Assignment set 3

● *Submit your answers 3. March at the latest. Like for assignment sets 1 and 2, you can work as groups of 2-4 students*

● Note a separate lecture slide set *Coalescence theory and selection tests. In case you find mistakes, please submit a question to course Moodle!*

3.1. DNA sequence data from a population, four individuals as a sample:

A	A	G	A	T	G	A	C	A	G	A	T	A	G	G	C	A
C	T	G	G	T	G	A	C	T	G	A	T	A	G	G	C	A
C	T	G	G	T	G	A	C	T	G	A	T	A	G	G	C	T
C	A	G	A	T	G	A	C	T	G	A	T	A	G	G	C	T

● How many segregating sites are there?
● What is the average number of of pairwise differences (Π) ?
● Calculate two estimates of Θ, one based on Watterson´s estimator and one based on Tajima´s estimator.
● Do the data contain more, fewer, or the same number of singletons as expected under the standard neutral coalescence model?

3.2. One might think that samples from an infinite-alleles neutral model should contain roughly equal numbers of alleles represented. This is not true. The expected sample configurations are very unequal, because the representation of each allele depends on the time in evolutionary history when it was created by mutation and the manner in which its frequency was affected by random genetic drift. To take a specific example, consider a sample of size $n = 6$ from a population evolving according to the infinite-alleles neutral model, and suppose that the sample contains only $k = 2$ different alleles. Let the configuration of alleles in the sample be represented as $(a_1, a_2, a_3, a_4, a_5)$, where a_i is the number of alleles represented exactly i times,with $\sum ia_i = 6$. It can be shown from from *Ewen´s sampling formula* that the probability of the configuration (a_1, a_2, a_3, a_4) equals

$$\text{Prob}\,\{a_1, a_2, a_3, a_4, a_5 \mid k = 2\} = 6! \,/\, (274 * 1^{a1} 2^{a2}\, 3^{a3}\, 4^{a4}\, 5^{a5} * a_1!\, a_2!\, a_3!\, a_4!\, a_5!)$$

In this case only three sample configurations are possible, namely $x = (1, 0, 0, 0, 1)$, $y = (0, 1, 0, 1, 0)$, and $z = (0, 0, 2, 0, 0)$.

● Calculate the probabilities of *x, y* and *z,* and the expected proportion of samples in which the numbers of the two alleles are not equal

3.3. For the infinite-alleles neutral model, the probability that a sample of size $n = 6$ contains exactly $k = 3$ alleles in the configuration (a_1, a_2, a_3, a_4) is given by

Prob $\{a_1, a_2, a_3, a_4 \mid k = 3\} = 6! / (225 * 1^{a1} 2^{a2} 3^{a3} 4^{a4} * a_1!\, a_2!\, a_3!\, a_4!)$

where a_i is the number of alleles represented i times in the sample, and $\sum ia_i = 6$.

● What sample configurations (a_1, a_2, a_3, a_4) are possible and what are their probabilities?

3.4. A 10kb (10 thousand bases) long DNA sequence from a single individual is known and 21 of the sites are heterozygous. The mutation rate in the region is 10^{-9} per site.

● Assuming an infinite sites model, provide an estimate of the effective population size of the population from which this individual has been sampled.

3.5. Five diploid individuals (10 DNA sequences) from a population with an effective population size of 20 000 individuals is known. The mutation rate for the DNA region of interest is 10^{-5} per generation,

● Assuming infinite site mutation model, how many segregating sites should be expected to be found in the data?

3.6. The mutation rate in a particuar gene is 10^{-9} per generation per base (bp). The gene is 800 bp long. Assume that both humans and chimpanzees have a generation time of twenty years, and that each mutation will create a new nucleotide (base) difference between humans and chimpanzees. The divergence time between humans and chimpanzees is 6 million years.

● How many nucleotide differences in this gene would be expected to be observed between humans and chimpanzees?