13.1 3 Fundamental Questions of Genome Assembly

There are 3 fundamental questions we want to answer regarding genome assembly:

1. What is the mean proportion of the target genome region covered by contigs?
2. What is the mean number of contigs?
3. What is the mean contig size?

Let the following variables and conditions hold:

- $c$ is the coverage of a DNA base, which is a Poisson distributed random variable
- $G$ is the length of the target DNA region
- $N$ is the number of reads each of which is of length $L$
- It follows that
  \[ c = \frac{NL}{G} \]
  - We say that $L << G$, meaning we ignore edge-effects.

Our goal is to define a stochastic process for genome sequencing and assembly. The stochastic process is one in which we work with the idea that fragments/reads are taken at random from the target DNA region as follows: the left endpoints of the reads are independently and uniformly distributed over the interval $[0, G]$. This means that any such left endpoint falls in an interval $(x, x + h)$ with probability $\frac{h}{G}$. If we consider the event of the left endpoint of a read falling in some determined interval a success, we have a binomial distribution for these left endpoints of reads; the mean of this binomial distribution is the number of events (namely the number of reads) times the probability of success:

\[ N \frac{h}{G} \]

If $N$ is large and $h$ is small, the binomial distribution described above is well approximated by a Poisson distribution with $\lambda = \frac{Nh}{G}$. This is related to the concept of the law of rare events.
13.2 Fundamental Question 1

Recall the first of our 3 fundamental questions regarding genome assembly: what is the mean proportion of the target genome region covered by contigs? Here is an outline leading us to the answer for this question:

- Recall the Poisson distribution, and consider the following formula:
  \[ P(Y = k) = \frac{e^{-c} c^k}{k!} \]

  where \( \lambda = c = \frac{NL}{G} \) refers to our coverage variable described earlier. Here, \( Y \) is a random variable equal to the number of reads whose left endpoint is located inside an interval of length \( L \). Because of the stochastic approach described before, the random variable \( Y \) is Poisson distributed with mean \( c \).

- Note the probability of no left endpoints being in an interval of length \( L \):
  \[ P(Y = 0) = e^{-c} \]

- Next, consider the probability that there is at least one left endpoint of a read in the interval of length \( L \):
  \[ 1 - P(Y = 0) = 1 - e^{-c} \]

- The mean portion of the target genome covered by a contig (i.e. covered by one or more reads) is the probability that a point chosen at random from the target genome is covered by at least one read/fragment that has its left endpoint in the interval of length \( L \) immediately to the left of this point (recall how all our reads are of length \( L \)).

- This probability, which represents the mean portion of the target genome covered by a read, and which is the answer to our first fundamental question is: \( 1 - e^{-c} \).

13.3 Fundamental Question 2

The second of our 3 fundamental questions is: what is the mean number of contigs? Here is an outline leading us to the answer for this question:

- Consider a binomial distribution, with \( n \) events, each with a probability of success \( p \), and the mean number of successes being \( np \).

- Each contig has a unique rightmost read/fragment.

- The probability that a fragment is the right most member of a contig (its “signature”) is that of there being only right-hand points of a read in an interval of length \( L \), or, in other words, the probability that there are no left endpoints in an interval of length \( L \), which we had from finding our answer to question 1 as:
  \[ P(Y = 0) = e^{-c} \]

- The mean number of contigs is the number of reads/fragments, \( N \), multiplied by the probability that a fragment is the right most member of a contig (its “signature”).

- This gives us that the mean number of fragments is equal to:
  \[ N \times P(\text{probability of a read being the right most in a contig}) = Ne^{-c} \]