Problem Set 1

Gitanshu Munjal

Saturday, April 11, 2015

Problem 1

Calculate the allele frequenies for each locus.

I define an R function to calculate allele frequencies for a bi-allelic locus given genotype frequencies. When provided genotype frequencies in the order (homozygous1, heterozygous, homozygous2), the function returns frequency of allele1, frequency of allele2, and 1-frequency of allele1 (added to double check calc. allele 2 frequency). I recycle this function for later questions in the assignment.

```
################################################
# Allele Frequencies
################################################

allelefrequency <- function(ho1,het,ho2){
                                  a1 <- (2*ho1 + het)/(2*(ho1+het+ho2));
                                  a2 <- (2*ho2 + het)/(2*(ho1+het+ho2));
                                  a2c <- 1-a1;
                                  freqs <- c("allele1"=a1,"allele2"=a2,"1-allele1"=a2c)
                                  return(freqs)
                  }


#Locus1
ho1 <- cc <- 47     # allele1 = c
het <- ct <- 18
ho2 <- tt <- 35     # allele2 = t
allelefrequency(ho1,het,ho2)
```

```
##   allele1   allele2 1-allele1
##      0.56      0.44      0.44
```

So,the frequency of the C and T alleles at Locus 1 are 0.56 and 0.44 respectively.

```
#Locus2
ho1 <- aa <- 50     # allele1 = a
het <- ag <- 42
ho2 <- gg <- 8      # allele2 = g
allelefrequency(ho1,het,ho2)
```

```
##   allele1   allele2 1-allele1
##      0.71      0.29      0.29
```

So,the frequency of the A and G alleles at Locus 2 are 0.56 and 0.44 respectively.

What is the expected heterozygosity for each locus given the allele frequencies?

I define a function to return the expected frequency of heterozygous genotypes as 2pq when allele frequencies are p and q.

```
################################################
# Expected heterozygosity
################################################

ehetero <- function(p,q){2*p*q}

#Locus1
ehetero(0.56,0.44)
```

```
## [1] 0.4928
```

```
#Locus2
ehetero(0.71,0.29)
```

```
## [1] 0.4118
```

So, expected frequency of heterozygous genotypes at Locus 1 and Locus 2 are 0.4928 and 0.411 respectively.

Does the population significantly deviate from Hardy-Weinberg Equilibrium (HWE) at each locus?

```
################################################
# Test for Hardy-Weinberg Equilibrium (HWE)
################################################

#Locus1
#Observed data (genotype counts)
occ <- 47     # allele1 = c
oct <- 18
ott <- 35     # allele2 = t
total <- occ+oct+ott

#Expected frequency (under HWE)
p <- as.numeric(allelefrequency(occ,oct,ott)[1])
q <- as.numeric(allelefrequency(occ,oct,ott)[2])

#Expected genotype counts
ecc <- p^2 * total
ect <- 2*p*q * total
ett <- q^2 * total
c("ecc"=ecc,"ect"=ect,"ett"=ett)
```

```
##   ecc   ect   ett
## 31.36 49.28 19.36
```

```
#Matrix containing observed and expected data
hwedat <- matrix(NA,3,3)
colnames(hwedat) <- c("observed","expected","(o-e)^2/e")
rownames(hwedat) <- c("cc","ct","tt")
hwedat[,1] <- c(occ,oct,ott)
hwedat[,2] <- c(ecc,ect,ett)
obs <- hwedat[,1]
exp <- hwedat[,2]
hwedat[,3] <- ((obs - exp)^2)/exp
hwedat
```

```
##    observed expected (o-e)^2/e
## cc       47    31.36  7.800051
## ct       18    49.28 19.854675
## tt       35    19.36 12.634793
```

```
#Chi-square and p-value
chisq <- sum(hwedat[,3])
chisq                                #chi-squared value
```

```
## [1] 40.28952
```

```
pchisq(chisq,df=1,lower.tail=FALSE) #p-value
```

```
## [1] 2.189805e-10
```

Indeed, **the population significantly deviates from HWE at Locus 1** at the 95% confidence level.

```
#Locus2
#Observed data (genotype counts)
oaa <- 50     # allele1 = a
oag <- 42
ogg <- 8     # allele2 = g
total <- oaa+oag+ogg

#Expected frequency (under HWE)
p <- as.numeric(allelefrequency(oaa,oag,ogg)[1])
q <- as.numeric(allelefrequency(oaa,oag,ogg)[2])

#Expected genotype counts
eaa <- p^2 * total
eag <- 2*p*q * total
egg <- q^2 * total
c("eaa"=eaa,"eag"=eag,"egg"=egg)
```

```
##   eaa   eag   egg
## 50.41 41.18  8.41
```

```r
hwedat <- matrix(NA,3,3)
colnames(hwedat) <- c("observed","expected","(o-e)^2/e")
rownames(hwedat) <- c("aa","ag","gg")
hwedat[,1] <- c(oaa,oag,ogg)
hwedat[,2] <- c(eaa,eag,egg)
obs <- hwedat[,1]
exp <- hwedat[,2]
hwedat[,3] <- ((obs - exp)^2)/exp
hwedat
```

```
##    observed expected   (o-e)^2/e
## aa       50    50.41 0.003334656
## ag       42    41.18 0.016328315
## gg        8     8.41 0.019988109
```

```r
chisq <- sum(hwedat[,3])
chisq                                #chi-squared value
```

```
## [1] 0.03965108
```

```r
pchisq(chisq,df=1,lower.tail=FALSE) #p-value
```

```
## [1] 0.8421643
```

At the 95% confidence level, **the population does not significantly deviate from HWE at Locus 2.**

What are two possible reasons a population would deviate from HWE?

(1) Non-random mating
(2) Migration (I/E-mmigration)

Problem 2

How many of each haplotype were definitively observed in this population?

All genotypes except the CTAG double heterozygote contribute towardss the observed haplotypes. Double homozygotes contribute $2N$ haplotypes and heterozygotes for one or the other locus contribute N haplotypes where N is the number of individuals with the haplotype under consideration.

```
################################################
# Definitively Observed Haplotypes
################################################

CA <- (2*40) + (2) + (4)
TA <- (2*8) + (2) + (22)
CG <- (2*4) + (4) + (0)
TG <- (2*4) + (22) + (0)
Total <- CA + TA + CG + TG

#Matrix of Observed Haplotypes
obsdata <- matrix(NA,5,1)
colnames(obsdata) <- c("NumberObserved")
rownames(obsdata) <- c("CA","TA","CG","TG","Total")
obsdata[,1] <- c(CA,TA,CG,TG,Total)
obsdata
```

```
##       NumberObserved
## CA                86
## TA                40
## CG                12
## TG                30
## Total            168
```

The above matrix shows the number of each haplotype that were definitively observed.

Calculate D, D', and r2 for the C-A haplotype.

I use the definitive set of observed haplotypes to calculate allele frequencies.

```
################################################
# D, D', and r2 for the C-A haplotype
################################################

#Allele Frequencies
c <- (CA + CG) / Total
a <- (CA + TA) / Total
t <- (TA + TG) / Total
g <- (CG + TG) / Total
c("c"=c,"a"=a,"t"=t,"g"=g)
```

```
##         c         a         t         g
## 0.5833333 0.7500000 0.4166667 0.2500000
```

The allele frequencies are then used to calculate expected haplotype frequencies

```
#Expected Haplotype Frequencies
expca <- c * a
expta <- t * a
expcg <- c * g
exptg <- t * g
c("expca"=expca,"expta"=expta,"expcg"=expcg,"exptg"=exptg)
```

```
##     expca     expta     expcg     exptg
## 0.4375000 0.3125000 0.1458333 0.1041667
```

```
#Data Matrix
parta <- matrix(NA,5,3)
colnames(parta) <- c("ObservedNumber","ObservedFrequency","ExpectedFrequency")
rownames(parta) <- c("CA","TA","CG","TG","Total")
parta[,1] <- c(CA,TA,CG,TG,Total)
parta[,2] <- parta[,1]/Total
parta[,3] <- c(expca,expta,expcg,exptg,expca+expta+expcg+exptg)
parta
```

```
##       ObservedNumber ObservedFrequency ExpectedFrequency
## CA                86        0.51190476         0.4375000
## TA                40        0.23809524         0.3125000
## CG                12        0.07142857         0.1458333
## TG                30        0.17857143         0.1041667
## Total            168        1.00000000         1.0000000
```

```
#D for C-A haplotype
obsca <- parta[1,2]
D <- obsca - expca
D
```

```
## [1] 0.07440476
```

Thus, $\mathbf{D = 0.0744}$

```
#D prime for C-A haplotype
Dprime <- D / min(expcg,expta)
Dprime
```

```
## [1] 0.5102041
```

Thus, $\mathbf{D' = 0.510}$

```
#R squared for C-A haplotype
Rsquared <- (D^2) / (c*a*t*g)
Rsquared
```

```
## [1] 0.1214772
```

Thus, $\mathbf{R^2 = 0.121}$

Calculate Chi-square and determine if there is significant linkage disequilibrium (LD) at this locus.

We use expected and observed haplotype frequencies previously calculated to determine expected haplotype counts assuming the total allele count to be 168 and then compare these data with the observed data set to determine significance of LD using a chi-squared test.

```
################################################
# Chi-square test to determine significance of LD
################################################

chisqtab <- matrix(NA,4,3)
colnames(chisqtab) <- c("observed","expected","(o-e)^2/e")
rownames(chisqtab) <- c("CA","TA","CG","TG")
obs <- chisqtab[,1] <- parta[1:4,2]*168
exp <- chisqtab[,2] <- parta[1:4,3]*168
chisqtab[,3] <- ((obs - exp)^2)/exp
chisqtab
```

```
##    observed expected (o-e)^2/e
## CA       86     73.5  2.125850
## TA       40     52.5  2.976190
## CG       12     24.5  6.377551
## TG       30     17.5  8.928571
```

```
chisq <- sum(chisqtab[,3])
chisq                              #chi-squared value
```

```
## [1] 20.40816
```

```
pchisq(chisq,df=1,lower.tail=FALSE) #p-value
```

```
## [1] 6.256236e-06
```

Indeed, **there is significant LD at this locus** at the 95% confience level.

Problem 3

Calculate the p-value of association between genotype and phenotype using a Chi-square test on expected and observed allele counts.

We recycle the function we defined in question 1 and use total genotype counts (across cases and controls) to determine allele frequencies.

```
################################################
# Allele Frequencies
################################################

#Total counts
cc <- 150
ca <- 250
aa <- 100

#Allele Frequencies
allelefrequency(cc,ca,aa)[1] #allele1 = c ; allele2 = a)
```

```
## allele1
##    0.55
```

```
c <- as.numeric(allelefrequency(cc,ca,aa)[1])
a <- as.numeric(allelefrequency(cc,ca,aa)[2])
```

Thus, the frequencies of the C and A allele are 0.55 and 0.45 respectively. We can use these frequencies to determine the expected allele counts for cases and controls by multiplying them each to the number of cases and controls respectively, as follows:

```
#Matrix of observed allele counts
obsdata <- matrix(NA,2,2)
rownames(obsdata) <- c("C","A")
colnames(obsdata) <- c("Cases","Controls")
obsdata[,1] <- c(200,200)
obsdata[,2] <- c(350,250)
totalcasealleles <- sum(obsdata[,1])
totalcontrolalleles <- sum(obsdata[,2])
obsdata
```

```
##   Cases Controls
## C   200      350
## A   200      250
```

```r
#Matrix of expected allele counts
expdata <- matrix(NA,2,2)
rownames(expdata) <- c("C","A")
colnames(expdata) <- c("Cases","Controls")
expdata[,1] <- c(c,a)*totalcasealleles            #Expected case count
expdata[,2] <- c(c,a)*totalcontrolalleles         #Expected control count
expdata
```

```
##   Cases Controls
## C   220      330
## A   180      270
```

Next, we use the above presented observed and expected matrices to determine significance of association using a chi-squared test.

```r
################################################
# Chi-square test for allele counts to determine association
################################################

x <- ((obsdata - expdata)^2) / expdata
chisq <- sum(x)
chisq                                #chi-squared value
```

```
## [1] 6.734007
```

```r
pchisq(chisq,df=1,lower.tail=FALSE) #p-value
```

```
## [1] 0.009459189
```

It would appear that the genotype is indeed significantly associated with the phenotype at the 95% confidence level **(p-value: 0.009459189)**

If the researchers used a SNP chip with 100,000 SNPs. How many of these SNPs would you expect to be that significant by chance?

```r
pchisq(chisq,df=1,lower.tail=FALSE)*100000
```

```
## [1] 945.9189
```

I would expect 946 SNPs to be that significant just by chance.

Do you think this locus plays an important role in gluten sensitivity?

No, given the p-value and the number of SNPs used I do not feel very confident in the association of this locus with gluten sensitivity as even using the most naive adjustment (like Bonferroni) for the p-value, I would need to see a p-value much lower (to the order of p < 10^-6) to be fairly confident in the association.

Problem 4

```
################################################
# Characterizing subpopulation structure
################################################

#Raw data provided
rawdata <- matrix(NA,3,6)
colnames(rawdata) <- c("SNP","RA","SP1","SP2","SP3","SP4")
rownames(rawdata) <- c("L1","L2","L3")
rawdata[1,] <- c("A/T","A",0.8,0.7,0.9,0.2)
rawdata[2,] <- c("G/C","G",0.6,0.1,0.8,0.7)
rawdata[3,] <- c("C/A","C",0.7,0.8,0.2,0.6)
rawdata <- as.table(rawdata)
rawdata
```

```
##    SNP RA SP1 SP2 SP3 SP4
## L1 A/T A  0.8 0.7 0.9 0.2
## L2 G/C G  0.6 0.1 0.8 0.7
## L3 C/A C  0.7 0.8 0.2 0.6
```

```
#Reduced raw data
obsdata <- matrix(NA,3,4)
colnames(obsdata) <- c("SP1","SP2","SP3","SP4")
rownames(obsdata) <- c("L1","L2","L3")
obsdata[,1:4] <- as.numeric(rawdata[,3:6])
obsdata
```

```
##    SP1 SP2 SP3 SP4
## L1 0.8 0.7 0.9 0.2
## L2 0.6 0.1 0.8 0.7
## L3 0.7 0.8 0.2 0.6
```

```
#Expected heterozygosity matrix
exphs <- matrix(NA,3,4)
colnames(exphs) <- c("SP1","SP2","SP3","SP4")
rownames(exphs) <- c("L1","L2","L3")
exphs <- 2 * (obsdata) * (1 - obsdata)
exphs
```

```
##     SP1  SP2  SP3  SP4
## L1 0.32 0.42 0.18 0.32
## L2 0.48 0.18 0.32 0.42
## L3 0.42 0.32 0.32 0.48
```

What is F_{ST} between SP01 and SP03 at locus 1?

```
hsbar <- rowMeans(subset(exphs, select = c(SP1, SP3)))[1]
averagep <- rowMeans(subset(obsdata, select = c(SP1, SP3)))[1]
ht <- 2 * averagep * (1-averagep)
loc1fst13 <- (ht -hsbar)/ht
loc1fst13
```

```
##         L1
## 0.01960784
```

Thus, F_{ST} **between SP01 and SP03 at locus 1 is 0.0196**

What is F_{ST} between SP02 and SP04 at locus 2?

```
hsbar <- rowMeans(subset(exphs, select = c(SP2, SP4)))[2]
averagep <- rowMeans(subset(obsdata, select = c(SP2, SP4)))[2]
ht <- 2 * averagep * (1-averagep)
loc2fst24 <- (ht -hsbar)/ht
loc2fst24
```

```
##    L2
## 0.375
```

Thus, F_{ST} **between SP02 and SP04 at locus 2 is 0.375**

What is F_{ST} among all four subpopulations at locus 3

```
hsbar <- rowMeans(exphs)[3]
averagep <- rowMeans(obsdata)[3]
ht <- 2 * averagep * (1-averagep)
loc3fst <- (ht -hsbar)/ht
loc3fst
```

```
##        L3
## 0.2122762
```

Thus, F_{ST} **among all populations at locus 3 is 0.212**

What is the likelihood that this individual originated from SP01? What about SP02, SP03 and SP04?

The genotype AA GC CA translates to homozygosity for the reference allele at locus 1 and heterozygosity for the other two loci. If p_i were the frequency of the ref. allele at loc_i then the frequency of the observed genotype would be $(p_1^2)(2 * p_2 * (1 - p_2))(2 * p_3 * (1 - p_3))$

```
probAA1 <- obsdata[1,]^2
probGC2 <- 2 * obsdata[2,] * (1-obsdata[2,])
probCA3 <- 2 * obsdata[3,] * (1-obsdata[3,])
probAAGCCA <- probAA1 * probGC2 * probCA3
probAAGCCA
```

```
##      SP1      SP2      SP3      SP4
## 0.129024 0.028224 0.082944 0.008064
```

The above matrix shows the population-wise likelihoods for the genotype.

What is the posterior probability that the individual originated from SP01? What about SP02, SP03 and SP04?

```
#Priors
priormatrix <- matrix(NA,4,1)
rownames(priormatrix) <- c("SP1","SP2","SP3","SP4")
priormatrix[,1] <- c(0.1,0.1,0.7,0.1)
colnames(priormatrix) <- c("priorprobability")
priormatrix
```

```
##     priorprobability
## SP1              0.1
## SP2              0.1
## SP3              0.7
## SP4              0.1
```

```
#Bayes Theorem Numerator
bayesnumerator <- probAAGCCA * priormatrix
colnames(bayesnumerator) <- c("bayesnumerator")
bayesnumerator
```

```
##     bayesnumerator
## SP1      0.0129024
## SP2      0.0028224
## SP3      0.0580608
## SP4      0.0008064
```

```
#Bayes Theorem Denominator
bayesdenominator <- sum(bayesnumerator)
bayesdenominator
```

```
## [1] 0.074592
```

```
#Posterior Probabilities
posteriorprob <- bayesnumerator/bayesdenominator
colnames(posteriorprob) <- c("posteriorprob")
posteriorprob
```

```
##     posteriorprob
## SP1    0.17297297
## SP2    0.03783784
## SP3    0.77837838
## SP4    0.01081081
```

The above matrix shows the population wise posterior probability for the given genotype.

Problem 5

I wrote a python program to analyze the given sequence data. The program (also submitted) returns pairwise differences, s, π(rounded), and the sequences reduced to segregating sites only(plus 1st base).

How many segregating sites (s) are present in these data?

10

What is pi (π) in these data?

```
Dij <- 66
n <- 6
pi <- Dij/((n*(n-1)) / 2)
pi
```

```
## [1] 4.4
```

Thus, π=**4.4**

What are s and π expressed in per site values?

```
s <- 10
pi <- 4.4
bp <- 50

spersite <- s/(bp)
spersite
```

```
## [1] 0.2
```

Thus, **s per site = 0.2.**

```
pipersite <- pi/(bp)
pipersite
```

```
## [1] 0.088
```

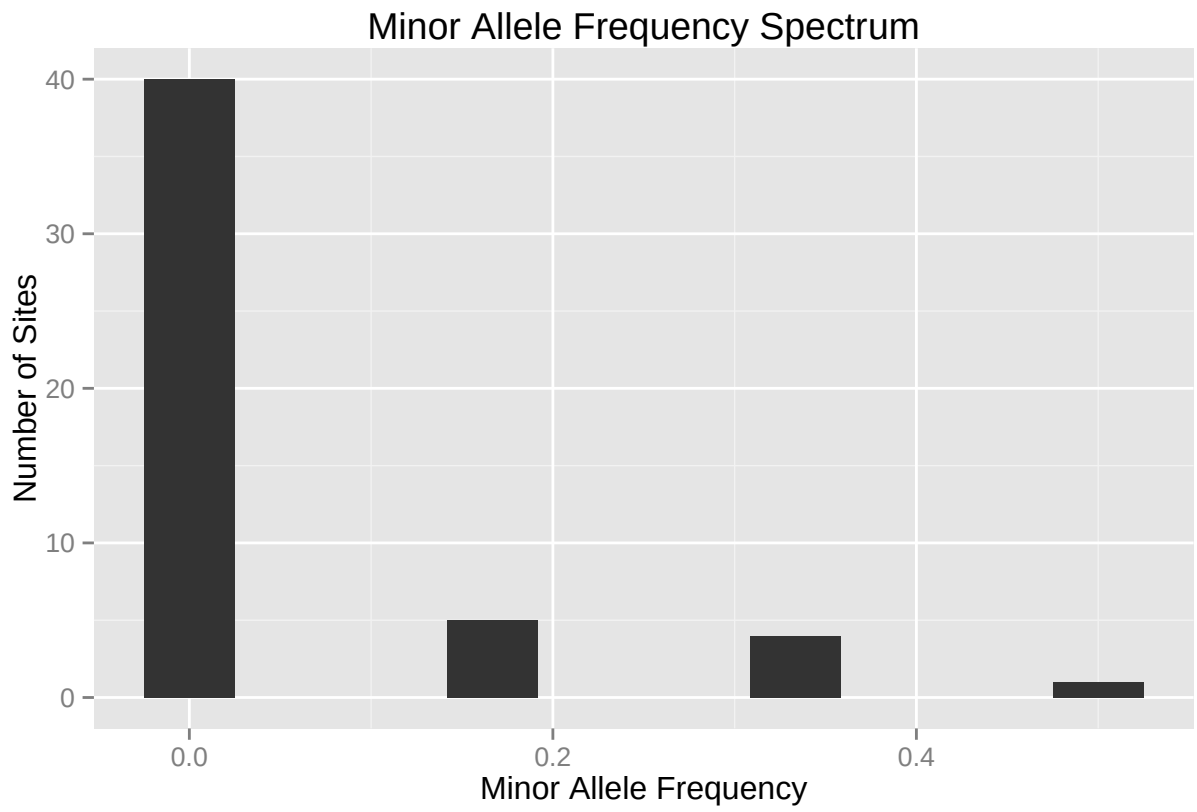
Thus, π **per site = 0.088.**

What is the minor allele frequency spectrum for these data?

```
sfsdata <- matrix(NA,4,2)
colnames(sfsdata) <- c("MinorAlleleFrequency","NumberofSites")
sfsdata[,2] <- c(40,5,4,1)
sfsdata[,1] <- c(0,1/6,2/6,3/6)
sfsdata
```

```
##      MinorAlleleFrequency NumberofSites
## [1,]            0.0000000            40
## [2,]            0.1666667             5
## [3,]            0.3333333             4
## [4,]            0.5000000             1
```

```
library(ggplot2)
sfsdata <- data.frame(sfsdata)
qplot(x=sfsdata$MinorAlleleFrequency,y=sfsdata$NumberofSites,geom="bar",
      stat="identity",width=0.05,xlab="Minor Allele Frequency",ylab="Number of Sites",
      main="Minor Allele Frequency Spectrum")
```



What is the derived allele frequency spectrum for these data?

```
sfsdata <- matrix(NA,6,2)
colnames(sfsdata) <- c("DerivedAlleleFrequency","NumberofSites")
sfsdata[,2] <- c(40,4,2,1,2,1)
sfsdata[,1] <- c(0,1/6,2/6,3/6,4/6,5/6)
sfsdata
```

```
##      DerivedAlleleFrequency NumberofSites
## [1,]              0.0000000            40
## [2,]              0.1666667             4
## [3,]              0.3333333             2
## [4,]              0.5000000             1
## [5,]              0.6666667             2
## [6,]              0.8333333             1
```

```
library(ggplot2)
sfsdata <- data.frame(sfsdata)
qplot(x=sfsdata$DerivedAlleleFrequency,y=sfsdata$NumberofSites,geom="bar",
      stat="identity",width=0.05,xlab="Derived Allele Frequency",ylab="Number of Sites",
      main="Derived Allele Frequency Spectrum")
```