Khan Inan
Professor Truong
BI-GY 7683

Experiment week 4

Rationale:
My experiment for week 4 was that I wanted to determine the genome size of the blue whale. Implementing my new knowledge of sequence alignment, genome assembly and phylogenetics, I can explore this topic much more in-depth than before. These new methods will allow me to analyze the gathering of sequence reads that have been obtained from tissue and DNA collection.

Hypothesis:
My previous hypothesis was that an organism's size is not an accurate predictor for its genome size. My reasoning for this is that there are many exceptions to this notion and examples of organisms where the genome size is distinctly much larger than one would imagine, like amoebas. Having said this however, genome size is definitely not random and there are other factors that contribute to how large the genome size of an organism is, beyond just the length or mass of the organism.

Experimental implementations:
Sequence alignment: alignment would be very useful in this case simply because this process allows us to locate instances of substitution, deletion and insertion of nucleotides within the genome of an organism. All three of these examples of mutations/transformations all contribute to the length of a sequence, which consequently determines genome size.To get into it more specifically, the best reference genome for the blue whale would be the fin whale, because they are closely related and still living today. Obtaining reference reads from a fin whale would allow for a comparison of sequences, and more importantly help us determine whether or not the genome of a blue whale is smaller or larger than a fin whale.

Genome assembly: The processes of tissue analysis, DNA collection and high throughput sequencing would not give us the complete genome of a blue whale but instead the “building blocks” from which we can assemble a map. One would think that determining genome size from having a large collection of sequence reads would be as simple as lining up the sequences regardless of order and measuring the end to end length of the series of nucleotides. This however is not the case because order is important in determining sections of overlap within the assembly, and it is safe to say that ignoring overlaps would make the genome seem larger than it actually is, due to recurring nucleotides and bases.

Phylogenetics: This last method gives a visual representation of where the blue whale falls in the larger biological tree. Phylogeny in the case of the blue whale is interesting because going further along the branches of the family tree of blue whales shows that they are fairly closely related to other aquatic mammals that one wouldn’t expect like hippopotamuses. In the case of

my experiment, phylogenetics mainly serves a reference tool for determining how the blue whale differs in its genetic makeup and genome size compared to other closely related (or even unrelated) species.

Challenges and solutions:
I would say that the biggest challenge in this case would be dealing with an incomplete collection of sequence reads. This is because it would give us a false understanding of the genome size of a blue whale. Although methods like De bruijn graph genome assembly would allow us to work with the fragments and sequence reads that we already have, there would be additional steps to determine what's missing from the complete picture. Another challenge would be having small sized contigs. This would be a problem, since if our DNA and sequence reads collection methods gave us contigs that are too small it would negatively affect our confidence and trust in the accuracy of our genome assembly.

Experiment week 5

Rationale:
My experiment from week 5 was that I wanted to observe the genetic similarities between the blue whale and other species of baleen whales. I saw through articles on the topic of comparative genomics that blue whales are extremely closely related to other aquatic mammal species to the point of even being able to breed with them successfully, and I wanted to delve in to the topic further to explore why and how this is the case

Hypothesis:
The blue whale shares a lot of genetic similarities with other baleen whales and aquatic mammals because of their close biological relation on the phylogenetic tree. If the evolution of the blue whale and their split into their own independent species occurred farther back in the evolutionary timeline then they would be more genetically different. However, since the phylogenetic evolutionary split of the blue whale was historically fairly recent, (at least compared to other species) they are more closely related to other organisms in the same family

Experimental implementation:
sequence alignment: alignment in this case would be useful for determining just how similar our collected blue whale sequence reads are with another reference genome. To quantify this similarity/difference and put it into a scientific value, we can use genomic distances. The closer our calculated genetic distance value is to 0, the closer our blue whale is related to this reference organism (at least on a genetic level.) Conversely, if The value of θ is closer to 0.5, then their genes are very far apart on the phylogenetic tree.

genome assembly: assembly in this case would be useful for converting our fragments of data and sequence reads into something more usable. sequence read lengths are shorter than genomes and genes and if we want to efficiently and correctly do sequence alignment analysis, then we want to construct a longer DNA sequence. We can do genome assembly by a couple

different methods like de bruijn graph and overlapping but I think the de bruijn graph method is better because it handles redundancy better in a large collection of HTS data.

Phylogeny:Lastly, phylogeny as I explained earlier will shed light into how long ago the blue whale split from another species like the hippopotamus, or the manatee or the fin whale. If there is extreme similarity between the blue whale and another species then they we also be fairly closely linked on the branches of the phylogenetic tree

Challenges and solutions:
The one major challenge is my experiment for determining genetic similarity of the blue whale with other species is that there will be a lot of gene sequence mutations/transformations. If there are many substitutions/deletions/insertions then it may give us a false read on the similarity or difference between our sequence reads and the reference reads. Unfortunately, there isn't an easy solution to this problem and the best thing to do would be to simply be very careful in our calculations on genomic distance and not jump to any early conclusions