Steps to simulate data:

I have implemented the model given in our recent work [1]. There were mainly four steps:

- **(1) Firstly, I simulated the total number of bases or alleles at each site i, i.e., n_T^i:**

  As we know that number of times a base is sequenced follows a **Poisson** distribution, when reads are distributed randomly across the genome, I drew n_T^i for each site coming from a **Poisson** distribution with parameter **DoC (Depth of Coverage).**

  Therefore, the average value of n_T^i will be equal to DoC.

  Outcome of this step: n_T^i for each site.

- **(2) Fixing/controlling the contamination fraction or the number of alleles or bases at each site coming from the contaminating population ($PopC$):**

  In order to fix the contamination fraction, I drew the number of alleles or bases from $PopC$ for each site, from a **binomial distribution** with parameter (n_T^i, c).

  I chose the binomial distribution here, since there exists only 2 possibilities, i.e., either the base will be from $PopC$ or from the endogenous individual. The other underlying assumption is that there exists a large pool of contaminants.

  Outcome of this step: fixing the contamination fraction (c) or $nCont^i$ which represents the number of bases at site i that are from the $PopC$.

- **(3) Simulating the counts of naturally-segregating alleles A and C for each polymorphic site i, present among $nCont^i$:**

  Among $nCont^i$, the number of allele A will follow a **binomial** distribution with parameter $(nCont^i, f_A^i)$. f_A^i is the frequency of allele A at site i.

  Choosing f_A^i (frequency of the allele A in the contaminating population):

  I considered two choices for the frequency distribution: uniform and a power law.

  Outcome of this step: number of allele A and allele C, i.e., n_A^i and n_C^i among the bases or alleles coming from $PopC$ at each site, i.e., present among $nCont^i$

- **(4). Simulating the counts of naturally-segregating alleles A and C present in endogenous DNA:**

  I drew a random number from a **uniform** distribution (min = 0, max = 1) for each site i.

  Depending on whether this random number is greater than 0.5 or not,

  $n_A^i = n_A^i + n_T^i - nCont^i$, and $n_C^i = n_C^i$

  OR
  $n_C^i = n_C^i + n_T^i - nCont^i$, and $n_A^i = n_A^i$

  Outcome of this step: number of allele A and allele C, n_A^i and n_C^i among the bases or alleles coming from the endogenous DNA

  OUTCOME OF ABOVE 4 steps: n_A^i, n_C^i, and f_A^i for each site i.

  Note that: n_A^i+ $n_C^i = n_T^i$, (Zero Error Case).

- In order to get the value of contamination rate I optimized the **likelihood function** derived in our recent work [1], to get the c_{mle}.

[1]. https://doi.org/10.1093/bioinformatics/btz660