Homework #2

AUTHOR
Benjamin Zeiger

Question 1

In a sample of 5 gene copies, what is the expected total length of the coalescent tree? (Hint: see slide 5 in the lecture from Jan 28; your answer will be a function of N) .

Answer

If mean length of genealogy is represented by the simplified equation:

$$4N(1+\frac{1}{2}+...+\frac{1}{n-1})$$

Then in a sample of 5 gene copies, the expected length of the coalescent tree should be:

$4N(1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}) = 4N(2.083) =$ ***8.332N generations***

Question 2

A researcher sequences a 10-kb-long DNA sequence from a single individual. The mutation rate in the region is 10^{-9} per site. The researcher finds 21 sites that are heterozygous. Assuming an infinite sites model and a standard neutral coalescent model, provide an estimate of the effective population size (N) of the population from which the individual has been sampled. How would your estimate change if the mutation rate was 10^{-8} per site? (Hint: see slide 6 from the lecture on Jan 28)

Answer

If the expected segregating sites are represented by:

$$S = 4N\mu(1+\frac{1}{2}+...+\frac{1}{n-1})$$

And we assume one diploid individual provides 2 DNA sequences, then we can solve for N:

$$S = 4N\mu(1) = 4N\mu$$

$$N = \frac{S}{4\mu}$$

so we would expect 21 segregating sites if the effective population size is:

$N = \frac{21}{4*10^{-9}} = 5.25*10^9$ or ***5.25 billion***.

A mutation rate of 10^{-8} would reduce N by a factor of ten (525 million).

Question 3

A researcher sequences 5 diploid individuals (10 DNA sequences) from a population with N=20,000 individuals. The total mutation rate for the region is 10^{-5} per generation. Assuming a standard coalescent model and infinite sites mutation, how many segregating sites should the researcher expect? (Hint: see slide 6 from the lecture on Jan 28)

Answer

Using our previous equation, we expect:

$$S = 4(20000)(10^{-5})(1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{9})$$

```r
vec <- seq(1,9,1)
vec2 <- 1/vec

a1 <- sum(vec2)
popN <- 20000
mu_rate <- 10^(-5)

S <- (4 * popN * mu_rate * a1)

S
```

```
[1] 2.263175
```

$$S = 2.263$$

The researcher should expect two segregating sites, possibly three.

Question 4

Consider the following sample of 6 gene sequences, with the alignment below showing only the variable positions along a sequence of 1000 bases.

Sample	Sequence
1	AAGCCTGTGT
2	AAGCCTGTAT
3	AAGCTTGTAT
4	AGATTTACAC
5	TAATTCACAC

Sample	Sequence
6	TAATTCACAC

a. Calculate the proportion of segregating sites (S) and the nucleotide diversity (= the average number of pairwise mismatches).

Answer

Allele	1	2	3	4	5	6	7	8	9	10
a	A	A	G	C	C	T	G	T	G	T
b	A	A	G	C	C	T	G	T	A	T
c	A	A	G	C	T	T	G	T	A	T
d	A	G	A	T	T	T	A	C	A	C
e	T	A	A	T	T	C	A	C	A	C
f	T	A	A	T	T	C	A	C	A	C
	8	5	9	9	8	8	9	9	5	9

There are 10 segregating sites here. The sum of pairwise mismatches is 79, and the nucleotide diversity is $(\frac{79}{10} = 7.9)$

b. Give two estimates of θ that can be derived from these data.

Answer

Waterson’s estimate of θ:

$$\hat{\theta}_W = \frac{S}{a_1} = \frac{10}{2.283} = 4.38$$

Tajima’s estimate of θ:

$$\hat{\theta}_\tau = \frac{k}{n(n-1)/2} = \frac{79}{15} = 5.27$$