Here is a series of high-level unit tests to validate various parts of the simulation software implementation.
These tests are based on classical population genetics results.
Note 1: This probably needs to be completed by low-level unit tests at the function level.
Note 2: The current text file is expected to be transformed in something more useful (Jupyter notebook or documentation
for the unit test module)

* Test A1: Genetic drift and random mating; new mutation

- Starting population: A population of N diploid individuals, with a unique TE (no selection, no transposition)
inserted randomly in a unique individual.
- Simulation setting: a large number R > 10000 simulation runs. Each run should last long enough to reach loss / fixation of the TE (in practice, 100N generations should be enough). Try at least over 2 orders of magnitude for N : N=10, N=100, N=1000
- What to track:
-- 1) frequency of the runs in which the TE has been fixed (= present in all individuals of the population) = r/R
-- 2) number of generations before loss or fixation
- Expected theoretical results: (from Crow & Kimura 1970, page 432)
-- The TE should reach fixaton with a frequency 1/2N (exact result)
-- In simulations where the TE is lost, the average time before lost should be 2log(2N) generations (approximated result)
-- In simulations where the TE is fixed, the average time before fixation should be 4N, with a standard deviation or 2.15N (approximated result)

* Test A2: Genetic drift and random mating; intermediate frequencies

- Starting population: A population of N diploid individuals, with identical TEs (no selection, no transposition) inserted at a random site at a given frequency 1/2N <= p <= 1-1/2N. The initial distribution of copies among individuals should follow Hardy-Weinberg frequencies, i.e. p^2 individuals should be homozygous for the TE insertion, 2p(1-p) should be heterozygous, and (1-p)^2 should not have any TE copy.
- Simulation setting: a large number of simulation runs. Same idea as in A1. The test should explore various N and various p (e.g. p=0.1, p=0.25, p=0.5, p=0.75, and p=0.9)
- What to track: the same as in A1
- Expected theoretical results (from Crow & Kimura 1070, page 431):
-- The TE should reach fixation with frequency p (exact result)
-- In simulations where the TE is lost, loss should happen in average after -4N(p/(1-p))log(p) generations (approx)

-- In simulations where the TE is fixed, fixation should happen in average after (-1/p)(4N(1-p)log(1-p)) generations (approx)

* Test B1: Genetic drift, random mating, and recombination

- Starting population: A population of N diploid individuals, with two elements (no selection, no transposition) inserted in two consecutive loci with a recombination rate c, in the same chromosome of the same individual.
- Simulation setting: similar to A1 and A2. Test with various N and various values for c (e.g. c=0, c=0.1, c=0.5)
- What to track : the frequency of the fixation probabilities of all four combinations: g1 = Both TEs, g2 = TE1 alone, g3 = TE2 alone, g4 = no TEs, with g1 + g2 + g3 = g4 = 1.
- Expected theoretical results, from Ohta 1968, Theor Appl Genet 38:243--248 (exact?):
-- g1 = 1/2N - c(1-1/2N)/(2Nc+1)
-- g2 = g3 = c(1-1/2N)/(2Nc+1)
-- g4 = 1-1/2N-c(1-1/2N)/(2Nc+1)
- Note: we would have more statistical power with initial frequencies larger than 1/2N, but initializing the starting population controlling the initial linkage disequilibrium might not be trivial.

* Test C1: Genetic drift, selection

- Starting population: a population of N diploid individuals with an TE inserted at a random site at a frequency p (see A2). The only difference is that the TE now has an effect s on fitness.
- Simulation setting: similar to A2. Try different values for s: s = +0.1, s=+0.01, s=-0.01
- What to track: fixation and loss frequencies
- Expected theoretical results (from Crow & Kimura 1970, page 426):
-- fixation probability = (1-exp(-4Nsp))/(1-exp(-4Ns))) (approximation)
- Note: negative values of s may easily lead to fixation probabilities close to 0, there might be little power to explore the parameter space when Ns is very negative.

* Test D1: transposition

- Starting population: a population of N (large N, to limit genetic drift) individuals with n0 TEs randomly inserted anywhere in the genome of all individuals. TEs have no fitness effect, and a transposition rate u per copy per generation.
- Simulation setting: short simulations will probably be enough (in long simulations, TE copy number will increase without limits).
- What to track: the average number of TEs per individual
- Expected theoretical results
-- The TE copy number should increase exponentially: log(n) = a t + b, where t is the time in

generations, a = log(1+u), and b = log(n0).

* Test D2: transposition and selection

- Starting population: the same as in D1, but TEs now have a fitness disadvantage of -s
- Simulation setting: the same as in D1, perhaps longer simulations if an equilibrium can be reached
- What to track: the same as in D1
- Expected theoretical results (approximation from Charlesworth & Charlesworth 1983)
-- The result depends on the fitness model, and only holds in large populations.
-- For multiplicative fitness, the dynamics is exponential : log(n) = a t + b, where t is the time in generations, a=log(1+u+s), and b = log(n0).