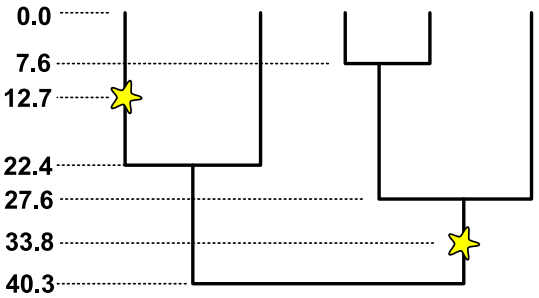
COMPUTATIONAL POPULATION GENETICS — EXERCISE SHEET 2

1. Calculate with your pocket calculator (or only using operations on your computer that also exist on your pocket calculator) Watterson's θ_W and Tajima's π for the following dataset:

```
ACGCTTCTATTCTTATTAACCAAACGGGTTGCAACTCTCTAGGA
...T.........................A.............
........G.............................T.......
........G......G..............A.....T.......
..A.....G......G..............A.....T.......
```

(Dots stand for the same nucleotide as in top line.)

2. Five gametes have been sampled from different individuals of population of an effective size of 10,000 indviduals.



A diploid gene locus of a length of 1 kb has been sequenced. The figure to the left shows a hypothetical genealogy of the locus, with the stars indicating mutations that occurred in this gene locus. The time scale is in units of 1,000 generations before present.

   (a) Define the time scale for which the rate of coalescence is 1.0 for each pair of lineages and convert the given times accordingly.

   (b) Assume that $\theta = 4N_e\mu = 5$, where μ is the mutation rate per generation and per kb. Calculate the probability density of the hypothetical genealogy with mutations as shown in the Figure.

3. A researcher wants to calculate the total bio mass of five species A, B, C, D, E of known population sizes 1000 (A), 800 (B), 1100 (C), 300 (D) and 700 (D). Thus, the total number of individuals is 3900. A sample of 130 individuals was taken, and their total mass was 58121 kg. Within each species sampling was representative but the sample sizes for the species did not reflect the population sizes. The following table shows the total mass of individuals sampled from each species, sample sizes per species, and the standard deviation of the weights from the sample.

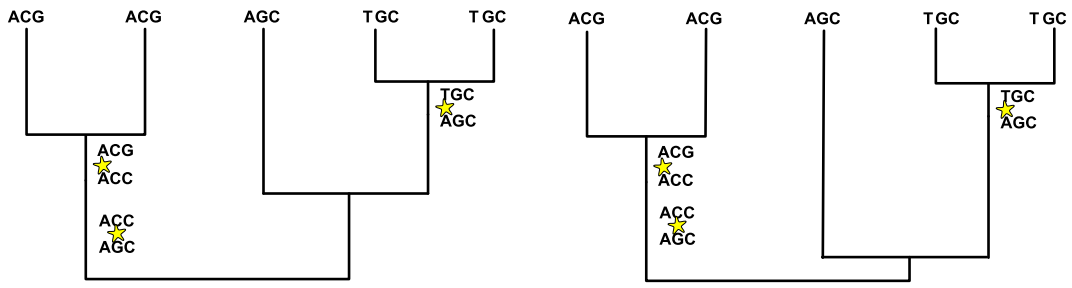
	A	B	C	D	E
total mass of sampled individuals	57	4272	76	53270	446
sample size	10	80	20	100	20
sd	0.8	6.3	0.7	64.7	3.4

   (a) Calculate a reasonable estimator for the total biomass of the five species, correcting for the bias of the sample sizes.

   (b) Discuss whether this was a clever sampling scheme or whether sample sizes should have been the same for all five species or proportional to the population sizes.

(c) What does this have to do with the contents of the lecture? Compare your calculation of the total biomass to the importance sampling formula.

4. The following two trees represent two out of 100 population genetic histories that were sampled in a Griffiths-Tavaré importance sampling scheme with an initial θ_0 of 1.0, assuming an infinite-sites mutation model and a constant-size panmictic population. Only segregating sites of the data are shown at the tips of the trees.



How much does each of the two histories contribute to the likelihood calculations of $\theta = 2.0$ and $\theta = 0.5$?

5. Two populations of size N exchange on average m migrants per generation in each direction. Assume that you randomly sample two alleles of an autosomal locus from the same island, and let T be the time to their most recent common ancestor, measured in units of $2N$ generations (neglect recombination). Let T' be the time to the most recent common ancestor for the case that the two alleles were sampled from different islands. Determine the expectation values $\mathbb{E}T$ and $\mathbb{E}T'$ (and if you are really ambitious also the standard deviations σ_T and $\sigma_{T'}$) by mathematical derivation or by computer simulation (e.g. with scrm or Hudson's ms) for all combinations of $N \in \{1000, 10^5, 10^8\}$ and $m \in \{0.1, 10, 1000\}$.

6. Simulate a dataset with 10 independent autosomal loci of lenght 1000 bp, 20 alleles sampled from each of two populations of $N_e = 10.000$ individuals that exchange 1 indivual in each direction every 5 generations. Use a DNA mutation model with double hits and $\theta = 10$ per locus. How accurately can LAMARC estimate θ and the migration rate? Explore also how long you should run LAMARC.

7. Simulate datasets of DNA loci sampled from three populations that exchange migrants at certain rates. Use pairwise migration rates lower than 1 migrant per generation. Explore how the accuracy of LAMARC estimates for the migration rates depends on the sample sizes, the number of available independent loci (assuming $\theta \approx 10$ per locus) and the LAMARC runtime. Use a mutation model that allows for double hits.

Due to the runtime of LAMARC, it may not be possible to present final results for exercises 6 and 7 within a week. However, you should perform preliminary test runs to estimate with which options and how long you need to run LAMARC, and start these final runs. Be prepared to present your approach and your initial findings next week.