A project on whole-genome assembly, annotation and phylogenetic analysis

Our project involved whole-genome reconstruction from short reads, genome annotation, extraction of meaningful and tractable information from the whole genome, and phylogenetic reconstruction.In this project, we used data generated from 2nd-generation sequencers (Illumina HiSeq).The samples we used for our project pertained to several isolates of *M. oryzae*. During the course of this mini-project we acquired skills in high throughput sequence analysis, data manipulation and phylogenetic analysis.

Introduction

Rice blast is a fungal disease caused by a plant pathogenic fungus called *Magnaporthe oryzae*.Rice blast is a major threat to global food security given the disease is responsible for approximately 30% of rice production losses globally—the equivalent of feeding 60 million people. The losses experienced as a result of rice blast has led to an increase of the global rice price and a reduction in consumer welfare and food security. With Rice being a staple crop for more than half of the world's population, management of rice blast disease would have far reaching beneficial effects on consumer livelihoods globally.(Lawton Nailey et al 2016).

Objectives

- To study the genetic diversity of Magnaporthe oryzae in our samples
- To compare our rice blast sequences with other rice blast accessions from Africa and Asia

Materials and Methods

A. Quality Control and trimming

The original data used can be found on /var/scratch/jb/Magnaporthe_project_data/ on ILRI cluster's compute05 host. It consisted of **5 samples**, each sample consisting paired-reads sequence data, in compressed FASTQ file format. These files were meant to contain reads of DNA pertaining to *Magnaporthe oryzae.* Upon obtaining our data from the hpc var scratch environment we proceeded to perform a FASTQC analysis on the samples. The FASTQC was performed using version 0.11.7 of FASTQC software on ILRI's HPC environment. The report generated from the FASTQC analysis was used to inform our next quality control steps (trimming). We went on to trim off the first 7 bases of each of our forward and reverse read in each sample. For this process we used trimmomatic. (trimmomatic/0.38). The Script for this process has been provided on our Scripts folder on our GitHub repository.

A FASTQC analysis was then performed on our trimmed data to ensure that our data was of good quality. Our analysis on the fastqc_trimmed_data confirmed that we had indeed trimmed off our data and that our dataset was clean and ready for the next process of assembly.

B. Genome Assembly

We indexed our short paired end reads using bwa. The script for this been provided on our GitHub repository under the scripts section.We then aligned our reads to a reference genome (Magnaporthe oryzae). The link for our reference genome can be found on the reference section of this write up.We went ahead to assemble our reads. Assembly was performed in a two part process: using velveth and velvetg at Kmer lengths 41, 49 and 55. We compared the N50 values of each of the kmer lengths. We finally settled on kmer length 49 (This decision was made after performing a trial run on one of the samples: more precisely sample 1.)

C. Genome annotation

After obtaining our contigs we annotated using the Geneious software(*Geneious Prime® 2019.2.1 Build 2019-06-17 11:03 Java Version 11.0.3+7 (64 bit*) . Our objective was to extract six genes as prescribed on the six_gene research paper (Ning Zhang et al 2011). Genes identified included : the largest subunit of RNA polymerase II gene, A DNA replication licensing factor gene (MCM7) ,18S rRNA gene(SSU), 28s rRNA (LSU), translation elongation factor I -alpha gene (TEF1) and internal transcribed spacer of the rRNA genes.

D. Retrieval of homologous sequences from the public databases

We blasted our contigs by drafting up a blast script and running it on the ILRI HPC. The script we ran was specified to the directory of each sample contig. We acknowledge the limitation of running a blast script this way. It does not encourage reproducibility or replicability. However due to time constraints we were not able to accomplish this task. The blasting script we used can be found on the scripts folder on our git repository.

Results and discussion

The fastqc report generated, clearly showed that the data was of good quality. The most striking issue was the per base sequence content of all the reads except one. Therefore,

we resolved to trim the first seven bases of each read. Upon performing a fastqc check on the trimmed data, we proved that indeed the bases were cut out.

We used BWA to index and map our reads to a reference genome (*Magnaporthe oryzae*). This generated a SAM file for each sample, with both the forward and reverse reads mapped on the reference genome. At the end of this step, we had a total of five SAM files.

We used velvet to assemble all our reads. We tried several kmers; 41, 49 and 55. We checked each assembly statistics and eventually settled of kmer 49. It had the best N50 value compared to the rest. Among the output were contigs.fa files. They comprised of nodes.

We used Geneious to annotate our contigs. The entire genome of *Magnaporthe oryzae* that we used as our reference was spread out across 7 separate chromosomes. We first assembled each sample's contigs onto each of the chromosomes. This generated consensus sequences. We then transferred the annotations from the references to each of the consensus sequences. Based on the genes highlighted on the literature paper provided, we had 6 genes to look out for. We only managed to find two of them. However, the regions on our consensus sequence expected to contain these genes had no coverage at all. We attributed this to the loss of most of the nodes during the assembly process. Frustrated by the outcome of events, we resorted to choosing other genes to use for the rest of the steps. This too was met by several roadblocks. For one, it was difficult to find a gene that was consistently represented in all the samples.

We tried blasting some genes, the hits mostly belonged to the earlier versions of our reference. Consequently, we could not construct phylogenetic trees.

Conclusion

Given more time, we could have repeated the assembly and see what we can find.