

Hardy-Weinberg Equilibrium Activity: Student Version

Norah Saarman

9/8/2022

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:

Male genotypes	Female genotypes		
	A ₁ A ₁ (P)	A ₁ A ₂ (H)	A ₂ A ₂ (Q)
A ₁ A ₁ (P)	P ²	PH	PQ
A ₁ A ₂ (H)	PH	H ²	HQ
A ₂ A ₂ (Q)	PQ	HQ	Q ²

The progeny produced by this set of matings would be:

Mating	Total Frequency	Progeny		
		A ₁ A ₁ (P)	A ₁ A ₂ (H)	A ₂ A ₂ (Q)
A ₁ A ₁ x A ₁ A ₁	P ²	P ²		
A ₁ A ₁ x A ₁ A ₂	2PH	PH	PH	
A ₁ A ₁ x A ₂ A ₂	2PQ		2PQ	
A ₁ A ₂ x A ₁ A ₂	H ²	H ² /4	H ² /2	H ² /4
A ₁ A ₂ x A ₂ A ₂	HQ		HQ	HQ
A ₂ A ₂ x A ₂ A 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 (Hexp) 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 can occur in cases of directional selection 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 N_e
- 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
# NOTE: "E" is for example, change to "A", "B", "C", or "D"
statsE <- basic.stats(myData[myData@pop == "E"])
```

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

Question 1: What is the first table? What are some trends that you see in the data?

Question 2: What is the second table? What is the overall expected heterozygosity, observed heterozygosity, and FIS in this dataset?

```
# plot the FIS per locus, remember to change "E" to match your assigned scenario
plot(statsE$perloc$Fis,main="Scenario E",xlab = "locus", ylab = "Fis")
abline(h=statsA$overall[9],col="red")
```

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 deficit (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

First, let's simulate inbreeding:

```
# simulate inbreeding
iSim <- sim.genot(size=100,nbal=8,nbloc=15,nbpop=1,N=1000,mu=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("Inbreeding",100)) # fill in "@pop" of genind object (for hierfstat)
```

Now, let's simulate restricted gene flow (population structure):

```
# simulate 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=TRUE),f=0.01,mu=0.01,t=1000) # simulate
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("Structure",100)) # fill in "@pop" of genind object (for hierfstat)
```

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 below 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 0.65 0.7635 0.7635  0 NA  NA  0 NaN 0.1487  NA
## loc02 0.34 0.4670 0.4670  0 NaN NaN  0 NaN 0.2719  NaN
## loc03 0.28 0.3420 0.3420  0 NaN NaN  0 NaN 0.1813  NaN
## loc04 0.47 0.6518 0.6518  0 NaN NaN  0 NaN 0.2789  NaN
## loc05 0.55 0.7048 0.7048  0 NaN NaN  0 NaN 0.2197  NaN
## loc06 0.59 0.6905 0.6905  0 NA  NA  0 NaN 0.1455  NA
## loc07 0.31 0.3813 0.3813  0 NA  NA  0 NaN 0.1869  NA
## loc08 0.48 0.6203 0.6203  0 NaN NaN  0 NaN 0.2261  NaN
## loc09 0.64 0.8158 0.8158  0 NA  NA  0 NaN 0.2155  NA
## loc10 0.38 0.4363 0.4363  0 NaN NaN  0 NaN 0.1290  NaN
## loc11 0.49 0.6203 0.6203  0 NA  NA  0 NaN 0.2101  NA
## loc12 0.60 0.7218 0.7218  0 NA  NA  0 NaN 0.1688  NA
## loc13 0.60 0.8118 0.8118  0 NA  NA  0 NaN 0.2609  NA
## loc14 0.64 0.7978 0.7978  0 NA  NA  0 NaN 0.1978  NA
## loc15 0.62 0.7622 0.7622  0 NA  NA  0 NaN 0.1866  NA
##
## $overall
##      Ho      Hs      Ht      Dst      Htp      Dstp      Fst      Fstp      Fis      Dest
## 0.5093 0.6391 0.6391 0.0000      NaN      NaN 0.0000      NaN 0.2031      NaN
```

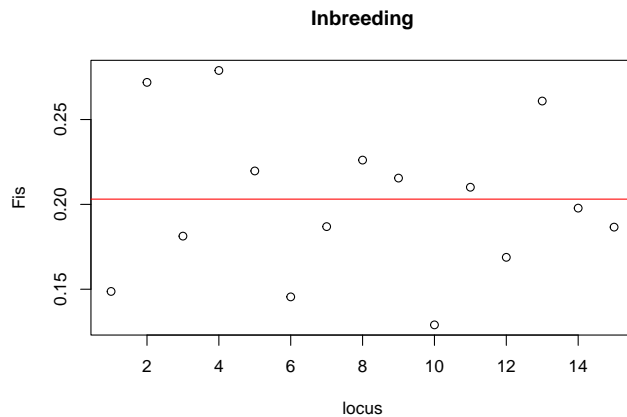
```
sStats <- basic.stats(sData)
sStats
```

```
## $perloc
##      Ho      Hs      Ht Dst Htp Dstp Fst Fstp  Fis Dest
## loc01 0.60 0.7457 0.7457  0 NA  NA  0 NaN 0.1954  NA
## loc02 0.76 0.7763 0.7763  0 NaN NaN  0 NaN 0.0210  NaN
## loc03 0.51 0.7693 0.7693  0 NaN NaN  0 NaN 0.3371  NaN
## loc04 0.64 0.6802 0.6802  0 NaN NaN  0 NaN 0.0590  NaN
## loc05 0.66 0.6561 0.6561  0 NaN NaN  0 NaN -0.0059  NaN
## loc06 0.47 0.7208 0.7208  0 NA  NA  0 NaN 0.3480  NA
## loc07 0.65 0.8344 0.8344  0 NaN NaN  0 NaN 0.2210  NaN
## loc08 0.69 0.8412 0.8412  0 NA  NA  0 NaN 0.1798  NA
## loc09 0.42 0.6723 0.6723  0 NaN NaN  0 NaN 0.3753  NaN
## loc10 0.57 0.7210 0.7210  0 NA  NA  0 NaN 0.2094  NA
## loc11 0.63 0.6551 0.6551  0 NA  NA  0 NaN 0.0382  NA
## loc12 0.72 0.7392 0.7392  0 NA  NA  0 NaN 0.0260  NA
## loc13 0.60 0.6990 0.6990  0 NaN NaN  0 NaN 0.1417  NaN
## loc14 0.44 0.6360 0.6360  0 NA  NA  0 NaN 0.3082  NA
## loc15 0.72 0.7027 0.7027  0 NaN NaN  0 NaN -0.0247  NaN
```

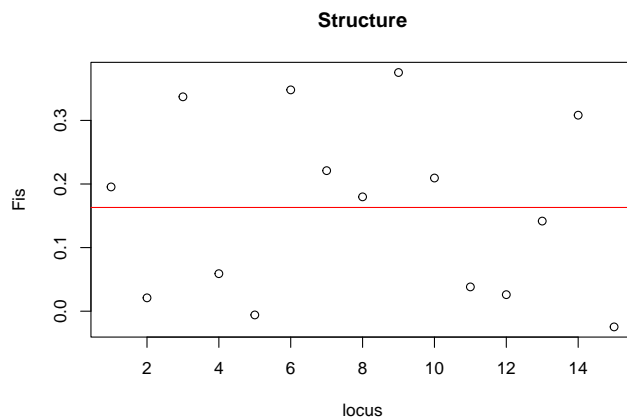
```
##
## $overall
##      Ho      Hs      Ht      Dst      Htp      Dstp      Fst      Fstp      Fis      Dest
## 0.6053 0.7233 0.7233 0.0000      NaN      NaN 0.0000      NaN 0.1631      NaN
```

```
# FIS plots
```

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



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



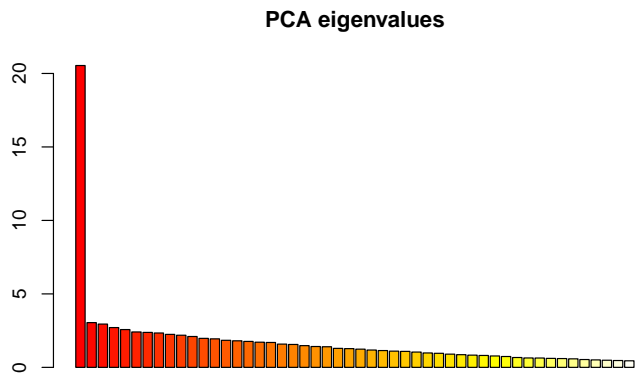
Question 8: Do you think there is an overall heterozygote excess or deficit in these two datasets? What do you think caused this excess/deficit in each case?

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)?

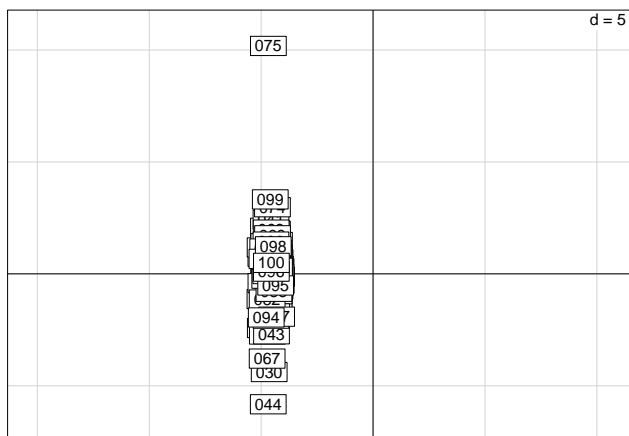
One idea for this is to run a principal component analysis on the simulated genotypes and visualize (plot) the eigenvectors to see how the individuals cluster... Is there any population structure?

First let's plot the PCA eigenvalues and eigenvectors from the simulation with inbreeding:

```
# PCA for simulation with inbreeding
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
```

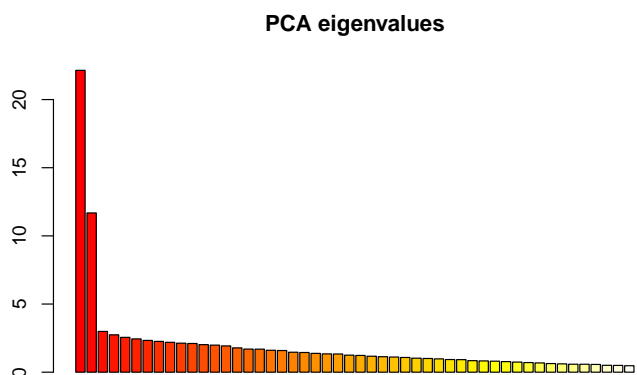


```
s.label(iPCA$li) # plot eigenvectors
```

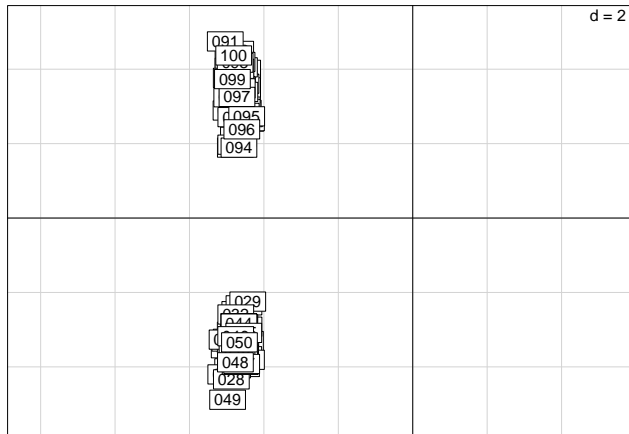


Now, let's plot the PCA eigenvalues and eigenvectors from the simulation with restricted gene flow:

```
# 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	0.50	0.5173	0.5173	0	NaN	NaN	0	NaN	0.0335	NaN
## loc02	0.68	0.7694	0.7694	0	NaN	NaN	0	NaN	0.1162	NaN
## loc03	0.56	0.5122	0.5122	0	NaN	NaN	0	NaN	-0.0932	NaN
## loc04	0.50	0.5178	0.5178	0	NaN	NaN	0	NaN	0.0343	NaN
## loc05	0.64	0.5608	0.5608	0	NaN	NaN	0	NaN	-0.1412	NaN
## loc06	0.70	0.7588	0.7588	0	NA	NA	0	NaN	0.0775	NA
## loc07	0.72	0.7671	0.7671	0	NA	NA	0	NaN	0.0615	NA
## loc08	0.80	0.8216	0.8216	0	NA	NA	0	NaN	0.0263	NA
## loc09	0.08	0.0784	0.0784	0	NaN	NaN	0	NaN	-0.0208	NaN
## loc10	0.78	0.7306	0.7306	0	NaN	NaN	0	NaN	-0.0676	NaN
## loc11	0.60	0.5845	0.5845	0	NaN	NaN	0	NaN	-0.0265	NaN
## loc12	0.74	0.7543	0.7543	0	NaN	NaN	0	NaN	0.0189	NaN
## loc13	0.70	0.7292	0.7292	0	NA	NA	0	NaN	0.0400	NA
## loc14	0.26	0.2776	0.2776	0	NA	NA	0	NaN	0.0632	NA
## loc15	0.66	0.6596	0.6596	0	NaN	NaN	0	NaN	-0.0006	NaN

```
##
```

```
## $overall
```

##	Ho	Hs	Ht	Dst	Htp	Dstp	Fst	Fstp	Fis	Dest
##	0.5947	0.6026	0.6026	0.0000	NaN	NaN	0.0000	NaN	0.0132	NaN

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

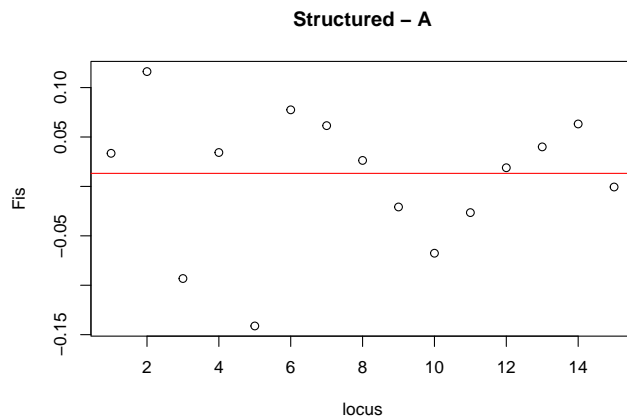
```
bStats
```

```
## $perloc
```

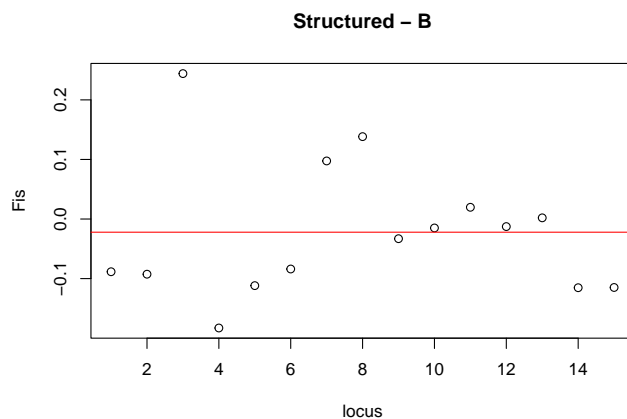
```
##      Ho      Hs      Ht Dst Htp Dstp Fst Fstp      Fis Dest
## loc01 0.70 0.6431 0.6431 0 NA NA 0 NaN -0.0885 NA
## loc02 0.84 0.7688 0.7688 0 NaN NaN 0 NaN -0.0926 NaN
## loc03 0.46 0.6086 0.6086 0 NaN NaN 0 NaN 0.2441 NaN
## loc04 0.78 0.6594 0.6594 0 NaN NaN 0 NaN -0.1829 NaN
## loc05 0.68 0.6116 0.6116 0 NaN NaN 0 NaN -0.1118 NaN
## loc06 0.24 0.2214 0.2214 0 NaN NaN 0 NaN -0.0839 NaN
## loc07 0.58 0.6427 0.6427 0 NaN NaN 0 NaN 0.0975 NaN
## loc08 0.58 0.6731 0.6731 0 NaN NaN 0 NaN 0.1383 NaN
## loc09 0.76 0.7357 0.7357 0 NaN NaN 0 NaN -0.0330 NaN
## loc10 0.36 0.3547 0.3547 0 NaN NaN 0 NaN -0.0150 NaN
## loc11 0.66 0.6733 0.6733 0 NaN NaN 0 NaN 0.0197 NaN
## loc12 0.70 0.6912 0.6912 0 NA NA 0 NaN -0.0127 NA
## loc13 0.50 0.5010 0.5010 0 NaN NaN 0 NaN 0.0020 NaN
## loc14 0.62 0.5559 0.5559 0 NaN NaN 0 NaN -0.1153 NaN
## loc15 0.78 0.6996 0.6996 0 NaN NaN 0 NaN -0.1149 NaN
##
## $overall
##      Ho      Hs      Ht      Dst      Htp      Dstp      Fst      Fstp      Fis      Dest
## 0.6160 0.6027 0.6027 0.0000      NaN      NaN 0.0000      NaN -0.0221      NaN
```

```
# FIS plots
```

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



```
plot(bStats$perloc$Fis,xlab = "locus", ylab = "Fis", main = "Structured - 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 simulation with inbreeding:

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

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

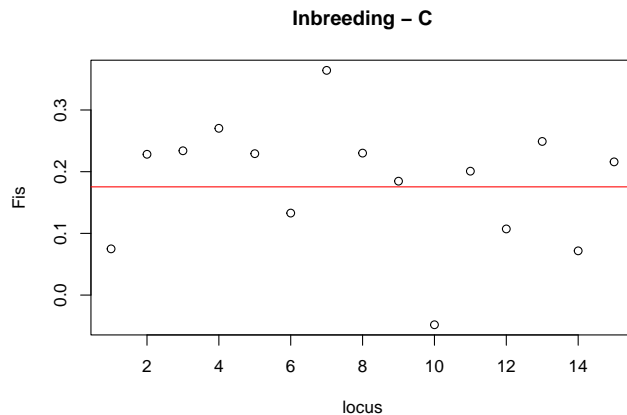
```
## $perloc
##      Ho      Hs      Ht Dst Htp Dstp Fst Fstp      Fis Dest
## loc01 0.7143 0.7721 0.7721 0 NaN NaN 0 NaN 0.0749 NaN
## loc02 0.3673 0.4760 0.4760 0 NA  NA 0 NaN 0.2282  NA
## loc03 0.3265 0.4262 0.4262 0 NA  NA 0 NaN 0.2339  NA
## loc04 0.4694 0.6433 0.6433 0 NaN NaN 0 NaN 0.2703 NaN
## loc05 0.5306 0.6884 0.6884 0 NA  NA 0 NaN 0.2292  NA
## loc06 0.6327 0.7296 0.7296 0 NaN NaN 0 NaN 0.1329 NaN
## loc07 0.1224 0.1926 0.1926 0 NaN NaN 0 NaN 0.3642 NaN
## loc08 0.4694 0.6097 0.6097 0 NaN NaN 0 NaN 0.2301 NaN
## loc09 0.6735 0.8259 0.8259 0 NaN NaN 0 NaN 0.1846 NaN
## loc10 0.4490 0.4284 0.4284 0 NA  NA 0 NaN -0.0481  NA
## loc11 0.5306 0.6639 0.6639 0 NaN NaN 0 NaN 0.2008 NaN
## loc12 0.6531 0.7315 0.7315 0 NA  NA 0 NaN 0.1072  NA
## loc13 0.5918 0.7881 0.7881 0 NA  NA 0 NaN 0.2490  NA
## loc14 0.6939 0.7474 0.7474 0 NaN NaN 0 NaN 0.0717 NaN
## loc15 0.6122 0.7808 0.7808 0 NA  NA 0 NaN 0.2159  NA
##
## $overall
##      Ho      Hs      Ht      Dst      Htp      Dstp      Fst      Fstp      Fis      Dest
## 0.5224 0.6336 0.6336 0.0000      NaN      NaN 0.0000      NaN 0.1754      NaN
```

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

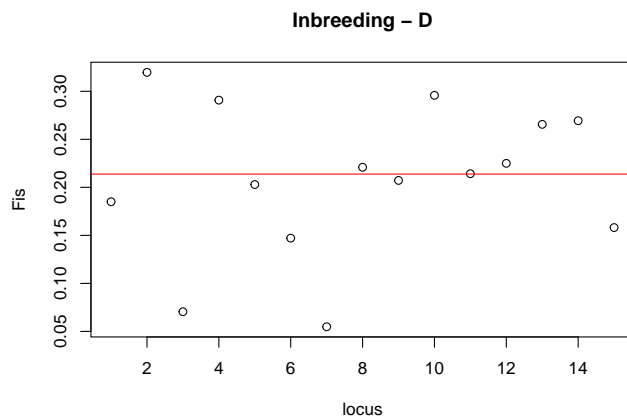
```
## $perloc
##      Ho      Hs      Ht Dst Htp Dstp Fst Fstp      Fis Dest
## loc01 0.5882 0.7218 0.7218 0 NaN NaN 0 NaN 0.1850 NaN
## loc02 0.3137 0.4612 0.4612 0 NaN NaN 0 NaN 0.3197 NaN
## loc03 0.2353 0.2531 0.2531 0 NaN NaN 0 NaN 0.0705 NaN
## loc04 0.4706 0.6635 0.6635 0 NA  NA 0 NaN 0.2908  NA
## loc05 0.5686 0.7133 0.7133 0 NaN NaN 0 NaN 0.2029 NaN
## loc06 0.5490 0.6437 0.6437 0 NaN NaN 0 NaN 0.1471 NaN
## loc07 0.4902 0.5186 0.5186 0 NaN NaN 0 NaN 0.0548 NaN
## loc08 0.4902 0.6292 0.6292 0 NA  NA 0 NaN 0.2209  NA
## loc09 0.6078 0.7667 0.7667 0 NA  NA 0 NaN 0.2072  NA
## loc10 0.3137 0.4455 0.4455 0 NaN NaN 0 NaN 0.2958 NaN
## loc11 0.4510 0.5739 0.5739 0 NaN NaN 0 NaN 0.2142 NaN
## loc12 0.5490 0.7084 0.7084 0 NaN NaN 0 NaN 0.2250 NaN
## loc13 0.6078 0.8276 0.8276 0 NaN NaN 0 NaN 0.2656 NaN
## loc14 0.5882 0.8051 0.8051 0 NA  NA 0 NaN 0.2694  NA
```

```
## loc15 0.6275 0.7453 0.7453 0 NaN NaN 0 NaN 0.1581 NaN
##
## $overall
##      Ho      Hs      Ht      Dst      Htp      Dstp      Fst      Fstp      Fis      Dest
## 0.4967 0.6318 0.6318 0.0000      NaN      NaN 0.0000      NaN 0.2138      NaN

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



```
plot(dStats$perloc$Fis,xlab = "locus", ylab = "Fis", main = "Inbreeding - 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?

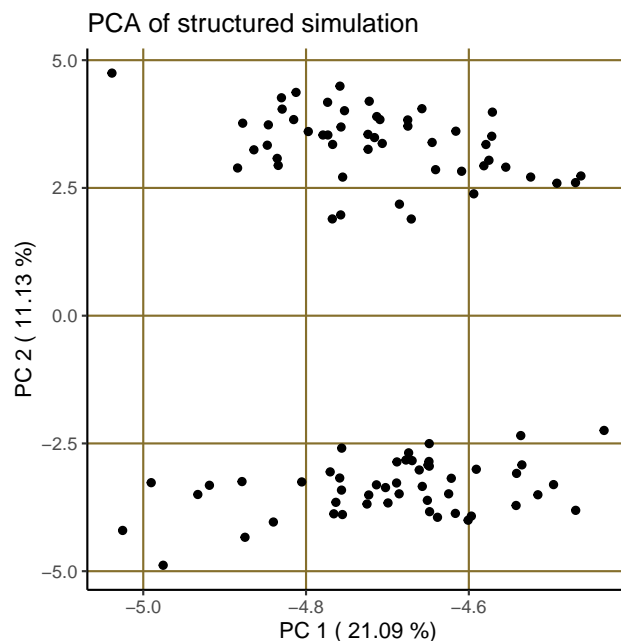
```
# Assign the PCA you want to plot: Structured simulation
pca1 <- sPCA

# 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.f
```



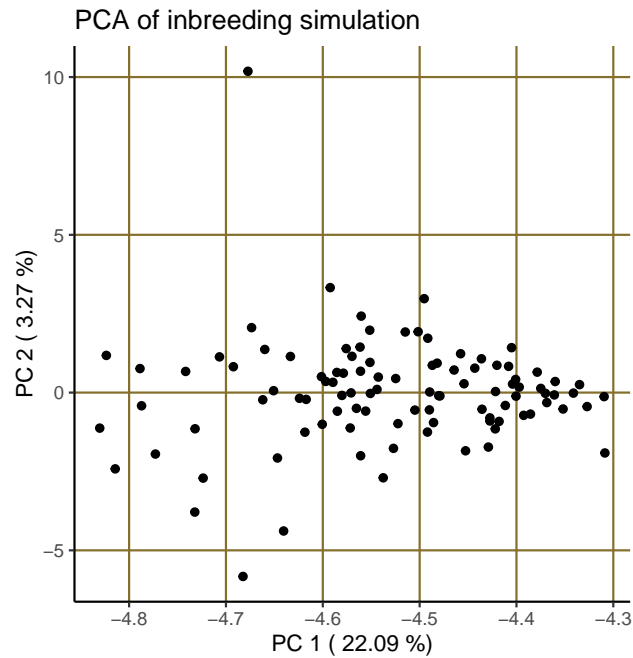
```
# Assign the PCA you want to plot: Inbreed simulation
pcai <- iPCA

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

# define what you want to plot and create a dataframe:
PC1i <- pc1i$li[,1]
PC2i <- pc1i$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 inbreeding simulation")+
  theme(panel.grid.major = element_line(colour = "#856f2c"), panel.grid.minor = element_blank(), panel.f
```

```
theme(panel.grid.major = element_line(colour = "#856f2c"), panel.grid.minor = element_blank(), panel.l
```



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.