Khan Inan
Excercise 10

M10_v1_description - Genome sequencing workflow

M10_v1_1) What is the basic workflow of the genome sequencing project?

Basically it involved obtaining a DNA clone, then sequencing the DNA clone and then and then assembling the resulting data from the numerous clones based on overlap to obtain and continuous sequence

M10_v1_2) What are the two main strategies of sequencing genomes?

The two main methods are hierarchical shotgun sequencing and whole genome shotgun sequencing

M10_v3_1) Why does the Map-based sequencing take more time and money?

This is because any map-based sequencing method requires sorting through extremely repetitive sequences that may or may not be accurate

M10_v3_2) What is RFLP and how do we use it to determine which segment overlaps one another?

RFLP stands for restriction fragment length polymorphism and it essentially plays the same role as radioactive markers in some other sequencing methods. RFLP are genetic markers and they show what DNA is passed down through families generation to generation

M10_v4_1) Why is it more computationally challenging to assembly data from shotgun sequencing compared to map-based sequencing

Shotgun sequencing uses computer processing power to align fragments correctly and search for overlaps. This requires complex software as well as computing power

M10_v4_2) What genome feature makes it challenging for shotgun sequencing method.

The fact that genomes contain many repetitive sequences makes is difficult for even the most meticulous softwares to determine where certain fragments belong on the genome

M10_v4_3) How do pair-end sequences help with this?

Paired end sequences help alleviate the issue of repetitive sequences by sequencing from both ends of the fragments in order to create data that is alignable and error correctable

M10_v6_description - Assembly assesment

M10_v6_1) In the equation c = LN/G , what do the different variables stand for?

L is the length, N is the # of reads, G is the genome length and C is the calculated coverage value

M10_v6_2) If our genome is 3 billion basepairs and we have sequenced 15 billion basepairs, how much of the genome will we have sequenced?

We would have sequenced the entire genome because 15 billion basepairs gives us a lot of repetition for problematic sequences and essentially reduces our margin of error by a lot for uncertain sections of the genome

M10_v6_3) Given the definition of N50, what is the definition of N20?

Since N50 essentially refers to the shortest possible contig length for which longer length contigs cover 50% of the genome assembly. N20 would be that value is a bit longer where any longer only covers 20% of the assembly

M10_v7_description - deBuijin Graphs
M10_v7_1) How do deBruijin graphs save memory requirements?

From what I understand de bruijn graphs dont store the actual sequences or fragments but instead hash table representations of the sequences along with vector data in order to save memory

M10_v7_2) What evidence can you use to collapse components in a graph?

The main evidence is the if you feel that any selections on the graph show signs of benign repetitive or redundant. Collapsing those components would reduce visual clutter as well as make things more concise

M10_V7_3) In transcriptomics, give one biological reason why certain graphs will not be collapsed.

This is most likely because it is difficult to eliminate variables or aspects of the graph especially with data from so many sources involved. Data can come from maps or reference genomes or databases so it is difficult to collapse something that is pulling from so many different sources.