Charlee Cobb - Transcriptomics, Exercise 7

Charlee Cobb

2023-03-20

What is our ultimate goal of the RNA-seq workflow? The goal of the RNA-seq workflow is to quantify how many reads are coming from a specific gene.This will lead us to understanding which genes are differentially expressed between conditions.

Why is it a good idea to perform quality control of our raw sequences? Quality control is needed because there are many known errors from sequencers and other technology that get introduced into the reads.

Why is it preferred to map to the reference genome instead of the predicted mRNA sequences? Using a reference genome is preferred in alignment because we don't have to depend on the quality of annotation in the predicted mRNA sequences, and we are not assuming that we know everything about the mRNA molecules in the experiment.

What is the difference between global and local alignment? The global approach to alignment tries to match the entire sequence from end to end. The local approach. Local alignment will find sequences of high similarity by focusing on regions in the sequence.

Why do we need heuristic methods to align NGS sequences? Heuristic methods are needed to align NGS sequences because we want to optimize speed and memory complexity when aligning reads.

How does BLAST manage to find an alignment so quickly? BLAST uses the Burrows-Wheeler transformation to run a heuristic search. BLAST scores it's findings and returns the highest scored match.

What is the difference between blastn and tblastx? blastn conducts a search of nucleotides from the input sequence. tbasltx translates the input sequence into proteins and searches for a matching protein sequence.

Why is BLAT a better choice (compared to BLAST) to look for alignment of short RNA sequences? How is the BLAT database created? BLAT takes advantage of a hash index and looks for a match of higher than 95% throughout the entire genome. BLAT therefore helps us find the location of the sequence better than BLAST.

List 3 different challenges with aligning short sequences 1)They can map to multiple regions in the genome 2)Errors can make a larger impact on alignment of shorter reads 3)Shorter sequence decreases the number of high quality nucleotide scores

How do you create a suffix array? A suffix array is made by taking the sequence and removing the first letter or nucleotide everytime you add the sequence into an index in the array.

What is a challenge when using suffix arrays to find sequence patterns? A challenge in using suffix arrays is that multiple copies of the genome will still take up space in the array, so you'll have to transform the array to optimize for space.

What are the steps to create a Burrow Wheeler's Transformation? You first align reads to the reference genome, but the reference genome has to be transformed in an indexed array.

What are the steps in finding a match using BWT? Once you have the table you introduce a special character to the end. Then you shift the input sequence to match the reference. If there is a miss match, it simply gets added to the end of the sequence. The special character tells us where to start again.

How many lines represent one sequence? And what is provided in each line? There is one line that represents a sequence, and each line has a read identifier and a read sequence quality.

What kind of information is provided in a SAM/BAM file? A SAM/BAM file cotains the alignment results of the alignment process. BAM is the binary file of SAM

How do the sequence aligners use GFF/GTF files? The gff and gtf files are used as arguments so aligners know where the genes are and where exons are spliced