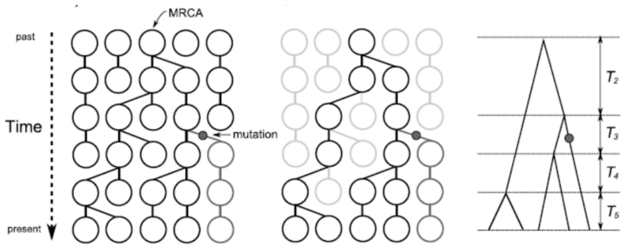
Tree theory: simulating data from stochastic processes

Lucas C. Monteiro - fc52849@alunos.fc.ul.pt

1.1 getTMRCA function

We started by creating the getTMRCA function that simulates the coalescent time of a sample of $\boldsymbol{n}$ lineages from a population with an effective size of $\boldsymbol{Ne}$ individuals.



$$n \to (n-1) \to (n-2) \to \cdots \to 2 \to 1\,(MRCA)$$

Figure 1. Representation of a coalescent process of n lineages at present time to one *MRCA* single gene, T_{MRCA} generations ago. The process requires $n-1$ coalescent events, $n \to n-1 \to \ldots \to 1$, therefore $T_{MRCA} = T_n + T_{(n-1)} + \ldots + T_2$.

The coalescent process can be seen as a process going from n genes at present time to a single gene sometimes in the past through a series of coalescent events. Theory says that each generation has a probability $\boldsymbol{p_n}$ of having a coalescent event therefore we can simulate a coalescent time $\boldsymbol{T_n}$ by generating a geometrically distributed random value. This probability, of n gene lineages passing to $n-1$ lineages is given by Equation 1.

$$p_n = \binom{n}{2}\frac{1}{2Ne} = \frac{n\,(n-1)}{4Ne} \qquad (1)$$

Note that the randomly generated geometrically distributed value is a valid simulation for the waiting time, in generations, of *one* coalescent event. However, the coalescent process until the ***Most Recent Common Ancestor***, or *MRCA*, of a sample of n lineages, consists of $n-1$ coalescent events, as represented in Figure 1, so $n-1$ coalescent event simulations are needed to simulate a full coalescent process. The respective coalescent times are added, yielding the total coalescent time estimation of the process $\boldsymbol{T_{MRCA}}$.

1.2 T_{MRCA} distribution

Having a population coalescent time simulator function, we can now simulate the time until MRCA distribution. In this exercise we simulate the T_{MRCA} distribution of one diploid individual, i.e $n = 2$ gene copies, in a population with an effective size of $Ne = 1\,000$. For this, we perform $10\,000$ coalescent simulations using *R*'s *replicate()* function. The distributions were then represented in histogram and density plots, as seen in Figure 2.

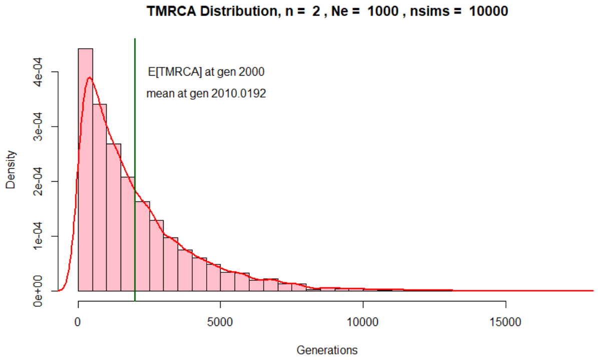


Figure 2. T_{MRCA} distribution in generations for one diploid individual, $n = 2$ genes, in a population with $Ne = 1000$. The histogram and density functions were achieved by running $10\,000$ coalescent time simulations and the mean and theoretical expected values are represented in dark green and blue lines, respectively.

In order to verify the trustworthiness of our results, we compare the mean of the distribution with the distribution's theoretical expected value, $\boldsymbol{E[T_{MRCA}]}$. In a population with a constant effective size Ne, from which we sample n lineages, the expected mean T_{MRCA} is given by Equation 2.

$$E\left[T_{MRCA}\right] = 4Ne\left(1 - \frac{1}{n}\right) \qquad (2)$$

Analysing Figure 2's mean and theoretical (mean) expected values, i.e $2\,000$ and $\approx 2\,010$ generations respectively, we see that they do not vary significantly from each other and conclude that our stochastic simulations represent well the coalescent process and yield a good T_{MRCA} estimation.

1.3 Sample and effective size effects

Now, having a reliable simulator, we can study how sample size and effective size influence T_{MRCA} distribution. We simulated T_{MRCA} distributions for different combinations of $n_{ind} = \{4, 10, 20\}, \Rightarrow n = \{8, 20, 40\}$, since we are studying diploid individuals, and $Ne = \{1000, 10000\}$. The density functions of the distributions, alongside the respective means, were then plotted as represented in Figure 3. Table I contains the means of the different distributions.

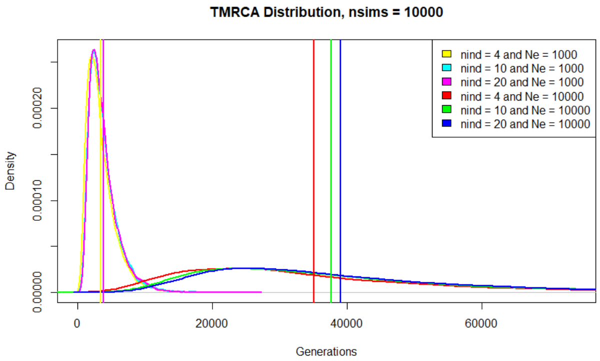


Figure 3. T_{MRCA} distributions for different combinations of n genes and Ne, in generations. The density functions were achieved by running 1 000 coalescent time simulations and the mean and theoretical expected values are represented in the respective color.

Analysing Figure 3 and Table I, it's clear that altering the effective size of the population alters substantially the distribution. The larger the Ne, the larger the mean T_{MRCA}, increasing linearly by factor of 4, a result that we can also obtain theoretically by Equation 1. Also, decreasing Ne seams to narrow the distribution thus making it more precise.

We can interpert this effect as smaller tree will coalesce more rapidly and doesn't have much possible coalescent events configurations until $MRCA$. A larger tree will coalesce more slowly and have more different configurations of coalescent events unitl $MRCA$, thus yieldind a wide and moved to the right T_{MRCA} distribution.

We see that increasing sample size doesn't effect much the T_{MRCA} distribution, however slightly increases its mean, especially for larger Ne values. This makes sense since the *effective* size of the population remains unchanged so the coalescent time its effectively the same. We could sample two individuals that, by chance, were close in the tree (are similar) so that the coalescent time between the two seams small in relation to the whole population T_{MRCA}. This, however, happens for small sample sizes since the probability of happening decreases significantly the larger the n.

Table I. Means of T_{MRCA} distributions for different sample and effective population sizes, in generations.

	nind = 4	nind = 10	nind = 20
Ne = 1000	3 494.08	3 878.539	3 846.774
Ne = 10000	33 635.78	37 584.659	38 096.886

For example, the $1 - \frac{1}{n}$ factor of Equation 2 is at 90% for $n = 10$. Here, increasing n will not change the factor much and preserve our results. So no further sampling effort is necessary since we would be sampling more and more similar individuals. Besides that, the factor is so high that the stochatic noise can easily overtake it, especially for smaller populations. This can be noticed in Table I at $Ne = 1000$, where the mean T_{MRCA} is larger for $n_{ind} = 10$ than for $n_{ind} = 20$.

2 Chimpanzee Ne estimation

We seek to estimate the effective size of two subspecies of chimpanzees - Central chimpanzees (*Pan troglodytes troglodytes*) and Western chimpanzees (*Pan troglodytes verus*) - based on the data from de Manuel et al. (2016) published in Science, using *Approximate Bayesian computation* (ABC) methods.

We first define the prior distribution for the parameter, i.e Ne, by generating 10 000 uniformely distributed random values between 10 and 100 000. To get the posterior distribution we simulate trees with mutations from our observed data sample size, number of sites and mutation rate. These variables are equal for both chimpazee sub-species, only varying the number of segregating sites.

For each simulation, we save the number of segregating sites and calculate the differences between number of segregating sites in the simulated data and the observed one. Keeping only the smallest ***tolerance***, or *tol*, percentage distances respective number of segregating sites, we obtain our posterior distribution and plot it alongside the prior as in Figure 4.

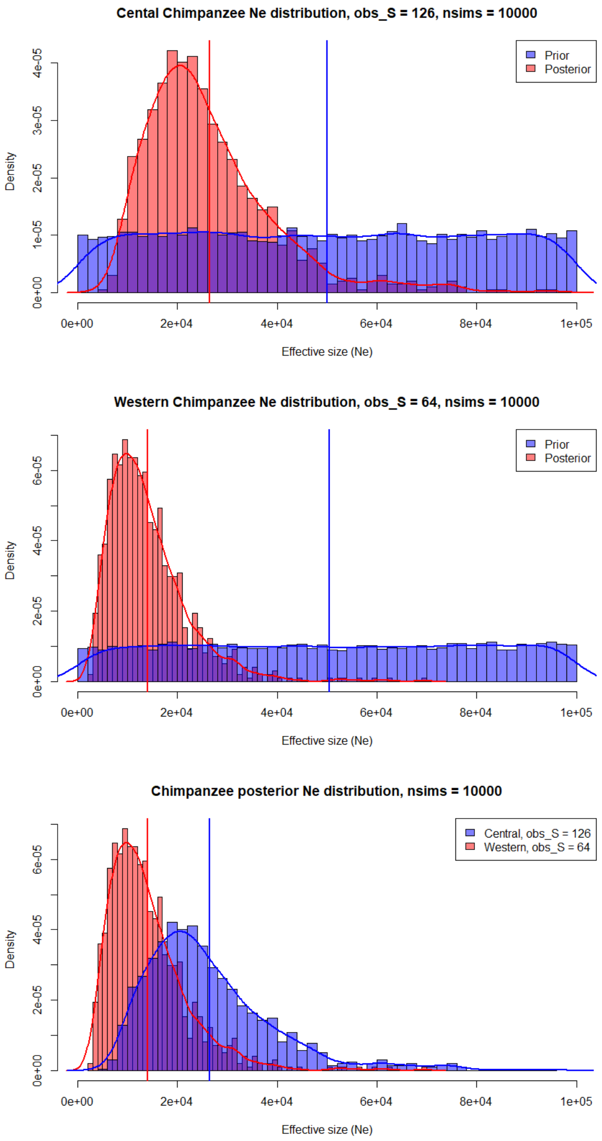


Figure 4. Histogram, density functions and distribution mean of simulated Chimpanzee population effective size distributions. *Top panel:* Prior and posterior distributions of Central chimpanzee sub-population. *Middle panel:* Prior and posterior distributions of Western chimpanzee sub-population. *Bottom panel:* Posterior distributions of Central and Western chimpanzee sub-populations.

We see that the observed data changes our prior to a nice skewed distribution, indicating that their was enought information in the data, i.e we simulated trees that matched our observation of segregating sites, in this particular case, yielding the estimated parameters, of population effective size. If the observation was fully disconnected from the simulations the distribution would take other shape, and would not give useful information about the estimated parameters.

Table II. Quantiles of the simulated posterior distributions of western chimpanzee sub-population effective size with different tolerance levels.

Quantile:	2.5%	97.5%
tol = 5%	1 957.373	37 414.932
tol = 20%	1 830.484	38 436.915

The distributions observed at Figure 4 show a desviated to the right central chimpanzee sub-population and with higher expected value, i.e mean, therefore a potencial larger effective size estimate, comparing with their western counterparts. This was expected since the number of observed segregating sites was larger.

Studying how the tolerance could affect our results, we varied the *tol* of our method to 5% and 20%, using the data of the western chimpanzee sub-population. As the method rejects the tolerance nearest simulated data of the observed one, we expect that increasing the tolerance allows more data to be accepted for the posterior thus yielding a broader distribution, being the limit situation where $tol = 1$ and all data is accepted so the posterior is equal to the prior.

And this is what we got as seen in Table II. The 2.5% and 97.5% quantiles were smaller and larger, respectively, with higher tolerance, evidencing a broader distribution. Notheless, in this case we increased the tolerance by 15% and the results were similar so we can say that for small tolerance levels the results aren't much affected. However, large tolerance levels lead to posterior distributions similar to prior so no results at all.