Assignment 9 – simulating genealogies

The coalescence theory provides a framework to connect genealogies of a sample with population properties. the most simple way to generate a coalescent tree is the following recipe (it might need some adjustment for Java)

1. Set the parameter: the population size N, the mutation rate μ, and the number of individuals in the sample n. In principle we could combine N and μ, but we don't want to do this here because the exercise should help to explore the interaction of n and μ.
2. Generate an array of n nodes
3. Draw a time to coalescence based on
$$\text{Prob}(u_k|N,\mu,k) = \exp\left(-u\frac{k(k-1)}{4N\mu}\right)$$
where k is the number of lineages (samples), for the first draw of a time $k = n$; simple rearrangement allows to get a time
$$u_k = -ln(r)\frac{4N\mu}{k(k-1)}$$
We know now that the next coalescent event happens u_k in the past
4. Pick two nodes at random from the array, create a new node connect the two nodes with it, add the time-interval u_k to the time of the node. then replace the first of the two picked nodes with the new node and shift the nodes to the right of the second node to the left. The nodes are now not part of the array anymore, but are still accessible through the newly formed node. Now our array is one element shorter than we started.
5. Continue at 2 until the array has only a single element left (this is the root of the tree).
6. Once the tree is in memory, simulate data on or print etc.

- Generate coalescent trees for a set of population sizes: 10000, 1000000 and 3 different mutation rates: 10^{-8},10^{-6}, this will be a 2x2 table, for each cell create at least two trees (better a hundred) and then simulate data on these trees. The more sites the better the result, with the small size and low mutation rate there is a good chance that in 500 base pairs there is no variable site.
- Count the number variable sites (averages). Describe the differences or similarities between your 4 cells.