 0.0000    NaN 0.2480    NaN
```

```
# FIS plots
plot(iStats$perloc$Fis,xlab = "locus", ylab = "Fis", main = "inbred")
abline(h=iStats$overall[9],col="red")
```



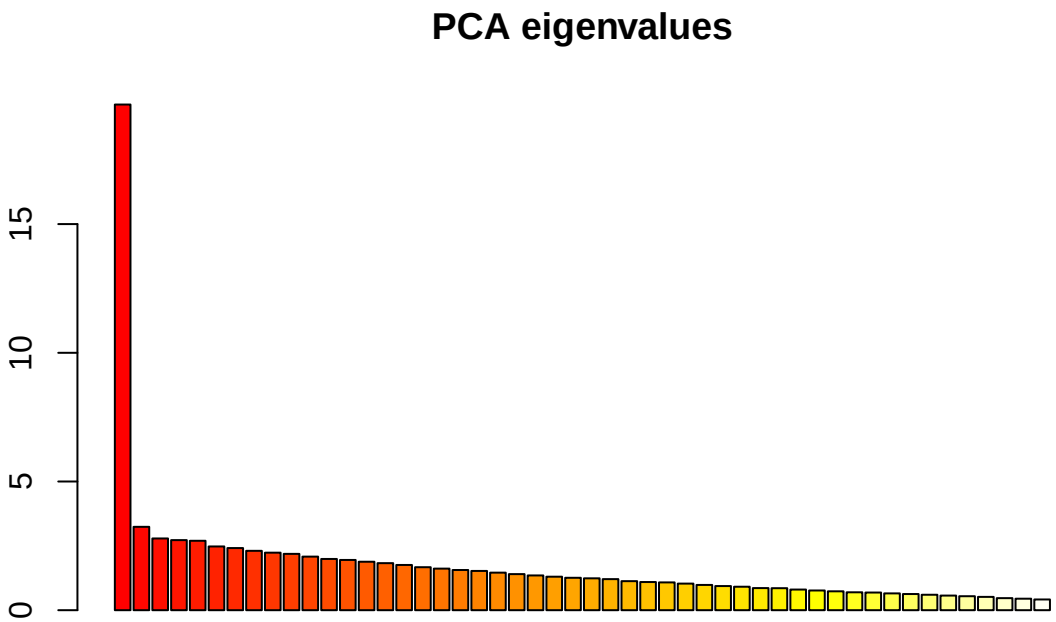

```
plot(sStats$perloc$Fis,xlab = "locus", ylab = "Fis", main = "structured")
abline(h=sStats$overall[9],col="red")
```



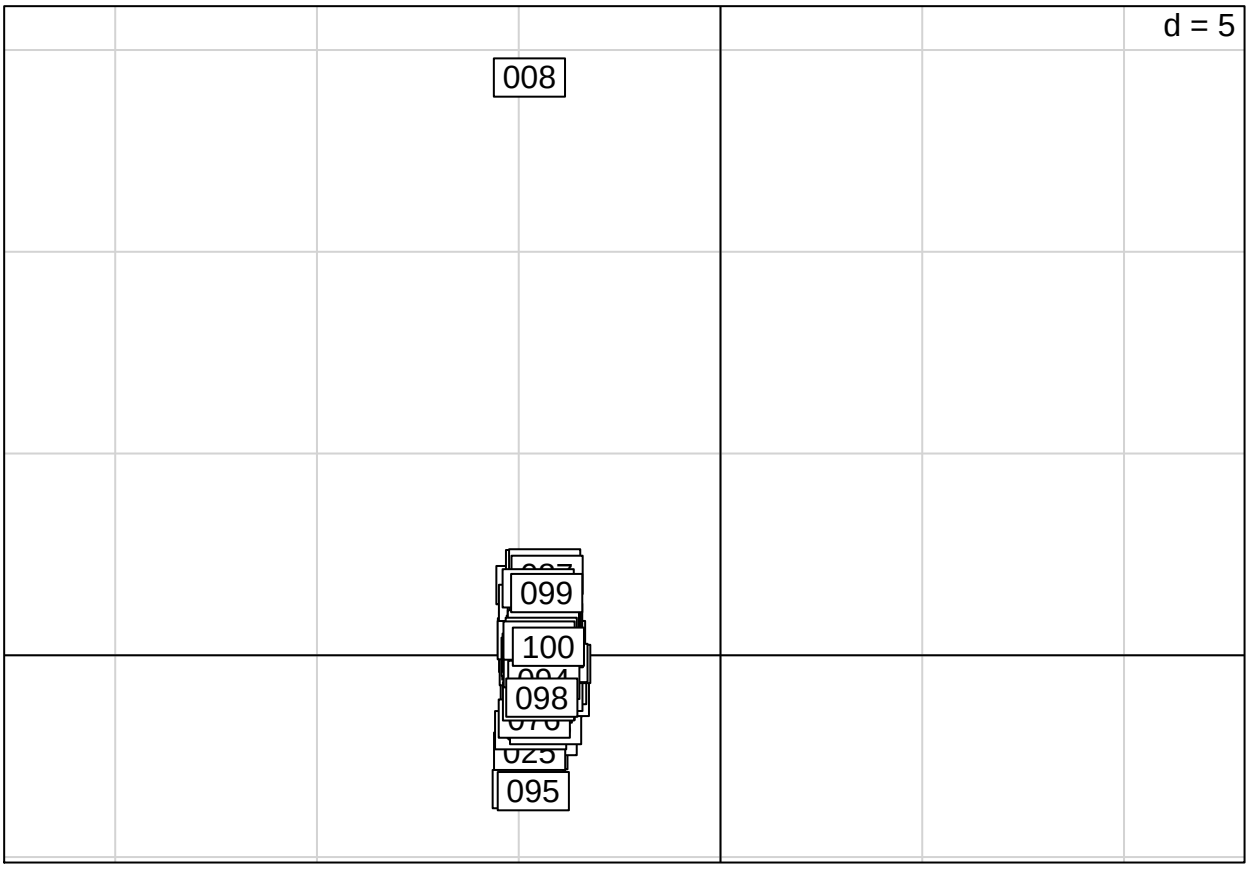
Question 8: Do you think there is an overall heterozygote excess or defecit in these two datasets? What do you think caused this excess/deficit in each case?

Question 9: Are there any obvious differences between the output for these different simulations? What can we do to distinguish between the possible causes (population structure and inbreeding)?

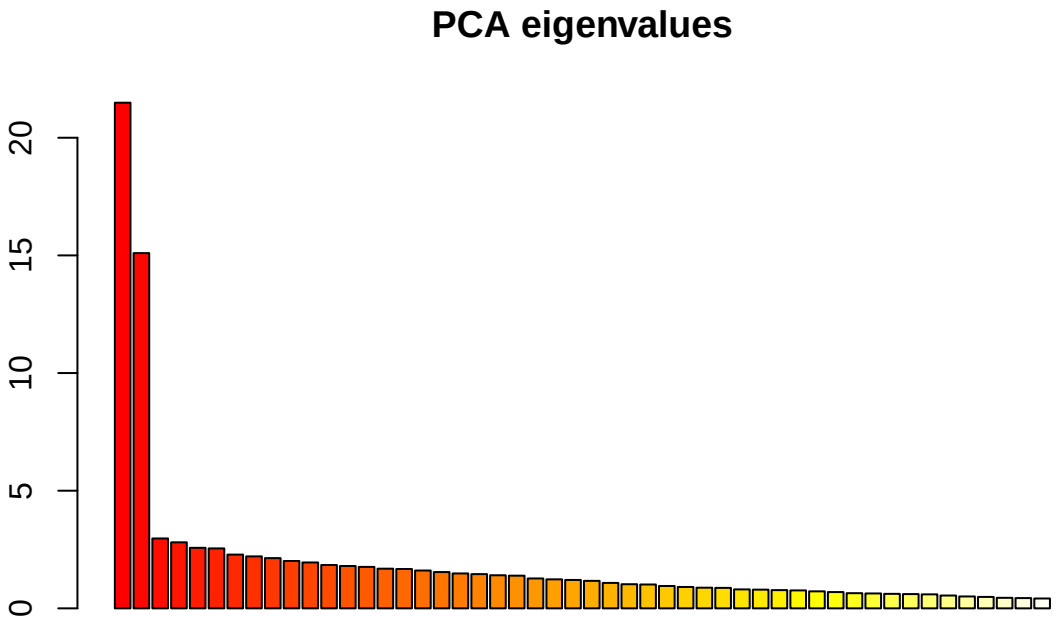
```
# PCA for inbred
iPCA <- dudi.pca(iData,cent=FALSE,scale=TRUE,scannf=FALSE,nf=4)
barplot(iPCA$eig[1:50],main="PCA eigenvalues", col=heat.colors(50)) # view eigenvalues
```



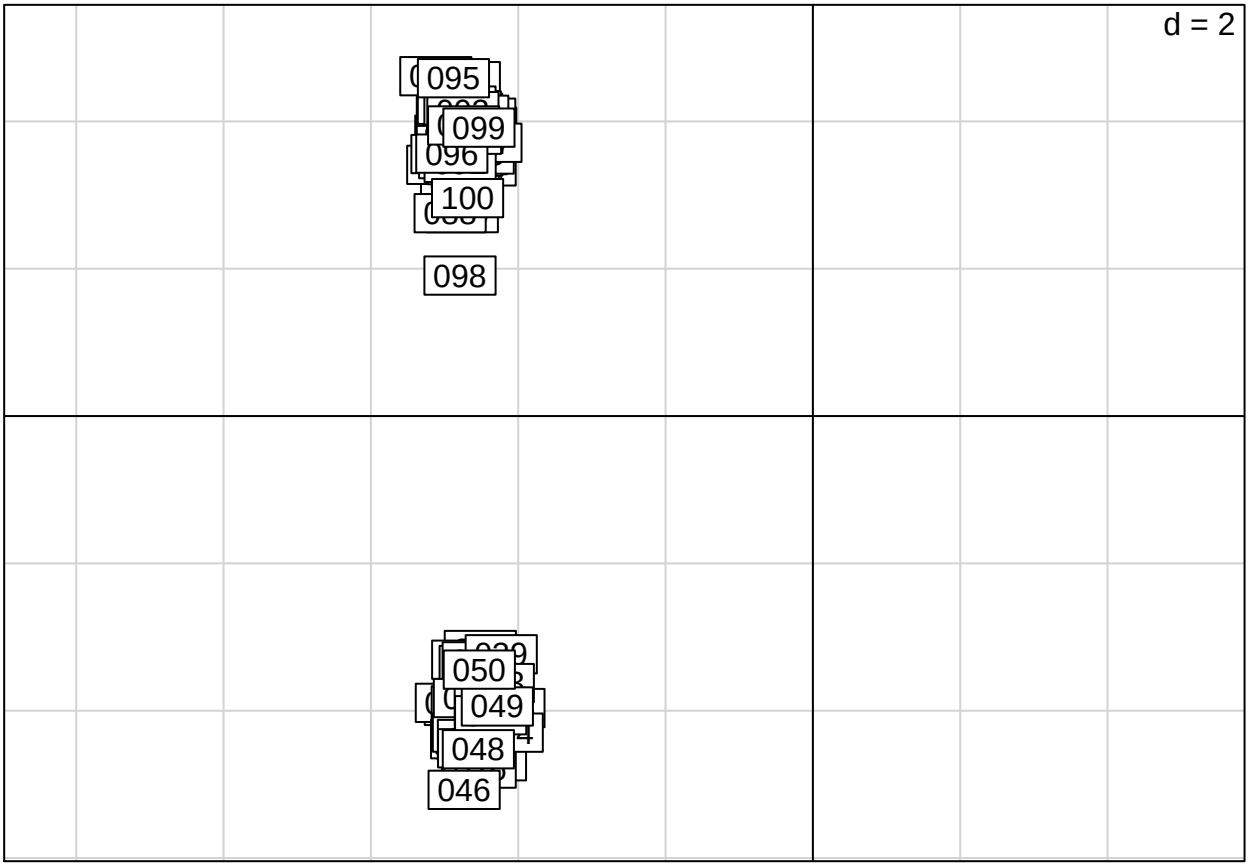

```
s.label(iPCA$li) # plot eigenvectors
```




```
# PCA for structured
sPCA <- dudi.pca(sData,cent=FALSE,scale=TRUE,scannf=FALSE,nf=4)
barplot(sPCA$eig[1:50],main="PCA eigenvalues", col=heat.colors(50)) # view eigenvalues
```



```
s.label(sPCA$li)
```



Question 10: Do you see the population structure in the one dataset and not the other? What other analysis would you want to do if this were your own thesis to prove to yourself there is population structure rather than inbreeding?

One idea here is to estimate FIS again after separating these putative clusters. Does the signal of FIS go away?

```
grpA <- sData[sPCA$li$Axis2>0]
grpB <- sData[sPCA$li$Axis2<0]

# basic stats
aStats <- basic.stats(grpA)
aStats
```

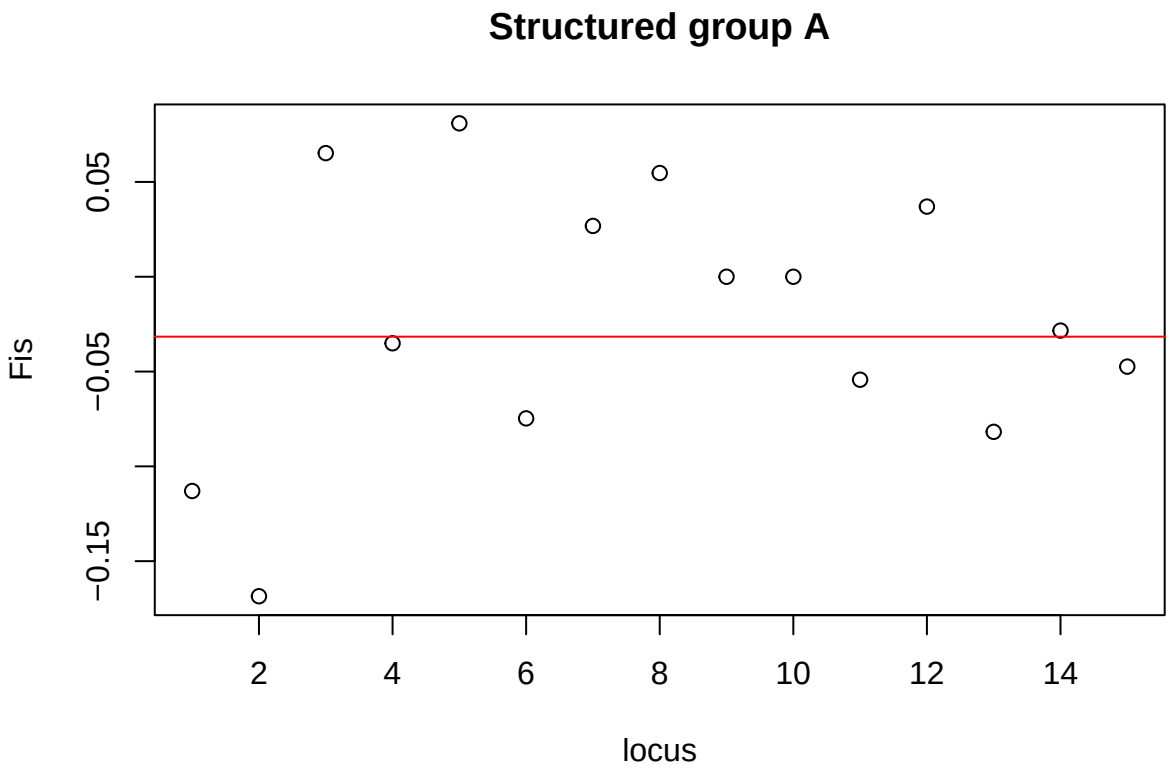
```
## $perloc
##         Ho     Hs     Ht Dst Htp Dstp Fst Fstp     Fis Dest
## loc01 0.78 0.7008 0.7008   0 NaN  NaN   0  NaN -0.1130  NaN
## loc02 0.62 0.5306 0.5306   0 NaN  NaN   0  NaN -0.1685  NaN
## loc03 0.50 0.5349 0.5349   0 NaN  NaN   0  NaN  0.0652  NaN
## loc04 0.56 0.5410 0.5410   0 NaN  NaN   0  NaN -0.0351  NaN
## loc05 0.38 0.4135 0.4135   0 NaN  NaN   0  NaN  0.0809  NaN
## loc06 0.74 0.6886 0.6886   0 NaN  NaN   0  NaN -0.0747  NaN
## loc07 0.20 0.2055 0.2055   0 NaN  NaN   0  NaN  0.0268  NaN
## loc08 0.36 0.3808 0.3808   0 NaN  NaN   0  NaN  0.0547  NaN
## loc09 0.74 0.7400 0.7400   0 NaN  NaN   0  NaN  0.0000  NaN
```

```
## loc10 0.02 0.0200 0.0200   0 NaN  NaN   0  NaN  0.0000  NaN
## loc11 0.86 0.8157 0.8157   0 NaN  NaN   0  NaN -0.0543  NaN
## loc12 0.52 0.5400 0.5400   0 NaN  NaN   0  NaN  0.0370  NaN
## loc13 0.34 0.3143 0.3143   0 NaN  NaN   0  NaN -0.0818  NaN
## loc14 0.76 0.7390 0.7390   0 NaN  NaN   0  NaN -0.0284  NaN
## loc15 0.74 0.7065 0.7065   0 NaN  NaN   0  NaN -0.0474  NaN
##
## $overall
##      Ho      Hs      Ht     Dst     Htp    Dstp     Fst    Fstp     Fis    Dest
##  0.5413  0.5247  0.5247  0.0000     NaN     NaN  0.0000     NaN -0.0316     NaN
```

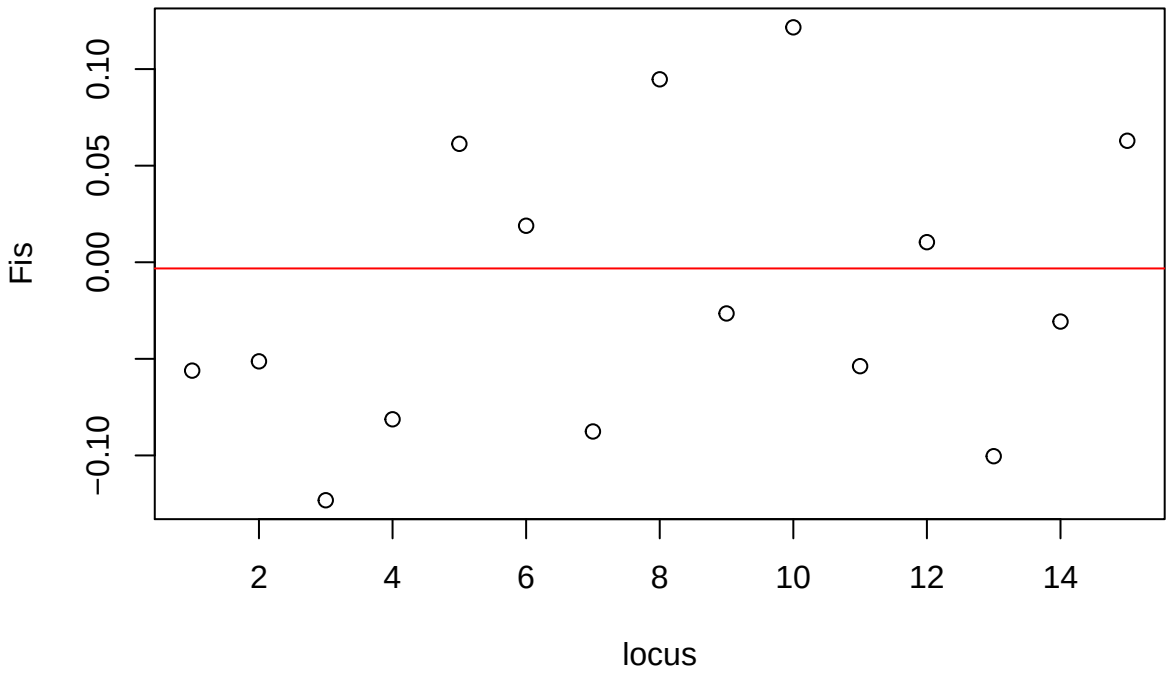
```
bStats <- basic.stats(grpB)
bStats
```

```
## $perloc
##         Ho     Hs     Ht Dst Htp Dstp Fst Fstp     Fis Dest
## loc01 0.78 0.7386 0.7386   0 NaN  NaN   0  NaN -0.0561  NaN
## loc02 0.28 0.2663 0.2663   0 NaN  NaN   0  NaN -0.0513  NaN
## loc03 0.56 0.4986 0.4986   0  NA   NA   0  NaN -0.1232   NA
## loc04 0.60 0.5549 0.5549   0 NaN  NaN   0  NaN -0.0813  NaN
## loc05 0.70 0.7457 0.7457   0 NaN  NaN   0  NaN  0.0613  NaN
## loc06 0.70 0.7135 0.7135   0 NaN  NaN   0  NaN  0.0189  NaN
## loc07 0.78 0.7171 0.7171   0  NA   NA   0  NaN -0.0876   NA
## loc08 0.64 0.7069 0.7069   0  NA   NA   0  NaN  0.0947   NA
## loc09 0.64 0.6235 0.6235   0 NaN  NaN   0  NaN -0.0265  NaN
## loc10 0.60 0.6831 0.6831   0 NaN  NaN   0  NaN  0.1216  NaN
## loc11 0.12 0.1139 0.1139   0 NaN  NaN   0  NaN -0.0538  NaN
## loc12 0.64 0.6467 0.6467   0  NA   NA   0  NaN  0.0104   NA
## loc13 0.32 0.2908 0.2908   0 NaN  NaN   0  NaN -0.1004  NaN
## loc14 0.24 0.2329 0.2329   0 NaN  NaN   0  NaN -0.0307  NaN
## loc15 0.62 0.6616 0.6616   0 NaN  NaN   0  NaN  0.0629  NaN
##
## $overall
##      Ho      Hs      Ht     Dst     Htp    Dstp     Fst    Fstp     Fis    Dest
##  0.5480  0.5463  0.5463  0.0000     NaN     NaN  0.0000     NaN -0.0032     NaN
```

```
# FIS plots
plot(aStats$perloc$Fis,xlab = "locus", ylab = "Fis", main = "Structured group A")
abline(h=aStats$overall[9],col="red")
```




```
plot(bStats$perloc$Fis,xlab = "locus", ylab = "Fis", main = "Structured group B")
abline(h=bStats$overall[9],col="red")
```



Question 11: After separating the dataset into two clusters and estimating FIS again for each putative cluster, does the signal of high FIS go away? Why/why not?

To illustrate what would happen if the underlying cause really were FIS not structured populations, we can run the same analysis with the inbred dataset:

```
grpC <- iData[iPCA$li$Axis2>0]
grpD <- iData[iPCA$li$Axis2<0]

# basic stats
cStats <- basic.stats(grpC)
cStats
```

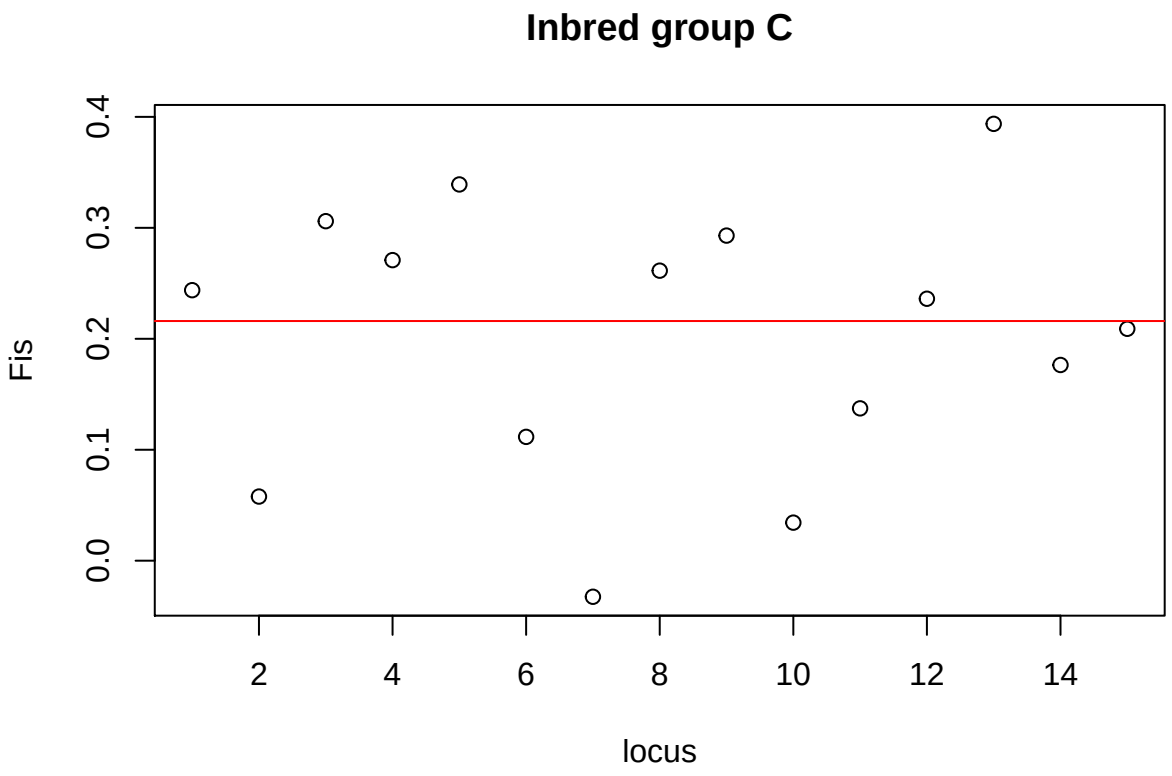
```
## $perloc
##           Ho     Hs     Ht Dst Htp Dstp Fst Fstp     Fis Dest
## loc01 0.2500 0.3306 0.3306   0  NA   NA   0  NaN  0.2438   NA
## loc02 0.2500 0.2653 0.2653   0 NaN  NaN   0  NaN  0.0578  NaN
## loc03 0.5455 0.7859 0.7859   0  NA   NA   0  NaN  0.3060   NA
## loc04 0.5000 0.6858 0.6858   0  NA   NA   0  NaN  0.2709   NA
## loc05 0.3864 0.5846 0.5846   0 NaN  NaN   0  NaN  0.3391  NaN
## loc06 0.5682 0.6395 0.6395   0 NaN  NaN   0  NaN  0.1116  NaN
## loc07 0.3864 0.3742 0.3742   0 NaN  NaN   0  NaN -0.0325  NaN
## loc08 0.3636 0.4923 0.4923   0  NA   NA   0  NaN  0.2614   NA
## loc09 0.5000 0.7072 0.7072   0 NaN  NaN   0  NaN  0.2930  NaN
## loc10 0.4318 0.4471 0.4471   0 NaN  NaN   0  NaN  0.0343  NaN
## loc11 0.5909 0.6850 0.6850   0 NaN  NaN   0  NaN  0.1373  NaN
## loc12 0.5909 0.7735 0.7735   0  NA   NA   0  NaN  0.2361   NA
## loc13 0.2955 0.4873 0.4873   0  NA   NA   0  NaN  0.3937   NA
## loc14 0.5455 0.6623 0.6623   0 NaN  NaN   0  NaN  0.1764  NaN
## loc15 0.5682 0.7183 0.7183   0  NA   NA   0  NaN  0.2090   NA
##
## $overall
##     Ho     Hs     Ht    Dst    Htp   Dstp    Fst   Fstp    Fis   Dest
## 0.4515 0.5759 0.5759 0.0000    NaN    NaN 0.0000    NaN 0.2160    NaN
```

```
dStats <- basic.stats(grpD)
dStats
```

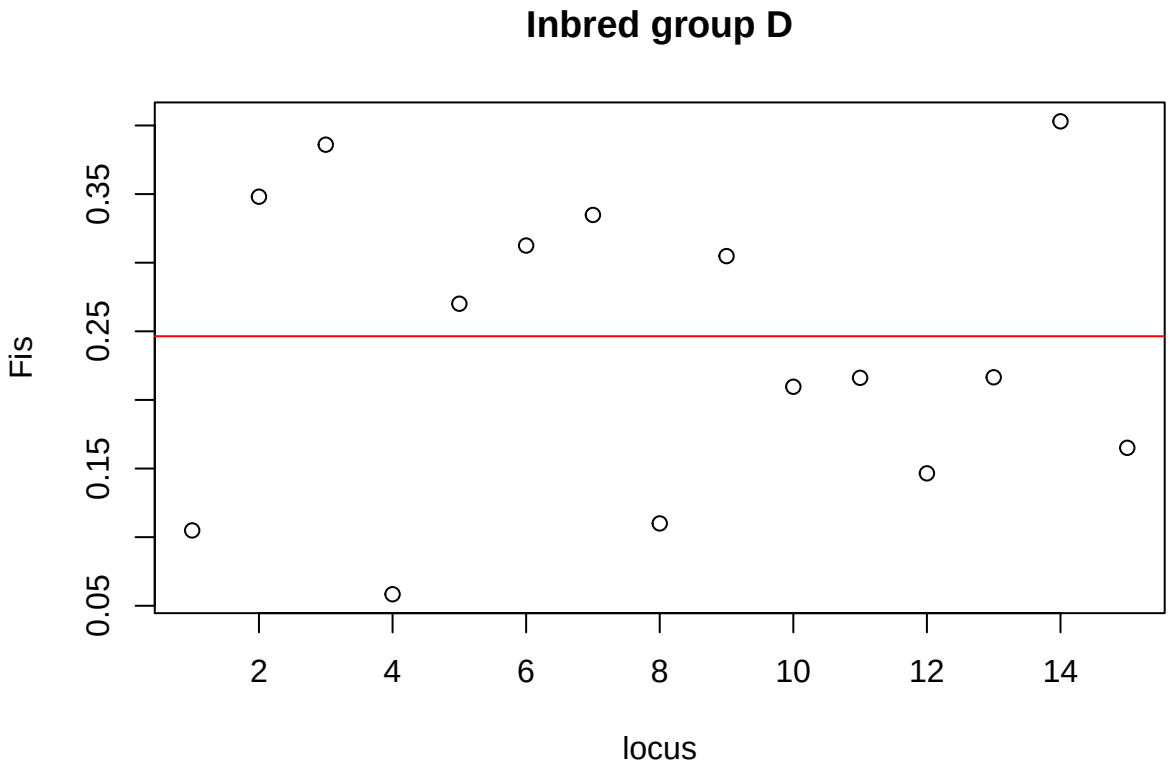
```
## $perloc
##           Ho     Hs     Ht Dst Htp Dstp Fst Fstp    Fis Dest
## loc01 0.4643 0.5187 0.5187   0 NaN  NaN   0  NaN 0.1049  NaN
## loc02 0.2143 0.3287 0.3287   0  NA   NA   0  NaN 0.3481   NA
## loc03 0.5000 0.8143 0.8143   0  NA   NA   0  NaN 0.3860   NA
## loc04 0.5179 0.5500 0.5500   0 NaN  NaN   0  NaN 0.0584  NaN
## loc05 0.4286 0.5872 0.5872   0 NaN  NaN   0  NaN 0.2701  NaN
## loc06 0.3393 0.4935 0.4935   0  NA   NA   0  NaN 0.3125   NA
## loc07 0.2857 0.4295 0.4295   0  NA   NA   0  NaN 0.3348   NA
## loc08 0.1786 0.2006 0.2006   0  NA   NA   0  NaN 0.1100   NA
## loc09 0.4643 0.6679 0.6679   0  NA   NA   0  NaN 0.3048   NA
## loc10 0.4107 0.5196 0.5196   0 NaN  NaN   0  NaN 0.2096  NaN
## loc11 0.5000 0.6378 0.6378   0  NA   NA   0  NaN 0.2161   NA
## loc12 0.6071 0.7114 0.7114   0  NA   NA   0  NaN 0.1465   NA
```

```
## loc13 0.5357 0.6838 0.6838   0 NaN  NaN   0  NaN 0.2165  NaN
## loc14 0.4286 0.7179 0.7179   0  NA   NA   0  NaN 0.4030   NA
## loc15 0.5000 0.5989 0.5989   0 NaN  NaN   0  NaN 0.1651  NaN
##
## $overall
##     Ho     Hs     Ht    Dst    Htp   Dstp    Fst   Fstp    Fis   Dest
## 0.4250 0.5640 0.5640 0.0000    NaN    NaN 0.0000    NaN 0.2464    NaN
```

```
# FIS plots
plot(cStats$perloc$Fis,xlab = "locus", ylab = "Fis", main = "Inbred group C")
abline(h=cStats$overall[9],col="red")
```




```
plot(dStats$perloc$Fis,xlab = "locus", ylab = "Fis", main = "Inbred group D")
abline(h=dStats$overall[9],col="red")
```



##

Question 12: This time, after separating the dataset artificially into two clusters and estimating FIS again for each putative cluster, does the signal of high FIS go away? Why/why not?

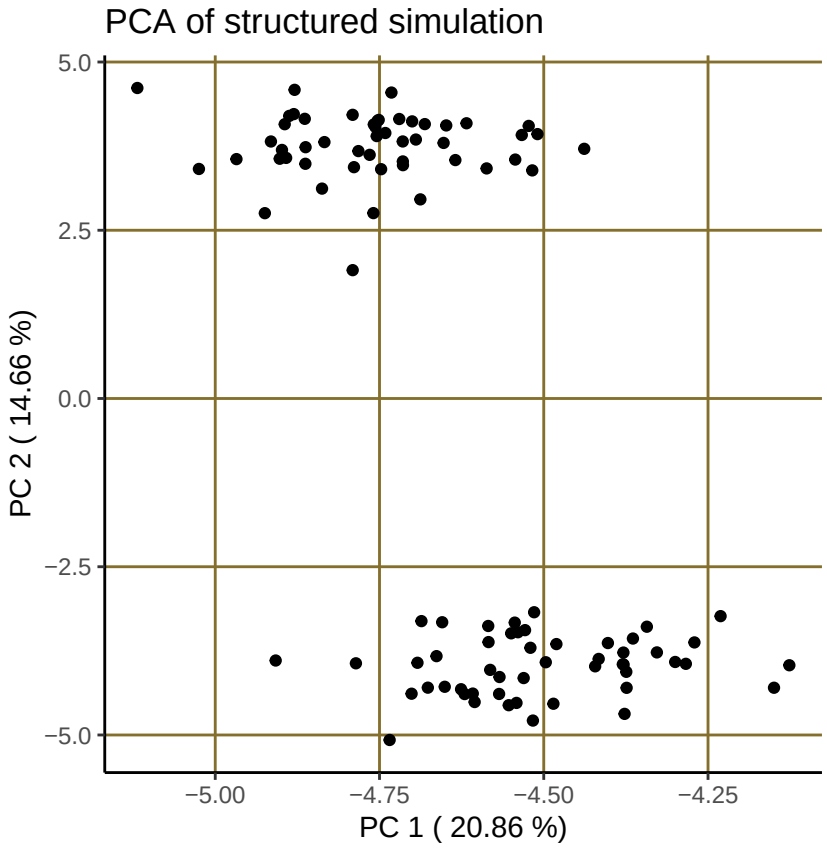
Question 13: Can you imagine a scenario where it may be difficult to distinguish the cause of a high value of FIS? What might this be?

```
# Assign the PCA you want to plot: Structured simulation
pca1 <- sPCA

# calculate percent variance of each component:
pc1 <- round(pca1$eig[1]/sum(pca1$eig)*100,digits=2)
pc2 <- round(pca1$eig[2]/sum(pca1$eig)*100,digits=2)

# define what you want to plot and create a dataframe:
PC1 <- pca1$li[,1]
PC2 <- pca1$li[,2]
df <- data.frame(PC1,PC2)

# plot:
ggplot(data = df, aes(PC1,PC2),)+
  xlab(paste("PC 1 (",pc1,"%)"))+
  ylab(paste("PC 2 (",pc2,"%)"))+
  geom_point(size=1.5)+
  ggtitle("PCA of structured simulation")+
  theme(panel.grid.major = element_line(colour = "#856f2c"), panel.grid.minor = element_blank(), panel.
```

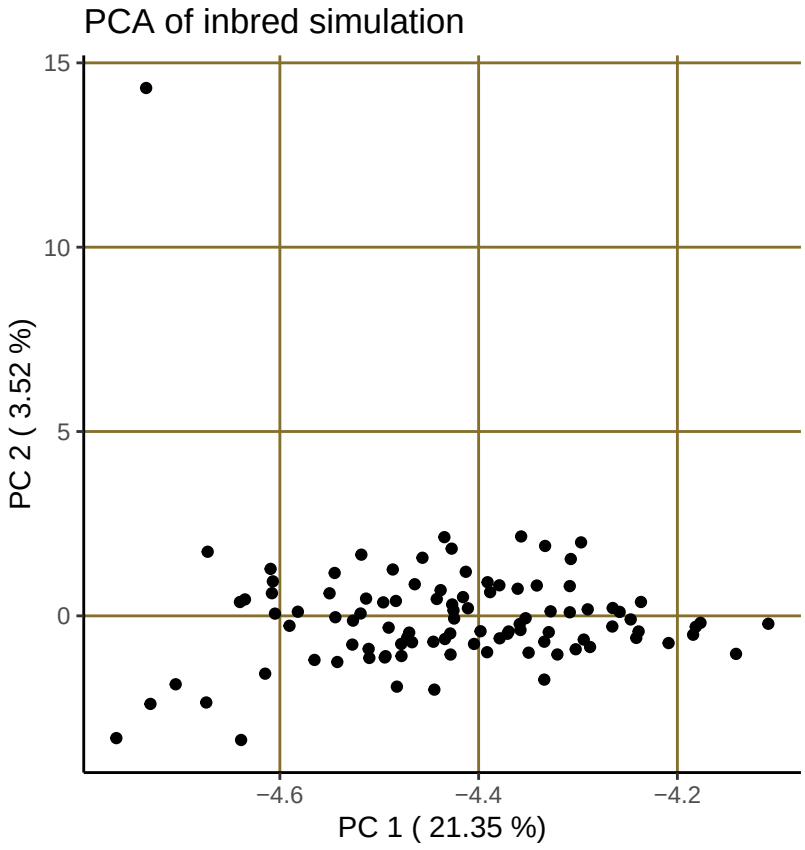



```
# Assign the PCA you want to plot: Inbreed simulation
pca1i <- iPCA

# calculate percent variance of each component:
pc1i <- round(pca1i$eig[1]/sum(pca1i$eig)*100,digits=2)
pc2i <- round(pca1i$eig[2]/sum(pca1i$eig)*100,digits=2)

# define what you want to plot and create a dataframe:
PC1i <- pca1i$li[,1]
PC2i <- pca1i$li[,2]
dfi <- data.frame(PC1i,PC2i)

# plot:
ggplot(data = dfi, aes(PC1i,PC2i),)+
  xlab(paste("PC 1 (",pc1i,"%)"))+
  ylab(paste("PC 2 (",pc2i,"%)"))+
  geom_point(size=1.5)+
  ggtitle("PCA of inbred simulation")+
  theme(panel.grid.major = element_line(colour = "#856f2c"), panel.grid.minor = element_blank(), panel.
```



Learning self-assessment questions (after):

Describe in your own words, what is the Hardy-Weinberg principle?

How might this principle be useful to you in understanding the health of a population of conservation concern?

Can you think of a scenario where it might be difficult to interpret what is the level of inbreeding in a population you have genotyped a sample from based on estimates of heterozygosity and FIS alone?

Did this activity improve your ability to answer these questions?

Did this activity improve your overall understanding of the utility of the Hardy-Weinberg principle? If so, how so?

References

Nei M. (1987) Molecular Evolutionary Genetics. Columbia University Press

This material drew inspiration and lecture note material from my Ph.D. advisor Dr. Grant Pogson's "Population Genetics" course from UC Santa Cruz, Spring Quarter, 2009.