Math 5233: Measures of genetic diversity and Tajima's $D$-test

Due 4/29. Group members (1 to 3): _____

(0) Consider the following five sequences of bases at six segregating sites ($S = 6$ in a genetic sample.

```
1 2 3 4 5 6
T A C G T C
G C A G C T
G A C T C C
G A C G C C
G C A G C T
```

(1) Calculate the genetic diversity parameter $\pi = (\sum_{i<j} k_{ij}) / \begin{pmatrix} n \\ 2 \end{pmatrix}$ for this sample.

($n = 5$ is the number of samples, $k_{ij}$ is the number of differences between sequence $i$ and sequence $j$.)

(2) Calculate the site-frequency spectra $\eta_i$, where $\eta_i$ is the number of sites with $i$ copies of one base and $n - i$ copies of another base. (For this example $i \in \{1, 2\}$.)

(3) Find a formula for the parameter $\pi$ in terms of the $\eta_i$ rather than the $k_{ij}$. (Hint: think about the contribution to $\pi$ from each column, rather than from two rows.) Doublecheck your formula by using it to calculate $\pi$ for the above example.

(4) (Optional) Tajima's $D$-test is a widely used statistical test for the null hypothesis that sites evolve neutrally in a constant-size population. Under the null hypothesis, the expected value of the two quantities $\pi$ and $S/a_1$ should be equal, where

$$a_1 = \sum_{i=1}^{n-1} \frac{1}{i}.$$

If the null hypothesis were true, we would expect that the coefficient of variation of $d = \pi - S/a_1$ would be small, usually between 0 and 1. Tajima's $D$-test studies a signed version of the coefficient of variation, denoted $D$:

$$D = \frac{d}{Var(d)} = \frac{\pi - \frac{S}{a_1}}{e_1 S + e_2 S(S-1)}$$

where $a_2 = \sum_{i=1}^{n-1} \frac{1}{i^2}$ and

$$e_1 = \frac{c_1}{a_1} \qquad e_2 = \frac{c_2}{a_1^2 + a_2}$$

in which

$$c_1 = b_1 - \frac{1}{a_1} \qquad c_2 = b_2 - \frac{n+2}{a_1 n} + \frac{a_2}{a_1^2}$$

which in turn need

$$b_1 = \frac{n+1}{2(n-1)} \qquad b_2 = \frac{2(n^2+n+3)}{9n(n-1)}.$$

If $D$ is not between $-2$ and 2, it constitutes strong evidence that the null hypothesis is false.

In order to boast to any population geneticists that you ever meet that you have calculated $D$ by hand, calculate $D$ for the example on the previous page.