Charlee Cobb - Transcriptomics, Exercise 10

Charlee Cobb

2023-04-10

M10_v1_description - Genome sequencing workflow

Genome sequencing is the process of extracting DNA and decoding the order of nucleotides in the given DNA to find genes and define nucleotide pattern.

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

In the general workflow fo the genome sequencing project, DNA is extracted from cells from an organism of interest and the DNA is then broken into fragments of various sizes. The fragments are then passed through Agarose gel electrophoresis and DNA is purified from the gel. Then a clone library is prepared and end sequences are obtained from DNA inserts. Finally, the short segments of DNA are decoded and the overlaps are aligned to assemble the sequence into contigs.

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

Whole genome shotgun sequencing and map based sequencing.

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

Map based sequencing takes more time and money because it requires making multiple copies of a segment of a genome, fragmenting those copies, aligning the fragments to each other, then connecting the overlaping fragments to other segments in the genome. These steps increase the cost of material and labor.

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

RFLP is a probe that hybridizes with DNA fragments after the DNA was seperated in a gel. RFLPs are used to measure the distance between DNA fragments based on the amount of shared nucleotides in the fragment.

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

It's more challenging because with shotgun sequencing you are trying to align fragments of the entire genome vs fragments of segments of a genome. Because there are a lot of repetitive regions in the genome, and to distingush between these regions you needed powerful algorithms that could be run on high speed computers.

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

Repetitive regions in the genome.

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

Paired-end sequences help orient the location of repatitve regions by adding more information to better map a sequence back to the genome.

M10_v6_description - Assembly assesment

Assembly assesment is the process of verifying and controlling the assembled DNA fragments that will produce a genome readout.

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

C stands for fold sequence coverage, L stands for length of reads, N stands for the number of reads, and G stands for genome size or length of the genome.

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?

around 99.4 percent of the genome

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

N20 would be where the size of the smallest contig of a set whose size makes up 20% of the genome.

M10_v7_description - deBuijin Graphs

A deBuijin Graph is a directed graph that can show overlaps between sequences of symbols.

M10_v7_1) How do deBruijin graphs save memory requirements?

The graphs save on memory requirements through indexing.

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

You can collapse non-branching paths in the graph into single edges

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

Because the gaps and non-branching paths may represent alternitave splicing. So the locus is the same but they don't ranch out in the given isoform.