Ken Roach
HST.508 PS#1

Problem 1

Fisher-Wright population with selection and mutation

Mutation Rate u1: A1 -> A2 u2: A2 -> A1

Selection $w_{11} = 1$ $w_{12} = 1 - s/2$ $w_{22} = 1 - s$

A) Obtain M(p) due to mutation

Model the number of transitions due to mutation as binomial random variables U_1 and U_2.

$$K \sim \binom{2N}{k} p^k (1-p)^{2N-k} \quad U_1 \sim \binom{K}{j} u_1^j (1-u_1)^{K-j} \quad U_2 \sim \binom{2N-K}{j} u_2^j (1-u_2)^{2N-K-j}$$

$$p = \frac{K}{2N} \quad p' = \frac{K'}{2N} \quad K' = K - U_1 + U_2$$

$$\begin{aligned} E[K'] &= E[K - U_1 + U_2] = E[K] + E[E[U_1|K]] + E[E[U_2|K]] \\ &= 2Np - E[u_1 K] + E[u_2(2N-K)] = 2Np - 2Npu_1 + 2N(1-p)u_2 \end{aligned}$$

$$M(p) = \frac{1}{2N}(E[K'] - E[K]) = (1-p)u_2 - pu_1$$

B) Obtain M(p) due to selection

$$M(p) = \frac{p(1-p)}{2\bar{w}} \frac{d\bar{w}}{dp}$$

$$\bar{w} = w_{11}p^2 + w_{12}2p(1-p) + w_{22}(1-p)^2 = p^2 + 2p(1-p)(1-s/2) + (1-p)^2(1-s)$$

$$\frac{d\bar{w}}{dp} = 2p + 2(1-s/2) - 4p(1-s/2) - 2(1-p)(1-s) = s$$

$$M(p) = \frac{sp(1-p)}{2(p^2 + 2p(1-p)(1-s/2) + (1-p)^2(1-s))} = \frac{sp(1-p)}{2(1-s+sp)}$$

C) Obtain V(p) after one generation

$$\begin{aligned} Var(K') &= E[Var(K'|K)] + Var(E[K'|K]) \\ &= E[Var(K - U_1 + U_2|K)] + Var(E[K - U_1 + U_2|K]) \\ &= E[Var(U_1|K) + Var(U_2|K)] + Var(K - u_1 K + u_2(2N-K)) \\ &= E[Ku_1(1-u_1) + (2N-K)u_2(1-u_2)] + Var(2Nu_2 + (1-u_1-u_2)K) \\ &= 2Npu_1(1-u_1) + 2N(1-p)u_2(1-u_2) + 2Np(1-p)(1-(u_1+u_2))^2 \end{aligned}$$

$$V(p) = Var\left(\frac{K}{2N}\right) = \frac{1}{2N}\left[pu_1(1-u_1) + (1-p)u_2(1-u_2) + p(1-p)(1-(u_1+u_2))^2\right]$$

D) Assuming no drift, find the steady state for mutation alone and for selection alone.

Using the Kolmogorov forward equation with V(p) = 0, steady state is achieved when M(p) = 0.

Mutation: $M(p) = (1-p)u_2 - pu_1 = 0$ therefore $p_{ss} = \frac{u_2}{u_1 + u_2}$

Selection: $M(p) = \frac{sp(1-p)}{2(1-s+sp)} > 0$ for $0 < p < 1$. $p_{ss} = 1$ since zero is unstable.

Problem 2

Obtain an expression for the mean time of fixation or loss starting with the equation below.

$$\bar{t}(x) = 1 + \sum_{\Delta x} \bar{t}(x+\Delta x) Pr(\Delta x)$$

$$= 1 + \sum_{\Delta x} \left[\bar{t}(x) + \frac{d\bar{t}(x)}{dx}\Delta x + \frac{1}{2}\frac{d^2\bar{t}(x)}{dx^2}\Delta x^2 \right] Pr(\Delta x)$$

$$= 1 + \bar{t}(x)\sum_{\Delta x} Pr(\Delta x) + \frac{d\bar{t}(x)}{dx}\sum_{\Delta x} \Delta x Pr(\Delta x) + \frac{1}{2}\frac{d^2\bar{t}(x)}{dx^2}\sum_{\Delta x} \Delta x^2 Pr(\Delta x)$$

$$= 1 + \bar{t}(x)\cdot 1 + \frac{d\bar{t}(x)}{dx}\cdot 0 + \frac{1}{2}\frac{d^2\bar{t}(x)}{dx^2}\cdot Var(x)$$

$$\frac{d^2\bar{t}(x)}{dx^2} = -\frac{2}{Var(x)} = -\frac{4N}{x(1-x)} = -4N\left[\frac{1}{x} + \frac{1}{1-x}\right]$$

$$\bar{t}(x) = -4N\left[x\ln x + (1-x)\ln(1-x) + ax + b\right]$$

Use the simple boundary conditions below to solve for integration constants a and b. The time to fixation or loss is zero if the allele is already fixed or lost.

$\bar{t}(0) = 0$ Therefore $b = 0$

$\bar{t}(1) = 0$ Therefore $a = 0$

Substitution in these values gives our final result.

$$\bar{t}(x) = -4N\left[x\ln x + (1-x)\ln(1-x)\right]$$

Problem 3

For two random alleles from the population, there is a 1 in 4 chance both came from males in the previous generation, in which case the probability of being identical is simply the standard recurrence relation for F but using the number of male alleles $2N_m$. The situation is similar if both alleles came from females, but this time $2N_f$ must be used. There is also a 1 in 2 chance the alleles came from different genders, in which case they could not have come from the same individual and the probability of being identical by descent is simply the homozygosity of the overall population in the previous generation.

$$F_{t+1} = \frac{1}{4}\left[\frac{1}{2N_m} + (1-\frac{1}{2N_m})F_t\right] + \frac{1}{2}F_t + \frac{1}{4}\left[\frac{1}{2N_f} + (1-\frac{1}{2N_f})F_t\right]$$

$$= \frac{1}{4}\left[\frac{1}{2N_m} + \frac{1}{2N_f}\right] + \frac{1}{4}\left[4 - \frac{1}{2N_m} - \frac{1}{2N_f}\right]F_t$$

Comparing this to the general formula using effective population size

$$F_{t+1} = \frac{1}{2N_e} + \left[1 - \frac{1}{2N_e}\right]F_t$$

We get the following formula for N_e.

$$\frac{1}{2N_e} = \frac{1}{4}\left[\frac{1}{2N_m} + \frac{1}{2N_f}\right] \quad \text{Which simplifies to} \quad N_e = \frac{4N_m N_f}{N_m + N_f}$$

Problem 4

A) Region A is likely a protein coding region. If position 1 in the sequence corresponds to the start of a translated codon, all three sites would be in wobble positions, giving them a high probability of being synonymous mutations. The mutations in Region B are evenly spread across the three codon positions and many would be non-synonymous. Region A also has 1/3 as many segregating sites as Region B, suggesting that the non-synonymous mutations may have been selected against and eliminated.

B) Tajima's D score is an appropriate metric to test for selection with this type of SNP data, assuming that the population has been stable over the time period of interest or can be modeled by an effective population. The D scores for the two regions are shown below.

Region A: $S = 3 \quad a_1 = 2.593 \quad \hat{\theta} = 1.157 \quad \pi = 0.750 \quad D = -1.236$
Region B: $S = 9 \quad a_1 = 2.593 \quad \hat{\theta} = 3.471 \quad \pi = 3.357 \quad D = -0.1483$

C) Region A shows a moderate excess of rare alleles, which likely indicates purifying selection in this case, but it could also result from a recent population increase. Region B is quite consistent with the neutral model and shows no evidence of selection.

Problem 5

Region A: 25% nucleotide difference between species
Region B: 5% nucleotide difference between species

A) This scenario is consistent with the neutral theory since the mutation rate is not necessarily the same in the two regions. It is well known that certain regions of the genome are more prone to mutation than others. If the mutation rates were known to be the same in the two regions, the results would suggest a selection process is taking place.

B) Calculate the expected number of segregating sites in a sample of 5 alleles given that 20 sites are observed in a sample of 10 alleles.

The number of segregating sites in the original set of samples can be used to estimate theta. Since it is a constant and does not depend on the number of samples, it can then be used to calculate the expected number of segregating sites in the sample of 5 alleles.

$\theta = 4\text{Nu} \quad E[S] = a_n\hat{\theta} \qquad a_{10} = \sum_{i=1}^{9} \frac{1}{i} = 2.828968$

$S_{10} = 20 = a_{10}\hat{\theta} \quad \hat{\theta} = 7.0697 \qquad a_5 = \sum_{i=1}^{4} \frac{1}{i} = 2.083333$

$S_5 = a_5\hat{\theta} = 14.73$

Problem 6

A simple Matlab script was used to calculate the D values for the provided SNP database. The script will be attached at the end of the problem set. The p values were calculated for a double sided test and assume that the D values follow a normal distribution.

A) Compare the obtained scores for African and European populations.

African: $S = 24189 \quad a_1 = 3.7343 \quad \hat{\theta} = 6477.5 \quad \pi = 4952.2$
$D = -0.9565 \quad p = 0.3388$

European: $S = 14767 \quad a_1 = 3.6908 \quad \hat{\theta} = 4001.0 \quad \pi = 3831.7$
$D = -0.1732 \quad p = 0.8625$

The African population has a much more negative D value than the European population, indicating a greater abundance of rare alleles. This could indicate either stronger selection pressures or a recent population increase. However, neither D value is statistically significant, at least with the normal approximation being used.

B) Repeat using only non-synonymous mutations.

African: $S = 498 \quad a_1 = 3.7343 \quad \hat{\theta} = 133.36 \quad \pi = 89.97$
$D = -1.3154 \quad p = 0.1884$

European: $S = 347 \quad a_1 = 3.6908 \quad \hat{\theta} = 94.02 \quad \pi = 69.52$
$D = -1.0593 \quad p = 0.2895$

Using only non-synonymous mutations, the D values were more negative for both populations. This suggests that non-synonymous mutations are under stronger selective pressure than synonymous or non-coding mutations. This makes intuitive sense, since changing even a single amino acid may have a significant effect on the function of the protein. However, the p values are not statistically significant, so the neutral model good still potentially hold.

C) Repeat using only synonymous mutations.

African: $S = 506 \quad a_1 = 3.7343 \quad \hat{\theta} = 135.50 \quad \pi = 104.96$
$D = -0.9112 \quad p = 0.3622$

European: $S = 304 \quad a_1 = 3.6908 \quad \hat{\theta} = 82.37 \quad \pi = 71.92$
$D = -0.5151 \quad p = 0.6065$

The D values are quite a bit less negative for the synonymous mutations than they were for the non-synonymous mutations. The D value for synonymous mutations in the African population is also very close to the value for all mutations in that population. Since selection is expected to be minimal for synonymous mutations, this suggests that other mechanisms like population change are likely to be responsible for the observed values.

```
% HST.508 PS1
load SNPData.mat;

% Set up constants
nafr = 24;
neur = 23;

% Calculate D for African population
[Safr,a1afr,thetaafr,piafr,Dafr,pvalafr] = tajima(nafr,x5,x9)

% Calculate D for European population
[Seur,a1eur,thetaeur,pieur,Deur,pvaleur] = tajima(neur,x6,x10)

% Include only nonsynonymous mutations
nonsynon = find(strcmp(x3,'NON-SYN     '));
f1afrnon = x5(nonsynon);
f2afrnon = x9(nonsynon);
f1eurnon = x6(nonsynon);
f2eurnon = x10(nonsynon);

[Safrnon,a1afrnon,thetaafrnon,piafrnon,Dafrnon,pvalafrnon] = tajima(nafr,f1afrnon,f2afrnon)
[Seurnon,a1eurnon,thetaeurnon,pieurnon,Deurnon,pvaleurnon] = tajima(neur,f1eurnon,f2eurnon)

% Include only synonymous mutations
synon = find(strcmp(x3,'SYNON       '));
f1afrsyn = x5(synon);
f2afrsyn = x9(synon);
f1eursyn = x6(synon);
f2eursyn = x10(synon);

[Safrsyn,a1afrsyn,thetaafrsyn,piafrsyn,Dafrsyn,pvalafrsyn] = tajima(nafr,f1afrsyn,f2afrsyn)
[Seursyn,a1eursyn,thetaeursyn,pieursyn,Deursyn,pvaleursyn] = tajima(neur,f1eursyn,f2eursyn)


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


% Calculate Tajima's D based on SNP frequencies
function [S,a1,theta,pi,D,pval] = tajima(n,f1,f2)

% Calculate pi using frequencies of alternative alleles
pi = sum(n^2 .* f1 .* f2) / (n*(n-1)/2);

% Only count sites where there is actual diversity in the population
S = sum((f1~=0).*(f1~=1));

% Below calculations are based on handout
a1 = sum(1./(1:(n-1)));
a2 = sum((1./(1:(n-1))).^2);
b1 = (n+1) / (3*(n-1));
b2 = 2*(n^2+n+3) / (9*n*(n-1));
c1 = b1 - 1/a1;
c2 = b2 - (n+2)/(a1*n) + a2/(a1^2);
e1 = c1 / a1;
e2 = c2 / (a1^2+a2);
theta = S/a1;
D = (pi - S/a1) / sqrt(e1*S + e2*S*(S-1));

% P-value for a double sided test assuming a normal distribution for D
pval = 2*normcdf(-abs(D));
```