Khan Inan
BI-GY 7633 transcriptomics
Exercise 7

What is our ultimate goal of the RNA-seq workflow?

The ultimate goal of RNA-SEQ should be to detect parts of sequence that are known and to also make new discoveries in terms of fusions, variants or mutations

Why is it a good idea to perform quality control of our raw sequences?

It is important to perform quality control because the sample from which the sequence was obtain may not be of the best quality or there may be an issue with flagging/parameter issues with the collected data

Why is it preferred to map to the reference genome instead of the predicted mRNA sequences?

The reference genome and the predicted mRNA are both estimates of what the genome will look like but using the reference genome is preferred because it is created from layering a large collection of sequencing information and thus is the best possible estimate

What is the difference between global and local alignment?

Global is essentially end to end of an entire strand while local is alignment of a small section or substring

Why do we need heuristic methods to align NGS sequences?

Using heuristic methods instead of something that is more exact is mainly done to save time and make the process a bit more efficient

How does BLAST manage to find an alignment so quickly?

BLAST pre-loads and pre-calculates information from the database to essentially have a record of each possible word/substring in both of the sequences. This makes alignment faster since the strings don't have to be combed and analyzed first

What is the difference between blastn and tblastx?

They differ in terms of what input they require and where they search. Blastx needs a nucleotide sequence and searches within a protein data while tblastn is the exact opposite

Why is BLAT a better choice (compared to BLAST) to look for alignment of short RNA sequences?

BLAT has their methods of reversing the splicing process that creates the short RNA sequences. The tool does this by determining which alignments use which of the bases in the mRNA one time, and then determines the splice sites bases on that.

How is the BLAT database created?

From what I understand the BLAT database is created by merging separate databases for each genome that the user is trying to search or index

List 3 different challenges with aligning short sequences

1. **Really short sequences can be in multiple positions of the same genome and repeating patterns**
2. **It is difficult to determine where the gaps are in the sequence and what parts are missing in the sequencing process**
3. **Errors and mutations are difficult to pinpoint and identify**

How do you create a suffix array?

You would first need a suffix tree, then you perform a depth first travel, which is essentially an algorithm that searches for all the possible combinations or connection/vetices within a graph or tree structure.

What is a challenge when using suffix arrays to find sequence patterns?

The biggest challenge is the scope/size of a suffix array. It can be extremely large and complex because there is a data point for each character within a sequence

What are the steps to create a Burrow Wheeler's Transformation?

1. **Perform cyclic rotations**
2. **Sort the rotations alphabetically**
3. **The final obtained column is the output**

What are the steps in finding a match using BWT?

1. **Find the range of a character within the first column**
2. **Obtain a range that matches in the last column**
3. **Find the pattern character and then find its range in the first row, followed by findings its subrange**

How many lines represent one sequence? And what is provided in each line?

There should be 16000 lines and each line should have 4000 reads

What kind of information is provided in a SAM/BAM file?

You can get alignment information for sequences of interest and the data that is gathered from mapping them to reference sequences

How do the sequence aligners use GFF/GTF files?

They are essentially a file where the text is separated via tabs, and they can be used to describe genomic features that are useful to tools that can align sequences