D. Metzler, WS 2022/2023

COMPUTATIONAL POPULATION GENETICS — EXERCISE SHEET 4

1. Assume you simulate an ancestral recombination graph (ARG) for a n whole chromosomes sampled from a neutral, constant-size population (with positive recombination rate). From all the trees that this ARG assigns to single nucleotide positions, you choose four:

   Tree A is the tree at position 1000.

   Tree B is the next tree after Tree A, that is, you move on from position 1000 until there is a recombination event somewhere on a branch of tree A. Tree B is then the tree of the nucleotide position right after the recombination event.

   Tree C is the 100th tree in the list of all trees appearing from left two right, that is, the tree after the 100th recombination event (where only recombination events that effect the current tree are counted).

   Tree D is the next tree after Tree C, that is, you move on until there is a recombination event somewhere on a branch of tree C. Tree D is then the tree of the nucleotide position right after the recombination event.

   Which of the following statements (a) to (k) are true? Substantiate your answers either logically or with computer simulations. (Start with $n = 2$ and $n = 3$). If you rely on computer simulations, try to find logical explanations for your observations.

   (a) The probability distribution of Tree A is that of a standard Kingman coalescent.

   (b) The probability distribution of Tree B is that of a standard Kingman coalescent.

   (c) The probability distribution of Tree C is that of a standard Kingman coalescent.

   (d) Tree B has the same probability distribution as tree A.

   (e) The expected total branch lenght of A is smaller than that of B.

   (f) The expected total branch lenght of B is smaller than that of A.

   (g) Tree C has the same probability distribution as tree A.

   (h) The expected total branch lenght of A is smaller than that of C.

   (i) Tree D has the same probability distribution as tree C.

   (j) The expected total branch lenght of C is smaller than that of D.

   (k) The expected total branch lenght of D is smaller than that of C.

2. You want to estimate $\theta = 4N_e\mu$ and the population growth rate g for a population from which you have sampled genetic data and calculated two summary statistics s and r. The values for the original data are $s_0 = 8.2$ and $r_0 = 4.6$. To carry out an ABC analysis you simulated five datasets according to the population model with parameter values that you sampled from your prior and calculated the summary statistics. The results were as follows:

θ	g	s	r
5.3	1.6	5.2	5.8
8.4	1.2	7.9	5.6
12.6	0.8	8.9	4.6
3.1	1.4	5.7	8.4
15.0	1.1	10.1	1.6

Do the following exercises without any ABC software package. (But you can of course use general R functions.)

(a) The very classical ABC method is applied to the data, with the euclidean distance as a distance measure on the summary statistics (without normalization) and a threshold of 2. Which combinations of parameter values are in the ABC posterior sample?

(b) ABC with local regression correction is applied with an Epanechnikov kernel with $\delta = 4$ for the local regression part. Which combinations of parameter values are used for the local regression and with which weights?

(c) Calculate for each of the two summary statistics the intercept and the slopes of the local-regression corrections. (Hint: you can do this either with matrix algebra as shown in the lecture or by using the `lm` command in R, which has the optional parameter "`weights`".)

(d) Use the results of the local regression to calculate the local-regression corrected values of the ABC-sampled parameter combinations.

(e) Calculate the approximated posterior distribution density of θ at the value 10 with the above local-regression ABC results and a bandwidth of $\Delta = 2.5$.

(f) Visualize the approximated posterior distribution density function of the population growth rate g with the above local-regression ABC results and a bandwidth of $\Delta = 0.25$.

3. Two populations emerged from a recent split of an ancestral population. There may still be a little amount of gene-flow between the populations.

(a) Explore for different sizes of datasets (start with 10 independent gene loci) how accurately ABC and at least one of IM/IMa, Lamarc and Jaatha can estimate the model parameters θ for the ancestral and the two descendant populations, time of populations split, migration rates and population growth rates.

(b) How can we deal with recombination within loci? Does recombination within loci increase or decrease the accuracy of parameter estimations?

4. Simulate datasets of sequence data sampled from two populations that stem from a recent joint ancestral popultion and compute the JSFS of the data (e.g. with the R package coala). Explore (e.g. by averaging over many simulations) how the expectation values of the entries of the JSFS depend on parameters like the time since the split, rates of geneflow between the population, population size ratios and population growth rates.

5. Four sequences have been sampled for each of two populations, and additionally an outgroup sequence from a closely related different species. Only sites that are segregating within the two populations are shown here:

```
sequence_1  pop_1  ACGGCAGCGAATGGGCTCA
sequence_2  pop_1  ACGATAGCGGGTGGTCCTA
sequence_3  pop_1  AGGGCAACGGGTGTTCCTA
sequence_4  pop_1  AGCGCAACGGGCGTTCTTG
sequence_5  pop_2  ACCGCGATAGGCGTTCTTA
sequence_6  pop_2  GCCGCGATAGGCGTTCCTA
sequence_7  pop_2  ACGGCGATAAGCGTTCTTA
```

```
sequence_8  pop_2  ACGGCGATAGGCGTTTTTG
outgroup           AGGGTGGCAGGCGTTCTTA
```

(a) Calculate the JSFS of this dataset.

(b) Calculate Tajima's π and Watterson's θ_W from the JSFS.

(c) Calculate separately for each of the two populations Tajima's π and Watterson's θ_W from the JSFS.

6. Implement ABC-PMC in R for a demographic model like in the above exercise and explore the performance of this approach. (For the actual ABC steps you can use the abc command from the abc package.)