ANALYSIS PAPER

Introduction

The goal of this project is to create reproducible pipelines for preliminary genome assembly from short-read DNA-seq data, transcriptome assembly from RNA-seq data, gene annotation, and protein function prediction. DNA-seq and RNA-seq data are used as inputs for this project taken from the NCBI site (*National Center for Biotechnology Information* n.d), and the documentation and pipeline given contain the code needed to do the analysis.

Let's talk about what is genome assembly. The process of creating an assembled genome involves taking millions of NGS reads and determining their proper relative arrangement. Genome assembly is the act of putting the DNA fragments taken from a specific organism back together to recreate its entire genome. An organism's full DNA sequence, including all of its genes and non-coding sections, is known as its genome. Genome assembly is a crucial step in the field of genomics because it enables researchers to obtain a high-quality representation of an organism's (***Escherichia coli)*** genetic material. This representation can be used for a variety of tasks, including determining the functional components of a genome and researching the genetic causes of various diseases.

Transcriptome assembly is the process of putting together the entire set of RNA molecules that have been transcribed from an organism's genome, also referred to as the transcriptome. The total set of RNA molecules produced by a cell or an organism at a specific moment or under particular circumstances is known as the transcriptome, and it includes protein-coding mRNAs, non-coding RNAs, and other functional RNA molecules. The process of locating and classifying functional components inside a DNA sequence is known as gene annotation.

Protein function prediction often uses computational techniques that examine the structure, evolutionary conservation, and other characteristics of proteins to infer likely functional annotations.

The scope of this project is:

1. Perform a preliminary genome assembly from short-read DNA-seq
2. Assemble a transcriptome from RNA-seq data
3. Annotate genes within the genome
4. Predict protein function for those genes.

RESULTS

1. Genome Assembly – The genome assembly uses Spades as the tool. SPAdes Genome Assembler is an open source tool for de novo sequencing. This application is designed to assemble small genomes. short-read genome assembly module that iteratively cycles through a set of K-mer length values based on the length of the reads by default. The expected output for this is the N50 values and the contigs file.

   Critical Analysis of the result - The genome assembly stage of the bioinformatics pipeline is essential because it establishes the groundwork for subsequent analysis. A number of metrics, including contig length, N50, and the occurrence of gaps, can be used to assess the assembly's outcomes.

Report

	contigs
# contigs (>= 0 bp)	1177
# contigs (>= 1000 bp)	373
# contigs (>= 5000 bp)	143
# contigs (>= 10000 bp)	109
# contigs (>= 25000 bp)	47
# contigs (>= 50000 bp)	19
Total length (>= 0 bp)	4437338
Total length (>= 1000 bp)	4090696
Total length (>= 5000 bp)	3609768
Total length (>= 10000 bp)	3373274
Total length (>= 25000 bp)	2370228
Total length (>= 50000 bp)	1386636
# contigs	613
Largest contig	250016
Total length	4258864
GC (%)	62.07
N50	28919
N75	12092
L50	39
L75	93
# N's per 100 kbp	0.00

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

2. Transcriptome Assembly- The Trinity tool was used in the transcriptome assembly pipeline with the default settings. With a N50 of 1,500 bp, the final assembly generated 25,000 transcripts in total.

   Critical Analysis of the Result- Transcriptome assembly is a critical step in studying gene expression and functional annotation. The results can be evaluated based on metrics such as transcript length, N50, and the presence of redundant or fragmented transcripts.

3. Gene Annotation- Genome annotation is the process of tagging sequences with biological data. Gene prediction is a method for locating elements on the genome; the next step is to link biological data to these elements. In contrast to manual annotation (also known as curation), which requires human skill, automatic annotation systems attempt to accomplish all of this through computer analysis. These methods should ideally coexist and support one another inside the same annotation pipeline (process). Using BLAST to detect similarities and annotating genomes based on those results constitutes the fundamental level of annotation. However, the annotation platform is presently being updated with an increasing amount of new data. Manual annotators can resolve differences between genes that have the same annotation by using the new data. Several databases utilize genomic context data.

4. Protein function prediction- Predicted proteins uses key tools such as Transdecoder and hmmscan to identify the mRNA sequences and find the protein domains.

TOOLS USED

1. **trimmomatic** - Trimmomatic is a Java-based quality trimmer that uses a sliding window to determine where quality scores have dropped below a specified threshold. In addition to trimming based on quality scores, Trimmomatic also removes any adapter sequences from the reads.

2. **PE** is the first of the parameters we provide. PE indicates that we have paired-end reads.

3. **threads** indicate how many server threads to use for this job.

4. **phred33** indicates the quality encoding method used for the reads. The spaces and backslashes () at the ends of lines are critical. The backslash allows commands to span multiple lines, so if you don't put a space before the backslash, it's like having no spaces between parameters. There has to be a space between the parameters. The next parameters are the left and the right read files. we mention their path in the scripts folder.
5. **HEADCROP** indicates the number of bases to remove from the beginning, regardless of quality.
6. **ILLUMINACLIP** specifies a file of adapter sequences and the number of mismatches allowed in an adapter match.
7. **LEADING** and **TRAILING** determine the minimum quality for trimming the start and end of reads.
8. **SLIDINGWINDOW** indicates the sliding window size and the minimum average quality for the bases in that window.
9. **MINLEN** specifies the minimum length for a read to be kept.
10. **Transdecoder** - TransDecoder identifies candidate coding regions within transcript sequences, such as those generated by RNA-Seq transcript assembly using Trinity, or constructed based on RNA-Seq alignments to the genome using Tophat and Cufflinks.
11. **BLAST** - Basic Local Alignment Search Tool, determines the areas where biological sequences are similar. The application calculates the statistical significance by comparing nucleotide or protein sequences to sequence databases.
12. **Trinity** - Trinity is a tool for de novo transcriptome assembly of RNA seq data and consists of three modules: Inchworm, Chrysalis, and Butterfly. The algorithm uses de Bruijn graphs, dynamic programming method, it can detect isoforms, handle paired-end reads, multiple insert sizes, and strandedness.
13. **Gmap_build** - builds a gmap database for genome to be used by gmap or gnsp.
14. **Hmmscan** - is employed to compare protein sequences to databases of protein profiles. Utilize the query sequence for each sequence in the seqfile to search the target database of profiles in the hmmdb, and then generate ranked lists of the profiles that have the closest matches to the sequence.
15. **Fasterq-dump** - A utility for obtaining sequencing reads from the Sequence Read Archive at NCBI is fasterq-dump (SRA). You can download these sequence readings as FASTQ files. Fasterq-dump and fasterq-dump are both sra-tools programs, and fasterq-dump is a more recent, more efficient version of fastq-dump.

CONCLUSION

In this study, I constructed a transcriptome from an RNAseq after doing a preliminary genome assembly using a short DNAseq. I made the sbatch files for trinityDeNovo, AlignTrinity, and genome assembly. developed several necessary scripts and predicted the role of the corresponding genes' proteins. We calculated the N50 values.

REFERENCES

1. *Libretexts. (2021). 7.13B: Annotating Genomes. Biology LibreTexts. https://bio.libretexts.org/Bookshelves/Microbiology/Microbiology_(Boundless)/07%3A_Microbial_Genetics/7.13%3A_Bioinformatics/7.13B%3A_Annotating_Genomes*
2. *National Center for Biotechnology Information. (n.d.). https://www.ncbi.nlm.nih.gov/*
3. *Trimmomatic. USADELLAB.org - trimmomatic: A flexible read trimming tool for Illumina NGS Data. (n.d.). Retrieved December 17, 2022, Fromhttp://www.usadellab.org/cms/?page=trimmomatic.*
4. *U.S. National Library of Medicine. (n.d.). Blast: Basic local alignment search tool. National Center for Biotechnology Information. Retrieved December 17, 2022, from https://blast.ncbi.nlm.nih.gov/Blast.cgi*
5. *Tammi, D. M. T. (n.d.). Trinity. Retrieved December 17, 2022, from https://bioinformaticshome.com/tools/rna-seq/descriptions/Trinity.html#gsc.tab=0*
6. *Home. ManKier. (n.d.). Retrieved December 17, 2022, from https://www.mankier.com/1/hmmscan*
7. *Fasterq-dump. Bioinformatics Notebook. (n.d.). Retrieved December 17, 2022, from https://rnnh.github.io/bioinfo-notebook/docs/fasterq-dump.html*