Biostatistics 666
Problem Set 2
Due October 2, 2017

1. Mutations in the G6PD gene, which maps to the **X chromosome**, are associated with resistance to infection by the malaria parasite. A stretch of 2000bp surrounding the gene was sequenced in a male susceptible to malaria and in another male who appeared resistant to malaria.

   Assume the effective population size is N = 10,000 individuals, that the mutation rate is 10^{-8} per base-pair per generation and that there is no evidence for recombination in the region.

   a) Given the gene is on the X chromosome and there are 10,000 individuals in the population, how many sequences are segregating in the population? State any assumptions about the number of males and females in the population.

   b) What is the expected time to the most recent common ancestor (MRCA) of the two sequences? Please state any assumptions you made for this calculation.

   c) What is the expected number of differences between the two sequences?

   d) When the two sequences were compared, 5 differences were identified. What is the probability of observing 5 or more differences between the two sequences? Could you interpret this result as evidence of natural selection at the locus?

   e) If your model allowed for recombination within this 2000 bp sequence, how might your answer to a), b) and c) above to change?

   f) In general, how do you expect patterns of genetic variation and linkage disequilibrium to compare between the X chromosome and autosomes? Do you expect to see more (or fewer) variants per base pair in one setting – or do you expect both to be about the same? Do you expect to see the same degree of linkage disequilibrium in both settings – or do you expect one to show greater linkage disequilibrium?

2. Consider the following bottlenecked population model:

Historical population size N_e = 10,000 sequences,

Followed by a bottleneck with N_e = 100 sequences
Lasting for 10 generations
And ending 2000 generations ago

Population size after bottleneck N_e = 1,000,000 sequences

To sample a coalescent time for a pair of sequences, consider the following pseudocode:

```
SampleCoalescenceTime()
      {
      TimeToCoalescence = SampleFromExponential(Mean = 1,000,000)

      If (TimeToCoalescence < 2,000)
            Return TimeToCoalescence;

      TimeToCoalescence = 2,000 + SampleFromExponential(Mean = 100)

      If (TimeToCoalescence < 2,010)
            Return TimeToCoalescence;

      TimeToCoalescence = 2,010 + SampleFromExponential(Mean = 10,000)

      Return TimeToCoalescence;
      }
```

a) Using your favorite programming language, implement this code, sample coalescence times for 1,000 pairs of sequences and plot a histogram to summarize their distribution.

b) How does this distribution of coalescence times compare to what you would expect with a constant population size N_e = 10,000 sequences?

c) If you were to count pairwise differences between sequences in a population like this one, what would you expect? How would this result differ from that for a constant sized population with N_e = 10,000 sequences?