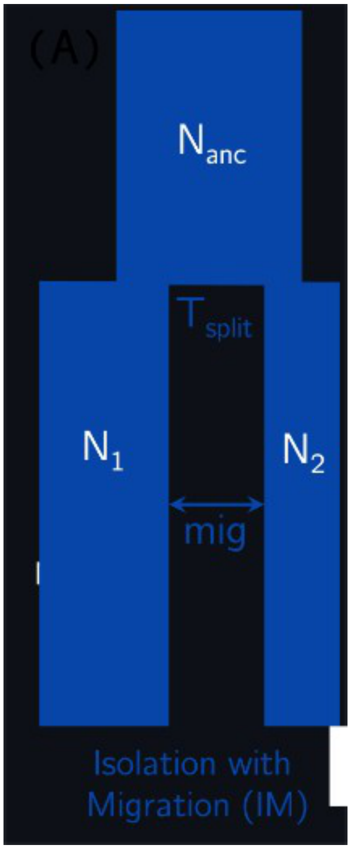
Practical – Day 3 (ABC)

We are analysing genomic data from two populations of *Anopheles gambiae* from two different locations. We have 50 diploid samples per location. We wish to estimate their demographic history following an Isolation-with-Migration (IM) model.



Task 1 (morning): Based on the following information, decide on suitable prior probabilities and perform random samples from them. You can use R or python but you need to check that you are covering the whole range of all prior distributions. Assume a mutation rate of 3.5e-9, a recombination rate of 8.4e-9, and a sequence length of 1,000 bp.

- **T_split:** We know that these populations have been separated by at least 1,000 generations and a major environmental change between these two locations happened around 8,000 generations ago.
- **N1/N2:** We know from capture-recapture data that population 1 is approximately 30 times larger than population 2; we also don't have a clear intuition of each value but from previous findings we expect values for population 1 to be between 50,000 and 200,000.

- **mig:** We expect either complete isolation after the split (rate=0) or pervasive migration (rate=0.1).
- **Nanc:** We know this value with high confidence to be around 7,000,000.

I suggest to start choosing one value for each parameter and run simulations under this model. Once this works well, then try to impose some distributions on each parameter as indicated and perform several simulations, say 100 or 1,000, by jointly randomly sampling from each distribution.

To use `msprime` for your simulations, start with the examples from yesterday and modify them accordingly. You can look for additional instructions from the documentation of msprime to implement your models: additional demographic models:
`https://tskit.dev/msprime/docs/stable/demography.html` other examples:
`https://tskit.dev/tutorials/popgen.html`.

You also need to implement how to draw random samples from a probability distribution using numpy package in python. For instance, draws from finite array can be done as
`https://numpy.org/doc/stable/reference/random/generated/numpy.random.choice.html` or from a uniform distribution as
`https://numpy.org/doc/stable/reference/random/generated/numpy.random.uniform.html` or from a Normal distribution as
`https://numpy.org/doc/stable/reference/random/generated/numpy.random.normal.html`

Task 2 (afternoon): Assume that the observed summary stats are in 'mosquito-observed.csv'

Fst, dxy, segsites1, segsites2, pi1, pi2, tajima1, tajima2

0.2134, 0.0978, 0.3797, 0.1013, 0.0914, 0.0355, 0.2847, 2.0788

These values are: F_ST (Fst) between the two populations, divergence between the populations (dxy), # segr. sites in each population (segsites1, segsites2), π in each population (pi1, pi2), and Tajima's D in each population (tajima1, tajima2).

Estimate the parameters of the IM model using ABC. As first trial, assume that we know all parameters but **T_split**. In particular, assume that there is no migration and **N1**=150,000 and **N2**=5000. Produce the posterior distribution of **T_split**.

You can look at this link to calculate summary statistics using msprime
`https://tskit.dev/tskit/docs/stable/stats.html#sec-stats`

I suggest that you simulate the datasets and compute the summary statistics in `Python`, and write the values to a file. Then load the values from the file in `R`. Inspect whether the range of the simulated summary statistics include the observed ones, how the summary statistics correlate with each other and the parameters, and then use the function `abc` from the library “abc” to perform the abc-estimation.

Task 3 (afternoon): Attempt to jointly infer **T_split**, **N1** and **N2**.