**BIOMI 609 Computational Genomics and Bioinformatics**

**Spring 2022**

**Dr. Arun Sethuraman**

**Assignment 3 - Population Genomics**

**Due via Canvas on Friday, 3/25/2022 at 11:59 PM**

**Total: 30 points**

1. A SNP in the FTO gene associated with individuals with high risk of obesity is present in African populations at a frequency of 0.471, in Europeans at a frequency of 0.426, and in East Asians at a frequency of 0.157. Describe the Wahlund effect with respect to these allele frequencies, assuming random mating within each continental subpopulation. (10 points)

Let’s call this SNP as A, and the other allele as G

Africa: P(A) = 0.471 = p1 (say) => P(G) = 1-P(A) = 1-0.471 = 0.529 = q1

Europe: P(A) = 0.426 = p2 (say) => P(G) = 1-0.426 = 0.574 = q2

East Asia: P(A) = 0.157 = p3 (say) => P(G) = 1-0.157 = 0.843 = q3

If we consider this as a single population (without accounting for population structure), the average A allele frequency = (p1+p2+p3)/3 = 0.351 = p’, and average G allele frequency = 1-0.351 = 0.649 = q’. Therefore the expected heterozygosity in this “total” population = 2p’q’ = 2\*0.351\*0.649 = 0.456.

On the other hand, if we account for the presence of each subpopulation, the average expected heterozygosity would be (2p1q1+2p2q2+2p3q3)/3 = (2\*0.471\*0.529+2\*0.426\*0.574+2\*0.157\*0.843)/3 = 0.417.

The “reduction” in heterozygosity due to the presence of subpopulation structure, also called the Wahlund Effect is evident here, where 0.417 < 0.456.

2. Locus 1: Locus 2:

Individual 1 AGG GTA CAA Individual 1 ACC GCC TTT

Individual 2 AGG GTC CAA Individual 2 AGC GCC TTT

Individual 3 CGG GTA GAA Individual 3 ACG GCC TTT

Individual 4 CGC GTC GAC Individual 4 ACC GCG TTT

Individual 5 AGG GTA GAA Individual 5 ACC GCC TTC

Compute the following at these two loci:

(a) Theta (Watterson) - 5 points

(b) Theta (Tajima) - 5 points

Locus 1:

# of segregating sites = S = 5

n = number of individuals (sequences) = 5

Theta(Watterson) = = 5/(1/1 + ½ + 1/3 + ¼) = 2.4

Theta(Tajima) = = (1+3+3+4+2+4+1+5+2+1)/10 = 2.6

Locus 2:

S = 4

n = 5

Theta(Watterson) = 4/(1/1 + ½ + 1/3 + ¼) = 1.92

Theta(Tajima) = (1+2+2+2+1+2+2+1+2+1)/10 = 1.6

3. Write a simple script in a language of your choice to simulate genetic drift under a binomial sampling process (Wright-Fisher model). Then vary N (size of the population) as 10, 100, 1000, with starting allele frequency of one allele (p) as 0.2, and plot its allele frequency over 100 generations (10 points).

I wrote mine in R, using the sample() function, but you’re welcome to use whatever language/logic here.

N=10 #Size of the diploid population

allfreqs<-c() #Empty array to store allele frequencies

allfreqs[1]=0.2 #Allele frequency in the first generation, set to 0.2 here = p

plot(1,type="n",xlim=c(0,100),ylim=c(0,1),main="N=100",xlab="Time in generations", ylab="Allele frequency")#Creating an empty plot

for(sims in 1:10) { #Just repeating the below simulations 10 times

for(t in 1:100){ #Loop over 100 generations

gen<-sample(c(0,1),2\*N,replace=TRUE,prob=c(allfreqs[t],1-allfreqs[t]))

#Sampling with replacement (simulating drift)

allfreqs[t+1]<-length(which(gen==0))/(2\*N)

#Compute allele frequency in next generation

}

points(allfreqs,col="red",type="l") #Plot this

}



Similarly, repeat with N = 100, N = 1000. Code stays the same except for the first line. As you see, the smaller the population, the greater the chance of drift.

