J. Jedediah Smith

Week 9 Paper Questions 2

BIFX-504

10/21/2021

Continue reading “Introduction to differential gene expression analysis using RNA-seq” and answer the following questions:

1. One of the important goals of performing quality control checks on raw sequencing data is to remove uninformative sequences. What types of uninformative sequences may be present in the data?

Types of uninformative sequences include: PCR duplicates, adapter contamination, rRNA and tRNA reads, and unmappable reads.

1. Why is the base content in the first 10-15 bases of RNAseq reads not uniformly distributed?

Due to the “random” hexamer priming step, which is evidently not as random as one might have hoped.

1. Describe the “seed and extend” approach for short read alignment

A subset of the read is chosen to be a “seed” that the tool finds the best possible match for in an index made up of the reference genome. Each matched seed is extended on both sides under certain constraints until as much of the read is aligned as possible.

1. Why is mapping of RNA-seq reads more challenging than genome sequencing (of DNA)?

Due to the spliced alignment of exon-exon-spanning reds and the presence of multiple different isoforms of the same genes.

1. Describe the approach used by pseudoaligners such as Salmon and Kallisto use to count the number of shared k-mers between sequences and transcript annotations.

The sequences for comparison are sliced up into groups of unique k-mers of a given length k. Then, for each pairwise comparison, the number of mutual occurrences of a specific k-mer are tallied. To assess similarity, some sort of distance function is employed. Two identical sequences should have a distance of zero.

1. What are the benefits and drawbacks of alignment-free methods compared to alignment-based methods of read mapping?

Alignment-free is much faster, but they are very reliant on a precise and comprehensive transcript annotation. When a sequenced fragment originates from a locus not part of the pre-defined DNA annotation, it may be incorrectly attributed to a certain region. Alignment-based tools have scoring systems to guard against such spurious alignments. But alignment-free tools typically lack these metrics and just go with the best match.

1. Why do RefSeq and Ensembl differ from each other? Why is this important?

RefSeq and Ensembl are based on different annotation pipelines that are both trying to determine the details about untranslated regions, introns, and exons. The use of different pipelines is a way to validate research. If both pipelines provide similar evidence that support the same conclusion, this adds more weight to it.