DE NOVO ASSEMBLY OF AN UNKNOWN GEMINIVIRUS

NURUL JANNAH BINTI MAT @ MOHAMAD

FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2014

DE NOVO ASSEMBLY OF AN UNKNOWN GEMINIVIRUS

NURUL JANNAH BINTI MAT @ MOHAMAD
(SGJ130002)

SUBMITTED TO THE
INSTITUTE OF BIOLOGICAL SCIENCES
FACULTY OF SCIENCE
UNIVERSITY OF MALAYA, IN PARTIAL
FULFILMENT OF THE REQUIREMENT FOR
THE DEGREE OF MASTER OF BIOINFORMATICS

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: NURUL JANNAH BT MAT @ MOHAMAD

(I.C/PASSPORT NO: 900902-11-5290)

Registration/Matric No: SGJ130002

Name of Degree: MASTER OF BIOINFORMATICS

De novo Assembly of an Unknown Geminivirus ("this Work"):

Field of Study: Bioinformatics, *De novo* Assembly, Geminivirus

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date

Subscribed and solemnly declared before,

Witness's Signature Date

Name:

Designation:

ABSTRACT

Next-generation sequencing (NGS) also known as high throughput sequencing is now fast and cheap enough to be considered part of the toolbox for investigating the unknown virus. Illumina Genome Analyzer is one of the developed next-generation sequencing platforms that produce a significant larger volume of sequence data. The short sequence reads generated from Illumina Genome Analyzer can be used to perform de novo assembly. Therefore, this study was conducted to perform de novo assembly of an unknown geminivirus using the sequence reads generated from Illumina Genome Analyzer. In this study, the de novo assembly was carried out using SOAPdenovo and it indicates that only one scaffold (C11095) that mapped into the geminivirus genomes. After the scaffold output was obtained, the gene was predicted using GeneMark.hmm. There were 5 open reading frames (ORFs) predicted as gene. The function of each predicted gene was annotated using three different annotation tools, InterPro, Gene Ontology (GO) and UniProt. For example, from the InterPro result, the gene 1 encodes the geminivirus AL3 coat protein, while the UniProt result shows that the gene 1 encodes the replication enhancement protein and the GO shows that the gene 1 was involved in the viral process (biological process). In this study, the predictive genes were compared with the geminivirus genomes using BRIG (BLAST Ring Image Generator). The BRIG image shows that the large sequence of the unknown geminivirus was missing between 1000 bp until 1300 bp. From the genes comparison result, it indicates the similarity between the unknown geminivirus and the geminivirus genomes where all the geminiviruses encode the coat protein and replication-associated protein. The differences between the unknown geminivirus and the geminivirus genomes were the unknown geminivirus encodes the replication enhancement protein (gene 1), the hypothetical protein (gene 3) and the glyoxylate carboligase (gene 5). The phylogenetic

result shows that the geminiviruses can be classified into the East Asia (China, Taiwan, and Japan) and the Southeast Asia (Malaysia, Indonesia, Philippines and Vietnam) viruses. The unknown geminivirus (candidate virus) was located in the Southeast Asia group. This phylogenetic tree indicates that the unknown geminivirus share common ancestor with Tobacco leaf curl Indonesia virus C1, V2, V1 genes for replication-associated protein, putative V2 protein, coat protein, partial and complete cds. The results of the phylogenetic tree suggest that the unknown geminivirus could be a Southeast Asia strain and it could be attack tobacco plants. The main point of this study was carried out to show the process in identifying an unknown sequence reads generated from Illumina Genome Analyzer.

ABSTRAK

Penjujukan generasi akan datang (NGS) juga dikenali sebagai pemprosesan tinggi penjujukan kini cepat dan cukup murah untuk dipertimbangkan sebahagian daripada alat untuk menyiasat virus yang tidak diketahui. Illumina Genome Analyzer adalah salah satu platform penjujukan generasi akan datang yang menghasilkan jumlah data jujukan yang lebih besar. Urutan pendek yang dihasilkan oleh Illumina Genome Analyzer boleh digunakan untuk melaksanakan perhimpunan de novo. Oleh itu, kajian ini dijalankan untuk melaksanakan perhimpunan de novo ke atas satu geminivirus yang tidak diketahui dengan menggunakan urutan pendek yang dihasilkan oleh Illumina Genome Analyzer. Dalam kajian ini, perhimpunan de novo telah dijalankan dengan menggunakan SOAPdenovo dan ia menunjukkan bahawa hanya satu scaffold (C11095) yang dipetakan ke dalam genom geminivirus. Selepas memperolehi scaffold, gen telah diramalkan dengan menggunakan GeneMark.hmm. Terdapat 5 bingkai bacaan terbuka (ORFs) yang diramalkan sebagai gen. Fungsi setiap gen ramalan telah diramalkan dengan menggunakan tiga alat yang berbeza iaitu InterPro, Gene Ontologi (GO) dan UniProt. Sebagai contoh, InterPro menunjukkan bahawa gen 1 mengekod kot protein geminivirus AL3, manakala UniProt menunjukkan bahawa gen 1 mengekod protein peningkatan replikasi dan GO menunjukkan bahawa gen 1 terlibat dalam proses virus (proses biologi). Dalam kajian ini, gen ramalan dibandingkan dengan genom geminivirus dengan menggunakan BRIG (BLAST Ring Image Generator). Imej BRIG menunjukkan bahawa urutan yang besar geminivirus yang tidak diketahui telah hilang antara 1000bp sehingga 1300bp. Dari hasil perbandingan antara gen, terdapat persamaan antara geminivirus yang tidak diketahui dengan genom geminivirus dimana semua geminivirus mengekod kot protein dan protein replikasi-berkaitan. Perbezaan antara geminivirus yang tidak diketahui dengan genom geminivirus adalah geminivirus

yang tidak diketahui mengekod protein peningkatan replikasi (gen 1), protein hipotesis (gen 3) dan glyoxylate carboligase (gen 5). Hasil filogenetik menunjukkan bahawa geminivirus boleh diklasifikasikan ke dalam Asia Timur (China, Taiwan, dan Jepun) dan Asia Tenggara (Malaysia, Indonesia, Filipina dan Vietnam) virus. Geminivirus yang tidak diketahui (calon virus) terletak dalam kumpulan Asia Tenggara. Pokok filogenetik menunjukkan bahawa geminivirus yang tidak diketahui berkongsi moyang yang sama dengan Tobacco leaf curl Indonesia virus C1, V2, V1 untuk replikasi-dikaitkan dengan protein, protein V2 yang diduga, kot protein, separa dan CDS lengkap. Keputusan pokok filogenetik mencadangkan bahawa geminivirus yang tidak diketahui boleh menjadi strain Asia Tenggara dan ia kemungkinan boleh menyerang pokok tembakau. Tujuan utama kajian ini dijalankan adalah untuk menunjukkan proses dalam mengenal pasti urutan yang tidak diketahui itu yang dihasilkan oleh Illumina Genome Analyzer.

ACKNOWLEDGEMENTS

I would like to thank all those people who made this thesis possible and an enjoyable experience for me. First of all I wish to express my sincere gratitude to my supervisor Dr. Rozaimi Razali that gives me an encouragement for this project.

I would also like to thanks to my co-supervisor Dr. Saharuddin Mohamad and the staff in the Department of Bioinformatics, Miss Devi and Mr. Ridzuan for their help and assistance in my project. Finally, I would like also to express my deepest gratitude for a constant support, emotional understanding and love that I received from my family. With the blessing of prayer from them, I can complete this thesis completely.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

AC1	DNA-A complementary sense 1
AL1	DNA-A left 1
AR1	DNA-A right 1
AV1	DNA-A virion sense 1
BC1	DNA-B complementary sense 1
BL1	DNA-B left 1
BLAST	Basic Local Alignment Search Tool
bp	Base pair
BR1	DNA-B right 1
BV1	DNA-B virion sense 1
C1	Complementary sense 1
DNA	Deoxyribonucleic acid
ICTV	International Committee on Taxonomy of Viruses
kb	Kilobyte
LIR	Large intergenic region
NCBI	National Center for Biotechnology Information
ND	No date
SIR	Small intergenic region
V1	Virion sense 1

CHAPTER 1

INTRODUCTION

1.1 Study background

Nowadays, there are lot of an unknown virus that cause disease in human, animal and also plant. Determining whole virus genome diversity is necessary for understanding the origin and the features of an unknown virus. Next-generation sequencing (NGS) also known as high throughput sequencing is now fast and cheap enough to be considered part of the toolbox for investigating virus. It has developed into a powerful tool that can be used to detect, identify and quantify novel viruses in one step (Dunowska *et al*., 2012). Besides that, this NGS can identify viral sequences without the background information on viruses.

Illumina Genome Analyzer is one of the developed next-generation sequencing platforms that produce a significant larger volume of sequence data than traditional Sanger sequencing (Illumina, ND). Many researcher uses the sequencing reads generated from Illumina Genome Analyzer to perform *de novo* assembly. *De novo* assembly is the process of merging overlapping sequence reads into contiguous sequences (contigs) without the use of any reference genome as a guide (Edwards and Holt, 2013). For example, Peng *et al*. (2014) performed *de novo* assembly of *Conyza canadensis* by combining data from multiple sequencing platforms (454 GS-FLX, Illumina HiSeq 2000 and PacBio RS). The present study will focus on an unknown Geminivirus genome.

Geminiviruses are a large family of plant viruses with circular, single-stranded DNA genomes that replicate through double-stranded intermediates. Usually these viruses are transmitted by insect. Padidam *et al*. (1995) stated that the viruses are

transmitted by whiteflies or leafhoppers which can cause significant diseases in many crop plants. There are a few types of Geminivirus. The Geminiviridae family that differ with respect to insect vector, host range and genome structure can be divided into four genera: Mastrevirus, Curtovirus, Topocuvirus and Begomovirus (Fauquet *et al*., 2008). This virus can attack many crops such as in pepper (Renteria-Canett *et al*., 2011), in tomato (Ghanim and Czosnek, 2000) and in tobacco (Paximadis and Rey, 2001).

At present, there is not much genetic information regarding an unknown geminivirus. To better understand the biology of the virus, we propose to perform de novo assembly of an unknown geminivirus using sequencing reads generated from Illumina Genome Analyzer.

1.2 Significance of study

To show the process in identifying an unknown sequence reads.

1.3 Objective of study

The major aims of this study were:

i) To perform de novo assembly of an unknown Geminivirus.

ii) To predict the number of genes in an unknown Geminivirus genome.

iii) To annotate the genes of an unknown Geminivirus.

iv) To compare the genes of an unknown Geminivirus with other viruses.

v) To perform phylogenetic tree for an unknown Geminivirus.

CHAPTER 2

LITERATURE REVIEW

2.1 The Geminiviridae

The family Geminiviridae is one of the largest and most important families of plant viruses. The geminivirus name comes from the unusual twin icosahedral (geminate) capsid structure of its members (Bisaro, 1996). The previous study has shown that each paired particle encapsidates a single molecule of covalently closed, circular, single-stranded DNA (ssDNA) (Bisaro, 1996; Zhou *et al*., 2001). The virion morphology is unique in the known viral world where two incomplete T = 1 icosahedra are joined together to form twinned particles (Krupovic *et al*., 2009). Their replication is using the rolling circle mechanism via double-stranded DNA (dsDNA) replicative form and using the cell replication system (Bisaro, 1996; Lapierre and Signoret, 2004). They have overlapping genes in different frames to efficiently code the proteins needed for replication, gene expression, encapsidation and movement (Saripalli, 2008).

2.2 Genomes of Geminivirus

Geminivirus have gained attention from the researchers around the world as the subjects of intensive research. This is due to their significance to plant pathology, plant molecular biology and plant biotechnology. Furthermore, their genomes have potential as vectors for the expression of foreign genes in plants, although this potential has yet to be fully exploited (Bisaro, 1996). The small, single-stranded DNA genomes of geminivirus encode four to eight proteins that are expressed from both strands of the double-stranded DNA replicative intermediate (Fondong, 2013). The genome of

geminivirus also encodes proteins that redirect host machineries and processes to establish a productive infection (Hanley-Bowdoin *et al*., 2013).

2.3 Classification of Geminivirus

There were four different genera in the family of Geminiviridae. In the previous study, the genus of Geminivirus was classified based on the taxonomy of insect vector, host range (monocot or dicot) and genome organization or arrangement. They can be divided into *Mastrevirus, Curtovirus, Topocuvirus* and *Begomovirus* (Padidam *et al.*, 1995; Fauquet *et al*., 2008). The geminivirus genome organization is composed of one or two components. The genome size of geminivirus ranges from 2.5 to 3.0 kb (Bisaro, 1996; Zhou *et al.,* 200; Hurst, 2011).

2.3.1 Mastreviruses

Mastreviruses have only one genomic component also known as monopartite genome and they are transmitted by leafhoppers. They have 2.6-2.8kb genome components (Cann, 2001). This virus mostly infects monocotyledonous plants such as maize (Lapierre and Signoret, 2004). Brown *et al*. (2012) quoted by Gaur *et al*. (2013) reported that this virus is the second largest genus of this family where it consists of 14 species according to a recent ICTV publication .Example species for mastreviruses are African streak virus group such as maize streak virus (MSV) which isolated from Africa and Indian Ocean islands such as Mauritius, La Reunion and Madagascar (Hughes *et al*., 1992).

The mastrevirus genome contains two intergenic regions, one large (LIR) and another small (SIR) (Bolok Yazdi *et al*., 2008). Besides that, the genome of mastreviruses consists of a single component encoding four proteins. The genome of these viruses encodes two proteins on the viral-sense (V-sense) strand that is the movement protein (MP) and the capsid protein (CP) that encapsidates, while it encodes RepA protein (exclusive of this genus) and the Rep protein on the complementary sense (C-sense) strand (Hurst, 2011). The viral genome is encapsidated by CP and the movement of virus is mediated by MP and also CP (Liu *et al*., 1999; Boulton, 2002; Hefferon and Dugdale, 2003). Meanwhile, Rep is required for viral DNA replication and RepA is used to interact with plant retinoblastoma-related proteins to regulate the host cell cycle (Boulton, 2002; Hefferon and Dugdale, 2003).

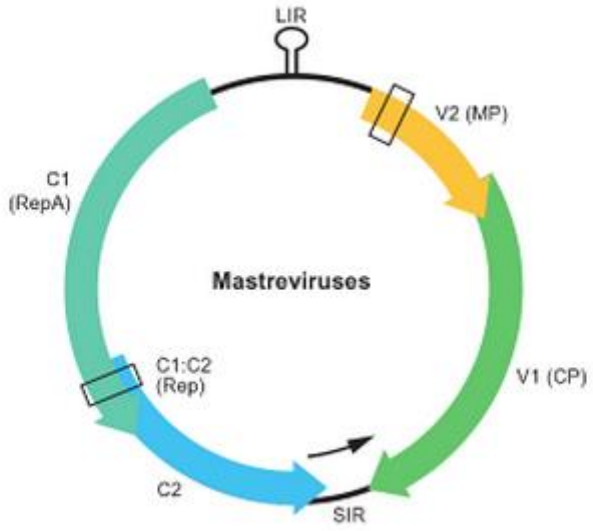

Figure 2.1: Genomic organization of mastreviruses (Hull, 2013).

2.3.2 Curtoviruses

Meanwhile, the curtoviruses also have a monopartite genome (one genomic component) of ~2.9-3.0kb (Cann, 2001; Chen *et al.,* 2011). They are transmitted by leafhoppers but these viruses prefer to infect dicotyledonous plants (Cann, 2001; Lapierre and Signoret, 2004). This virus has seven known species (Bolok Yazdi *et al*., 2008) such as spinach curly top virus from Southwest Texas (Baliji *et al*., 2004).

Their genomes encode up to seven proteins. Three encoded on the virion-sense strand are the coat protein gene (cp, V1) that encapsidates the virion-sense ssDNA and is involved in virus movement and insect vector transmission , a regulatory protein gene (reg, V2) that involved in the regulation of the relative levels of ssDNA and dsDNA, and a movement protein gene (mp, V3) for the virus movement. The complementary sense strand encodes a replication-associated protein gene (rep, C1) that required for the initiation of viral DNA replication, a gene expressing a protein that has silencing suppressor functions and acts as a pathogenicity factor in some hosts (ss, C2), a replication enhancer gene (ren, C3) and C4 protein is an important symptom determinant that is implicated in cell-cycle control (Cann, 2001; Varsani *et al.*, 2014).

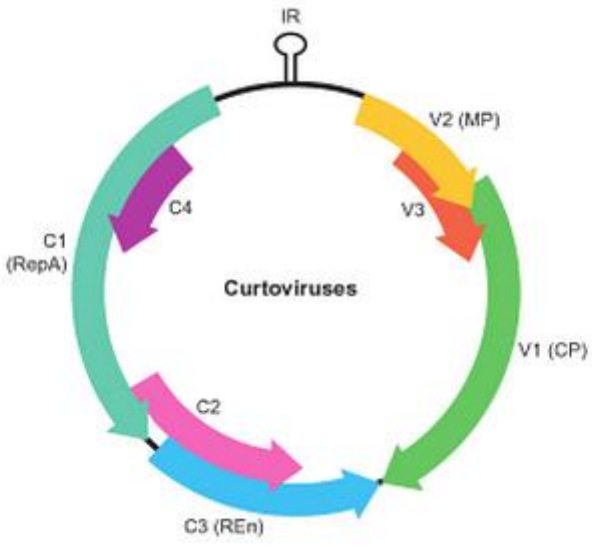

Figure 2.2: Genomic organization of curtoviruses (Hull, 2013).

2.3.3 Topocuviruses

The topocuviruses also have one genomic component which they infect dicots but they are transmitted by treehoppers (Lapierre and Signoret, 2004). The monopartite component is 2.8kb in size (Cann, 2001). The topocuviruses has only one species that is tomato pseudo-curly top virus (Saripalli, 2008; Hurst, 2011). The genome of this virus encodes for six proteins. Two proteins are encoded on the virion-sense strand, V1 is encoding for coat protein and V2 is encoding for movement protein. The genome encodes four proteins on the complementary sense strand that known as C1 encodes for replication-associated protein, C2 encodes for transcription activator protein (TrAP), C3 encodes for replication enhancer protein and C4 is implicated in cell cycle control (Hull, 2013).

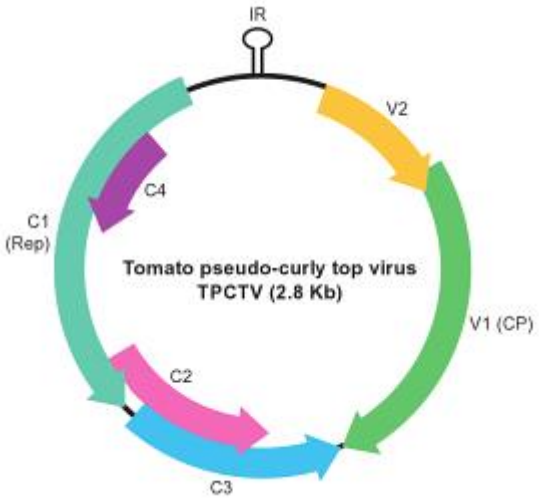

Figure 2.3: Genomic organization of tomato pseudo-curly top virus (Hull, 2013).

2.3.4 Begomoviruses

The begomoviruses are transmitted by whiteflies which infect dicotyledonous plants (Lapierre and Signoret, 2004). The genus of begomoviruses contains more than 200 species (Fauquet *et al*., 2008). The bean golden mosaic virus (BGMV) is an example for begomovirus species (Hurst, 2011). Based on the genome organization, begomoviruses are divided into two main groups that are bipartite and monopartite (Hurst, 2011; Kumar *et al*., 2014). Bipartite begomovirus genomes consists two circular ssDNA molecules (DNA-A and DNA-B) that encapsidated in separate particles. The genome component is 2.5-2.6kb in size. Meanwhile, monopartite begomovirus genomes consists of one circular DNA molecule (DNA-A) where the genome component size is 2.5-2.6kb (Hull, 2013).

Cann (2001) stated that the DNA-A component of the bipartite begomoviruses can replicate autonomously and produce virions but requires DNA B for systemic infection. Begomovirus DNA-A consists of six open reading frames (ORFs) which is AC1 (AL1) encodes the replication initiation protein (Rep) (Saunders *et al*., 2008). AC2 (AL2) encodes a transcription-activator protein (TrAP) and AC3 (AL3) encodes the replication enhancer protein (Tiendrebeogo *et al*., 2008). Besides that, begomovirus DNA-A also has AC4 that determines the expression of symptoms, AV1 (AR1) encodes the coat protein (CP) and AV2 encodes the movement of protein and it is also called the 'precoat' ORF (Gaur *et al.*, 2013).

Meanwhile, their DNA-B only has two open reading frames (ORFs) which is BV1 (BR1) encodes the nuclear shuttle protein (NSP) and BC1 (BL1) encodes the movement protein that involved in cell-to-cell transfer (Gaur *et al.*, 2013). The plus (+) virion-sense strand represented by AV1, AV2 and BV1, whereas the negative (-)

complementary sense strand represented by AC1, AC2, AC3, AC4 and BC1 (Yadava *et al.*, 2010).

Whereas, the monopartite component of begomoviruses equivalent to DNA-A of the bipartite viruses (Briddon *et al.*, 2001). Navot *et al*. (1991) stated that Tomato yellow leaf curl virus (TYLCV) is the most notable and economically most significant example of a monopartite begomovirus. According to the previous research, many monopartite begomovirus always associated with satellite DNAs. Many of this single component genome of begomovirus associated with satellite DNAs that are either required for development of typical disease symptoms (betasatellites) or have no apparent effect or modulate disease symptoms (alphasatellites) (Briddon *et al.*, 2010).

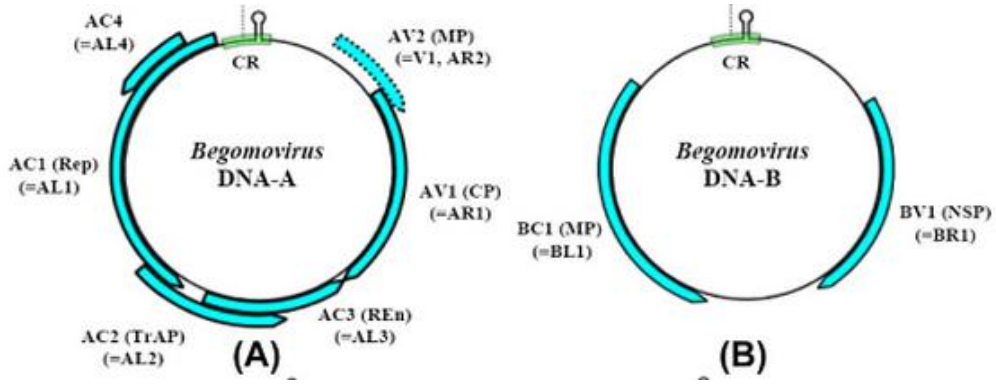

Figure 2.4: Genomic organization of begomoviruses (Gaur *et al*., 2013).

CHAPTER 3

METHODOLOGY

The raw data of geminivirus was provided by Dr Rozaimi Razali from Sengenics Sdn. Bhd. This raw data consists of sequencing short reads generated from Illumina Genome Analyzer (Kircher *et al*., 2009). The paired reads was used in this project. The raw data is in a fastq format files.

3.1 Performing the FastQC

The raw data was performed some simple quality control checks using FastQC (v0.11.2) which is able to report a wide range of information related to the quality of the reads (Leggett *et al*., 2013). The aim is to ensure that the raw data looks good and there are no problems or biases in our data which may affect the analysing result.

3.2 Performing the *De novo* Assembly

For the assembly, the sequencing reads generated from Illumina Genome Analyser was assembled by using SOAPdenovo (v2.04-r240). SOAPdenovo is a program in the SOAP package which is specially designed to handle the huge amounts of short reads generated by Illumina Genome Analyser (Li et *al*., 2008). SOAPdenovo is a novel short read assembly method that can build a *de novo* draft assembly for any genomes. This program also was the fastest assembler compared with others assembler. Besides that, this program was more efficient than other tools in terms of runtime and memory usage (Lin *et al*., 2011). There were two basic approaches in algorithms for

short read assemblers which are overlap graphs and De Brujin graph. SOAPdenovo is one of the assemblers that use De Brujin graph algorithm because overlap graphs do not scale well with increasing numbers of reads. De Brujin graphs reduce the computational effort by breaking reads into smaller sequences of DNA known as k-mers, where the parameter k denotes the length in bases of these sequences (Illumina, ND). The advantage in *de novo* assembly is the ability to identify novel genes that may reveal the unique characteristic about the species.

3.3 Predicting the Genes

After obtaining the assembled data from the *de novo* assembly, the scaffold output was used to predict the genes by using GeneMark. This program was designed and tuned for gene prediction in prokaryotic, eukaryotic and viral genomic sequences. There were two major programs in the GeneMark web software called GeneMark and GeneMark.hmm (Besemer and Borodovsky, 2005). The number of genes in the virus genome was predicted by using GeneMark.hmm (v3.25). The algorithm of GeneMark.hmm was designed to improve the gene prediction quality in terms of finding exact gene boundaries. It was shown that GeneMark.hmm program was more accurate in exact predictions compared with GeneMark program. Besides that, GeneMark.hmm program had the least number of missing genes and the highest percentage of annotated genes found exactly or partially (Lukashin and Borodovsky, 1998). The aim of gene prediction is to determine the ‘true’ functional sequence for the virus DNA sequence.

3.4 Annotating the Genes

After the genes were predicted by GeneMark, in order to determine the function of the predictive genes, the amino acid sequences of the predictive genes were used in this step. Three different tools, InterPro, UniProt and Gene Ontology were selected to perform the annotation of the protein sequences. These tools are all publicly available, and currently often used to determine the function of the protein sequence. InterPro is a resource that provides functional analysis of protein sequences by classifying them into families and predicting the presence of domains and important sites (InterPro, ND). This tool combines signatures (predictive models) from PROSITE, PRINTS, Pfam, ProDom, SMART and TIGRFAMs into a single searchable resource (Mulder *et al*., 2002). Meanwhile, the Universal Protein Resource (UniProt) provides the scientific community with a single, centralized, authoritative resource for protein sequences and functional information. It was formed by uniting the Swiss-Prot, TrEMBL and PIR protein database activities (Bairoch *et al*., 2005). The Gene Ontology (GO) gives a different result compared with the InterPro and UniProt. This tool is a collaborative effort to develop and use ontologies to support biologically meaningful annotation of genes and their products in a wide variety of organisms (The Gene Ontology Consortium, 2008). It provides a systematic language, or ontology, for the consistent description of attributes of genes and gene products that describe about their biological processes, molecular functions and cellular components (Rhee *et al.*, 2008; The Gene Ontology Consortium, 2008).

3.5 Comparing the Genes

After obtaining the protein functions of the predictive genes, the predictive genes were compared with genes from three different geminivirus reference genome. The BLAST Ring Image Generator (BRIG) tool was used to compare the genes. The aim is to determine genotypic differences between closely related viruses. The BRIG is a Java –based tool that enables rapid visualisation of BLAST comparisons to one or more central reference sequences using complete, draft or unassembled genome data (Alikhan *et al*., 2011). The results were plotted as a series of rings, each representing a query sequence, which are coloured to indicate the presence of hits to the reference sequence (Edwards and Holt, 2013).

3.6 Performing the Phylogenetic Tree

The scaffold sequence from the de novo assembly result was aligned using the BLAST program for phylogenetic analysis. The scaffold sequence was BLAST against the geminivirus reference genomes in the NCBI. Then, from the BLAST results, the phylogenetic tree was created. Cantor (1969), quoted by NCBI News (2006) stated that the phylogenetic tree display was created from genetic distances calculated using standard method from the aligned sequences that is Jukes-Cantor method for nucleotide comparisons. The phylogenetic analyses were done using the Neighbor-Joining method. This method was proposed for reconstructing phylogenetic trees from evolutionary distance data (Saitou and Nei, 1987). The aim to perform phylogenetic tree is to determine the evolutionary relationship between a set of homologous characters of one or several organisms (Schreiber, 2007). Besides that, the aim of the phylogenetic tree in this study was to determine the origin of an unknown geminivirus.

CHAPTER 4

RESULTS

4.1 Analysis of FastQC

Figure 4.1 shows that the average Phred score in a single read (1-100 base pair) from the fastq file. From the observation, the Phred score for all the bases in a single read was above 30. For example, base one until base three the Phred score is 34. The red lines represent the median for each base.

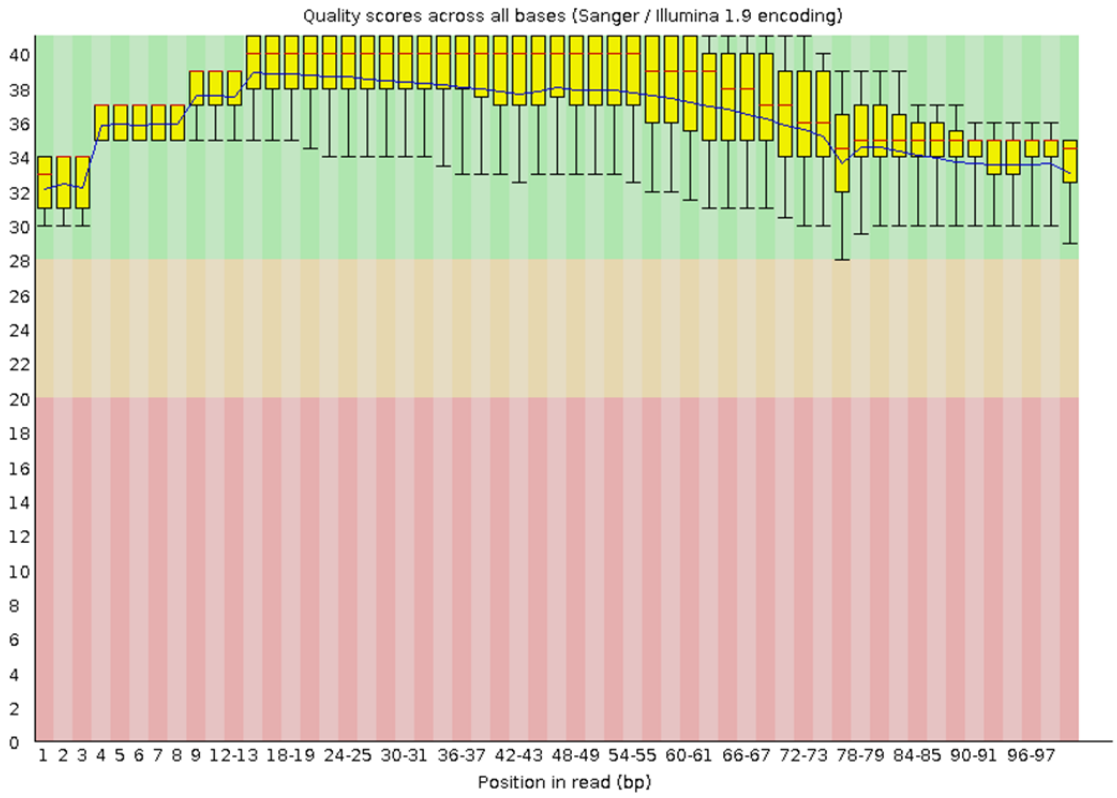

Figure 4.1: Per base sequence quality

Figure 4.2 shows that the average Phred score for 2 million sequences reads from the fastq file. From the observation, the average Phred score for 2 million sequences reads was between 38 until 40.

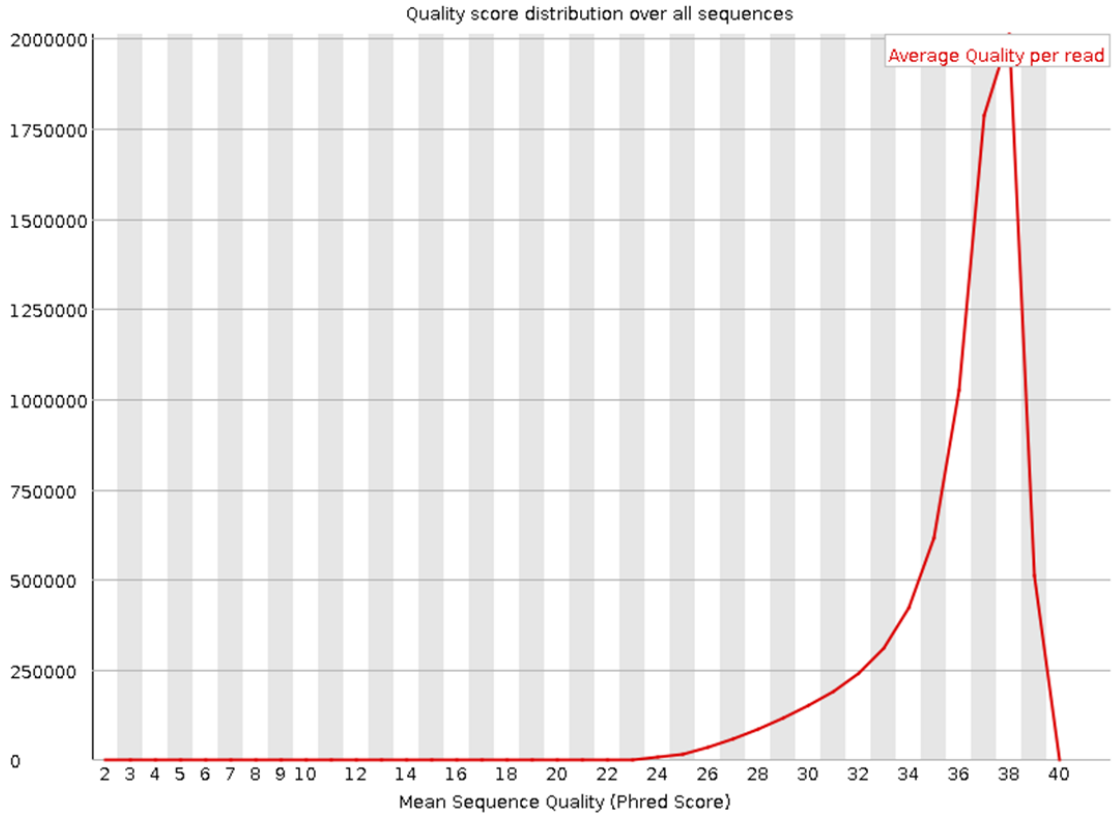

Figure 4.2: Per sequence quality scores

4.2 Analysis of De novo Assembly

Table 4.1 shows the *de novo* assembly statistic for an unknown geminivirus genome was carried out using SOAPdenovo software. The *de novo* assembly statistic of an unknown geminivirus was compared with the reference-guided assembly statistics of the geminiviruses reference genome (Table 4.2). The geminivirus reference genomes (all complete genome) were obtained from NCBI database. From the observation, the genome size of an unknown geminivirus was small (905kb) compared with the geminivirus reference genome. Besides that, the unknown geminivirus genome only has 5 genes compared with the geminiviruses reference genome which all have 6 genes. The result also shows that the unknown geminivirus has low GC content (39.17%), while the geminivirus reference genome has high GC content (41%-44%).

Table 4.1: The *de novo* assembly statistics for an unknown geminivirus genome.

Type of Geminivirus	Unknown Geminivirus
Genome size (Kb)	905
Number of contig	3120
Number of scaffold	2971
N50	699
GC content (%)	39.17
Number of genes	5

Table 4.2: Reference-guided assembly statistics for the geminivirus reference genomes.

Type of Geminivirus	Tobacco curly shoot virus, complete genome (NC_003722.1)	Papaya leaf curl China virus - [G8], complete genome (NC_005321.1)	Sweet potato leaf curl virus, complete genome (NC_004650.1)
Author name	Li *et al*. (2004)	Wang *et al*. (2004)	Lotrakul and Valverde (1998)
Genome size (Kb)	2.74	2.75	2.78
Number of contig	Not available	Not available	Not available
Number of scaffold	Not available	Not available	Not available
N50	Not available	Not available	Not available
GC content (%)	41.5	42.4	44.0
Number of genes	6	6	6

Table 4.3 shows the result of scaffold after blast against geminivirus reference genomes (all complete genome). There were 8 geminivirus reference genomes that obtained from the NCBI database. The result shows that only one scaffold (C11095) from 2971 scaffold that mapped to the geminivirus reference genomes. The geminivirus reference genome (AJ566744.1) has the highest identity percentage (92.23%) with the scaffold sequence. Meanwhile, the geminivirus reference genome (AB079689.1) has the lowest identity percentage (92.23%) with the scaffold sequence.

Table 4.3: The scaffold output blast against the geminivirus reference genomes.

Query ID	Scaffold ID	Percent Identity (%)	Length
AB079689.1	C11095	78.2	1381
AB055009.1	C11095	79.66	1367
NC_014596.1	C11095	85.91	1526
GU001879.1	C11095	79.38	1508
NC_009553.1	C11095	83.9	1491
HM164547.2	C11095	78.79	1386
AJ566744.1	C11095	92.23	1325
NC_002817.1	C11095	79.21	1371

4.3 Analysis of Gene Prediction

Table 4.4 shows that there were 5 predictive genes in the unknown geminivirus genome that was predicted using GeneMark.hmm software. The GeneMark.hmm also gives the output of the amino acid sequences and the nucleotide sequences of each predictive gene.

Table 4.4: The predictive genes for the unknown geminivirus genome.

Gene ID	Amino acids	Nucleotide sequence
Gene 1	TWRFLRVFRVQCMKYLD NLGVISINNVISACDHVL WNVLEKTEYVTQSSIIKF NLY	ACTTGGCGTTTCTTAAGAGTATTT AGGGTTCAATGTATGAAATATTT AGATAATTTGGGTGTAATTAGTAT TAATAATGTAATTAGCGCATGTG ATCATGTATTATGGAACGTATTGG AAAAAACAGAATATGTAACACAG TCTAGTATAATAAAATTCAATCTT TATTAA
Gene 2	MYRKPRLYRMYRSPDVP KGCEGPCKVQSYEQRHDI SHVGKVLCVSDVTRGNG LTHRVGKRFCVKSVYVL GKIWMDENIKTKNHTNT VMFYLVRDRRPYGSAMD FGQVFNMYDNEPSTATIK NDLRDRYQVLRKFTSTVT GGQYASKEQALVRKFMK INNYVVYNHQEAAKYDN HTENALLLYMACTHASN PVYATLKIRIYFYDSVQN	ATGTATCGCAAGCCCAGACTGTA CAGAATGTACAGAAGCCCTGATG TGCCCAAAGGTTGTGAAGGCCCG TGTAAGGTCCAATCGTATGAACA GAGGCACGACATATCCCACGTTG GTAAAGTATTATGTGTTAGTGATG TCACTCGTGGTAATGGGCTTACAC ATCGTGTGGGTAAGAGATTCTGT GTGAAATCTGTCTACGTGTTGGGT AAAATATGGATGGATGAAAATAT CAAAACCAAGAACCATACCAACA CTGTGATGTTTTATCTTGTTCGTG ATAGAAGGCCCTATGGTTCTGCT ATGGATTTTGGTCAGGTGTTTAAC ATGTATGATAATGAGCCCAGCAC TGCTACTATCAAGAATGATCTTCG AGATCGTTATCAAGTGTTAAGGA AATTCACTTCAACAGTTACCGGTG GTCAATATGCGTCTAAGGAGCAG GCATTGGTCAGGAAGTTTATGAA GATTAATAATTATGTAGTTTATAA TCATCAAGAAGCTGCTAAGTATG ACAACCATACTGAGAATGCGTTG TTATTGTATATGGCTTGTACTCAT GCCAGTAATCCAGTGTATGCTACT TTGAAGATCAGGATCTATTTTTAT GATTCTGTTCAAAATTAA

Table 4.4, continued.

Gene ID	Amino acids	Nucleotide sequence
Gene 3	MGLLVHVLLLLVTRTVG AAARLITGESKFRRRRTFE AGVEMTISAGRFDIIASTD Y	ATGGGCCTGTTGGTCCATGTCCT TCTTTTGTTGGTGACGAGGACAG TGGGGGCAGCAGCACGGCTCAT CACGGGGGAGTCGAAGTTCAGA CGGCGACGTACCTTCGAGGCGG GAGTGGAAATGACTATATCGGC GGGTCGCTTCGACATAATTGCG AGCACGGATTACTGA
Gene 4	MNTLRGPDSSADMAPPK RFLINCKNYFLTYPQCSLT KEEALSQLQNLNTPTNKK YIKICRELHEDGSPHLHVL IQFEGKYKCQNNRFFDLIS PTRSAHFHPNIQGAKSSSD VKSYIDKDGDTLEWGEFQ IDGRSARGGQQTANDAY AQALNSGSKSEALNVIKE LAPKDYVLQFHNLNANL DRIFAPPLEVFVCPFLSSSF DQVPEELEEWVSENVKD AAARPWRPKSIVVEGESR TGKTMWARSLGPHNYLC GHLDLSPKVYSNAAWYN VIDDVDPHYLKHFKEFMG AQRDWQSNTKYGRPIQIK GGIPTIFLCNPGPTSSYKE YLEEEKNSALKAWAIKN AEFITLTEPLYSGTHQSAT QNSQEETNPQAES	ATGAATACTTTGAGAGGACCTG ATAGCTCTGCTGACATGGCACCT CCAAAGAGATTTTTAATAAATTG CAAAAATTATTTCCTCACTTATC CACAGTGCTCTCTCACTAAGGA AGAAGCACTTTCCCAATTACAA AACCTAAACACACCAACAAATA AAAAATATATTAAAATCTGCAG AGAGCTTCACGAAGATGGGAGC CCTCATCTGCACGTGCTTATCCA GTTCGAGGGGAAATACAAGTGC CAGAATAACAGATTCTTCGACCT TATATCCCCAACCAGGTCAGCA CATTTCCATCCGAACATTCAGGG AGCTAAATCCAGCTCCGACGTC AAGTCTTATATCGACAAGGACG GAGACACCCTCGAATGGGGAGA ATTTCAGATCGACGGAAGATCT GCAAGAGGGGGGCAACAGACA GCCAACGACGCTTACGCCCAGG CGCTTAACAGCGGCAGTAAGTC AGAGGCTCTTAACGTAATTAAG GAGTTAGCTCCTAAAGATTATGT TTTACAATTTCACAATTTAAATG CTAATTTAGATAGGATTTTTGCA CCTCCTTTAGAGGTTTTTGTTTG TCCTTTTTTATCTTCTTCTTTCGA TCAAGTTCCCGAAGAACTTGAA GAGTGGGTTTCCGAGAATGTGA AGGATGCCGCTGCGCGGCCATG GAGACCCAAGAGTATAGTTGTA GAGGGCGAGAGTCGTACAGGGA AGACAATGTGGGCCAGATCATT AGGCCCACATAATTATCTGTGCG GTCACCTCGACCTGAGTCCTAAA GTGTACAGCAATGCTGCTTGGTA CAACGTCATTGATGACGTAGAC CCCCACTACCTAAAGCACTTTAA AGAATTCATGGGGGCCCAAAGG GACTGGCAAAGCAACACCAAGT ACGGGAGGCCAATTCAAATTAA

		AGGTGGAATTCCCACTATCTTCC TCTGCAATCCAGGCCCAACGTC ATCATATAAAGAGTACTTGGAA GAGGAAAAGAATTCCGCACTCA AAGCGTGGGCAATAAAGAATGC AGAATTCATCACCCTCACCGAA CCACTGTACTCAGGTACCCATCA AAGTGCAACACAGAATAGCCAA GAGGAGACCAATCCGCAGGCGG AGAGTTGA
Gene 5	MSPDLEAQKTLVYSQRFP QVVVELYLDRYDVVVGV EWPLVVLGYLEIEGISDRP SIHATLALSCSE	ATGAGTCCTGATCTGGAAGCTC AGAAAACACTGGTGTATTCCCA ACGCTTTCCTCAGGTTGTGGTTG AACTGTATCTGGATCGTTATGAT GTCGTGGTTGGTGTTGAATGGCC TCTCGTGGTGCTTGGTTATCTTG AAATAGAGGGGATTTCTGATCG TCCAAGTATACACGCCACTCTCG CATTGAGTTGCAGTGAGTAG

4.4 The Annotation of the Predictive Genes in the Unknown Geminivirus Genome

There were three different types of annotation tools that were used in this study in order to predict the function of the protein sequence of each predictive gene. The function of each protein sequence was annotated using the InterPro, Gene Ontology (GO) and UniProt. Table 4.5 shows that the annotation tools give different result about the protein function. However, only Gene Ontology gives specific information of the protein sequence about their biological process, molecular function and also cellular component.

Table 4.5: The annotation of the predictive genes in the unknown geminivirus genome.

Genes	Annotation Tools		
	InterPro	Gene Ontology (GO)	UniProt
Gene 1	1) Geminivirus AL3 coat protein (IPR000657)	**Biological Process:** 1)viral process (GO:0016032)	Replication enhancement protein
Gene 2	1)Geminivirus AR1/BR1 coat Protein (IPR000263)	**Molecular Function:** 1)structural molecule activity (GO:0005198) **Cellular Component:** 1)viral capsid (GO:0019028)	Coat protein
Gene 3	Not reported	Not reported	Hypothetical protein
Gene 4	1)Geminivirus AL1 replication-associated protein, catalytic domain (IPR022690) 2) Geminivirus AL1 replication-associated protein, central domain (IPR022692)	**Biological Process:** 1)DNA replication (GO:0006260) **Molecular Function:** 1)structural molecule activity (GO:0005198) 2)endodeoxyribonuclease activity, producing 5'-phosphomonoesters (GO:0016888) **Cellular Component:** 1)viral capsid (GO:0019028)	Replication – associated protein
Gene 5	Not reported	Not reported	Glyoxylate carboligase

4.5 Analysis of the Genes Comparison

Figure 4.3 shows that the image of genes comparison between the predictive genes and the reference genomes (all complete genome) that created using BRIG software. There were three different reference genomes that were obtained from NCBI database. The first ring represent GC content, the second ring represent Tobacco leaf curl Kochi virus-[KK] DNA, complete genome, isolate: TLCV-KK (AB055009.1), the third ring represent Tobacco leaf curl Japan virus-[JP3] DNA, complete genome, isolate: TLCV-Jp3 (AB079689.1), the fourth ring represent Tobacco leaf curl Pusa virus DNA-A, complete genome (NC_014596.1) and the last ring represent the unknown geminivirus genome that consists of 5 predictive genes.

Meanwhile, Table 4.6 shows the summary of the genes comparison between the predictive genes and the reference genomes. From the observation, the unknown geminivirus and the reference geminivirus has genes that encodes for replication-associated protein and coat protein. However, only the unknown geminivirus has a gene that encodes for glyoxylate carboligase.

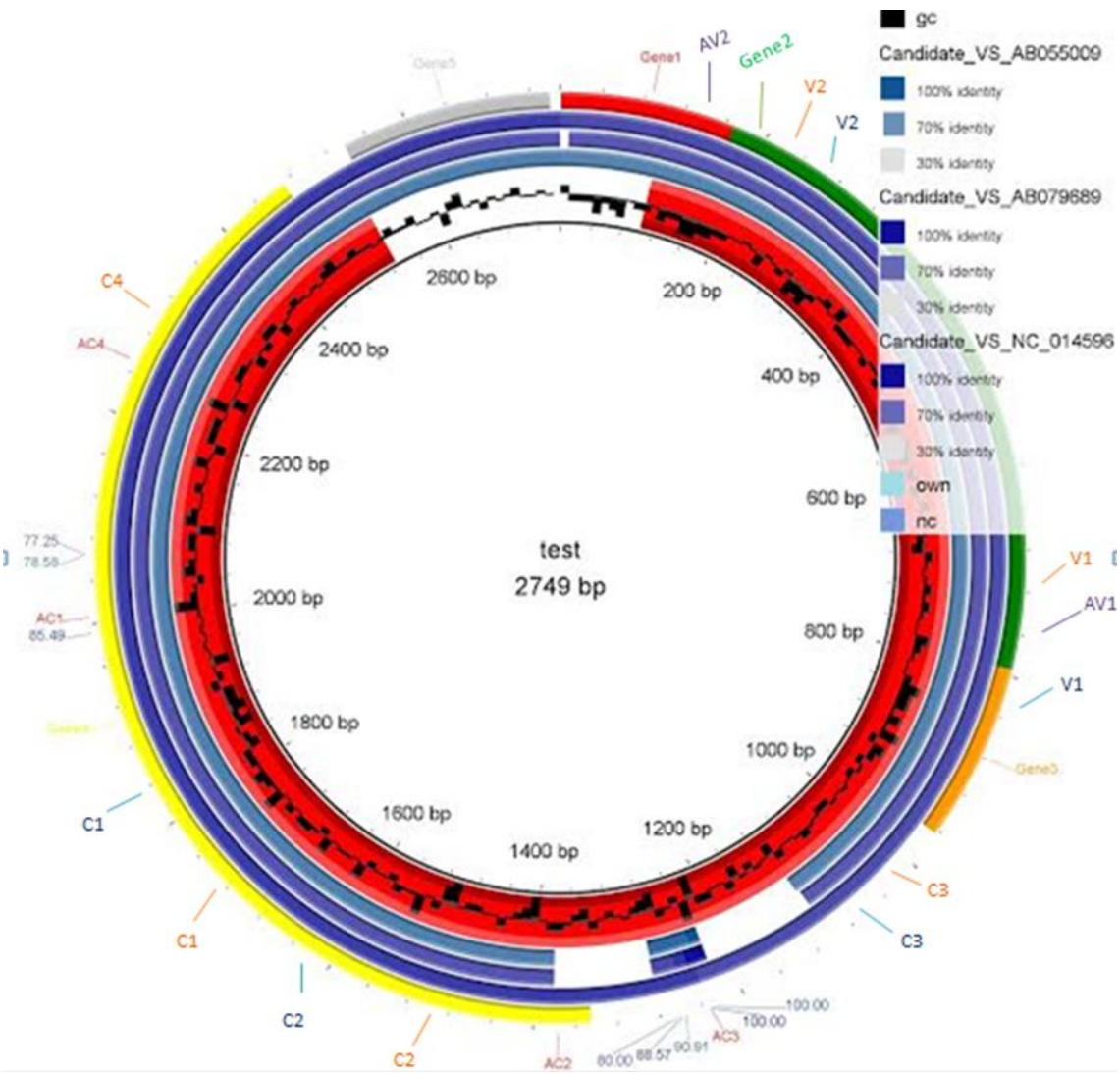

Figure 4.3: The genes comparison between the predictive genes and the reference genomes using BRIG.

Table 4.6: The summary of genes comparison between the predictive genes and the reference genomes.

Virus names	GenBank ID	Number of genes	Gene names
Unknown geminivirus	-	5	1) Replication enhancement protein (Gene 1) 2) Coat protein (Gene 2) 3) Hypothetical protein (Gene 3) 4) Replication –associated protein (Gene 4) 5) Glyoxylate carboligase (Gene 5)
Tobacco leaf curl Kochi virus-[KK] DNA, complete genome, isolate: TLCV-KK	AB055009.1	6	1) V2 protein (V2) 2) Coat protein (V1) 3) C3 protein (C3) 4) C2 protein (C2) 5) Replication-associated protein (C1) 6) C4 protein (C4)
Tobacco leaf curl Japan virus-[JP3] DNA, complete genome, isolate:TLCV-Jp3	AB079689.1	6	1) V2 protein (V2) 2) Coat protein (V1) 3) C3 protein (C3) 4) C2 protein (C2) 5) Replication-associated protein (C1) 6) C4 protein (C4)
Tobacco leaf curl Pusa virus DNA-A, complete genome	NC_014596.1	6	1) Pre-coat protein (AV2) 2) Coat protein (AV1) 3) C3 protein (AC3) 4) C2 protein (AC2) 5) Replication-associated protein (AC1) 6) C4 protein (AC4)

4.6 Analysis of the Phylogenetic Tree of an Unknown Geminivirus

Figure 4.4 shows that the phylogenetic tree for the unknown geminivirus with the geminivirus genomes from the NCBI database. The geminiviruses in the phylogenetic tree can be classified into two groups that are East Asia virus and Southeast Asia viruses. From the observation, the unknown geminivirus (candidate virus) was in a group of Southeast Asia virus. It share common ancestor with Tobacco leaf curl Indonesia virus C1, V2, V1 genes for replication-associated protein, putative V2 protein, coat protein, partial and complete cds.

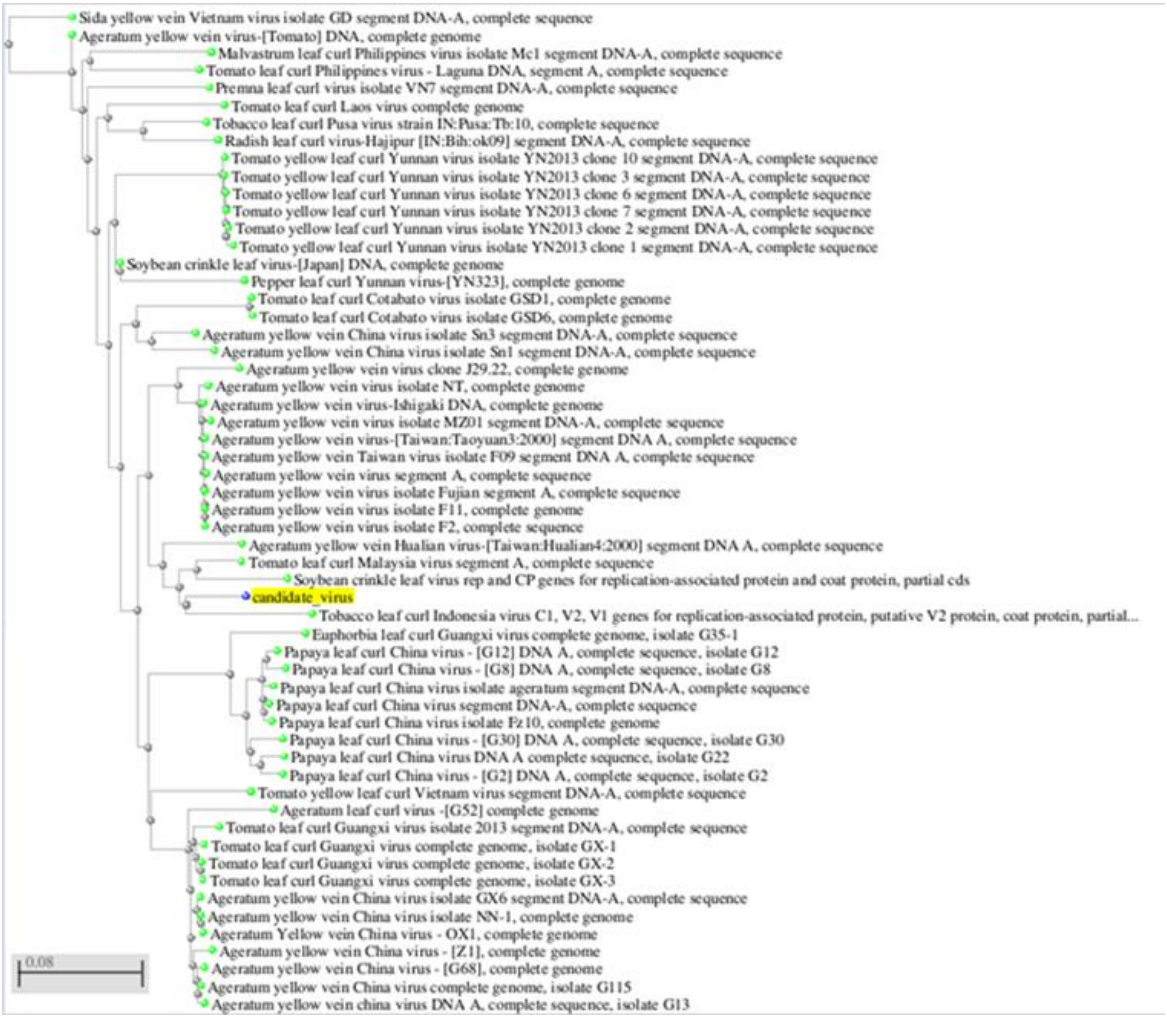

Figure 4.4: The phylogenetic tree for the unknown geminivirus with the geminivirus reference genomes.

CHAPTER 5

DISCUSSION

5.1 FastQC Evalution Data

In this study, the raw fastq was taken to perform the quality check using FastQC. The quality of the raw data or the sequencing accuracy was measured using Phred quality scoring (Q score). It indicates the probability that a given base is called incorrectly by the sequencer. The raw data shows a good quality because the average Phred quality score in a single read was above 30, while the average Phred quality score for 2 million reads was between 38 until 40. The short sequencing reads from Illumina has a good quality if the Q score is above 30 because the probability of incorrect base call is 0.001. Based on the previous research, Illumina sequencing by synthesis (SBS) technology delivers the highest percentage of error-free reads, with a vast majority of bases having quality scores above Q30 (Illumina, ND). For the Illumina platform, indel errors are rare, but the overall miscall error rate is typically around 1% (Nielsen *et al*., 2011).

5.2 Problem Arise

5.2.1 Detection of the Tobacco Genome

After performing the *de novo* assembly procedure using SOAPdenovo on the sequence reads (raw fastq), in order to identify the accuracy of the geminivirus, a few scaffolds were taken to perform blast against geminivirus genomes. We found out that majority of the scaffolds was mapped to the tobacco genomes instead of geminivirus genomes.

However, this problem can be solved by aligned the short sequencing reads (raw fastq) to the tobacco reference genome using bowtie (a next-generation specific read alignment program) instead of BLAST. This is because BLAST method is not specialized for the vast amount of data generated by next-generation sequencers. The Bowtie package enables ultrafast and memory-efficient alignment of large sets of sequencing reads to a reference sequence. This tool aligns reads with the aid of an index of the reference genome to achieve both speed and memory efficiency (Langmead, 2011).

In this step, we were looking for unaligned sequences between the fastq and tobacco reference genome. We were not interested in looking at aligned sequences between the fastq and the tobacco reference genome. This is because the unaligned sequences may contain our geminivirus genome. After obtaining the unaligned sequences, these unaligned sequences were used to perform the *de novo* assembly using SOAPdenovo.

5.3 *De novo* Assembly

After performing *de novo* assembly on the unaligned sequences, the scaffolds output was taken to perform BLASTN with a list of geminivirus genomes (all complete genome). The result shows that only one scaffold (C11095) from 2971 scaffolds was found to have high sequence alignment with the geminivirus reference genomes. Scaffold is an ordered assembly of contigs with gaps in between (Baker, 2012)

Besides that, from the assembly result using SOAPdenovo, the genome size of an unknown geminivirus was 905kb which is quite small compared with the genome size of the reference geminivirus. Mostly the reference genomes have a genome of 2.7kb in size. For example, the genome size of tobacco curly shoot virus, complete genome (NC003722.1) was 2.74kb. However, the genome size of an unknown geminivirus was quite similar with the genome size of nanovirus. Apart from geminivirus, nanovirus also the single-stranded DNA (ssDNA) plant virus, the nanovirus genomes consist of multiple circular ssDNA approximately 1kb in size (Timchenko *et al*., 2006).

5.4 The Gene Prediction

Once the ordered set of contigs (scaffold) has been obtained, the next step is to predict the number of genes for the draft genome. In this study, the number of genes was predicted using GeneMark.hmm. This tool can be used for gene finding in prokaryotes, eukaryotes and also viruses (Besemer and Borodovsky, 2005). When the GeneMark.hmm was applied to the scaffold sequence of an unknown geminivirus, there were 5 open reading frames (ORFs) predicted as genes were identified.

Previous study shows that when the GeneMark.hmm was applied to the E.coli genomic sequence, as many as 4440 genes were identified (Lukashin and Borodovsky, 1998). The nucleotide sequences of predicted genes and translated protein sequences are available as an output (Besemer and Borodovsky, 2005) to facilitate further analysis, such as BLAST searching (Altschul *et al.*, 1990).

5.5 The Gene Annotation

The output of translated protein sequences that provided by GeneMark was used to perform further analysis such as gene annotation. Annotation is the process of finding the function of the predictive genes. The translated protein sequences for each predictive gene were annotated using three different tools, InterPro, Gene Ontology (GO) and UniProt. The InterPro and UniProt tool only give the information about the product of the gene, while GO tool give more information about the biological process, molecular function and cellular component of the gene.

From the InterPro result, the gene 1 encodes the geminivirus AL3 coat protein, while the UniProt result shows that the gene 1 encodes the replication enhancement protein which is similar with the AL3 of begomovirus that encodes the replication enhancer protein (Tiendrebeogo *et al.,* 2008). The GO shows that the gene 1 was involved in the viral process (biological process).

Meanwhile, the gene 2 encodes the geminivirus AR1/BR1 coat protein (InterPro result) and the result from the UniProt also shows that gene 2 encodes the coat protein. Based on the previous research, AR1 of begomovirus encodes the coat protein and BR1 encodes the nuclear shuttle protein (NSP). The GO shows that gene 2 was involved in the structural molecule activity (molecular function) and also involved in the viral capsid (cellular component).

The InterPro result shows that the gene 4 encodes the geminivirus AL1 replication-associated protein, catalytic domain and central domain. The UniProt also shows that gene 4 encodes the replication-associated protein. The previous study shows that AL1 of begomovirus encodes the replication initiation protein (Rep) (Saunders *et al.,* 2008). From the GO result, the gene 4 was in the DNA replication (biological process), involved in the structural molecule activity and endodeoxyribonuclease activity which produces 5'-phosphomonoesters (molecular function), and this gene also involved in the viral capsid (cellular component).

Whereas for gene 3 and gene 5, there were no function reported from the InterPro and the GO. The function of the both genes was reported by the UniProt where the gene 3 was a hypothetical protein, while the gene 5 encodes glyoxylate carboligase. Both of these genes can be carried out for the further study to determine their function by using another annotation tool.

5.6 Comparison of the Predictive Genes with the Reference Genomes

The predictive genes were compared with the three reference genomes using the BRIG (BLAST Ring Image Generator). The BRIG image shows that the large sequence of the unknown geminivirus was missing between 1000 bp until 1300 bp. The result shows that only Tobacco leaf curl Pusa virus DNA-A, complete genome (NC_014596) has a complete sequence although another two reference genomes also has a complete genome. But the sequence of both reference genomes (AB055009.1 and AB079689.1) was missing between 1100bp until 1200bp and between 1300bp until 1400bp.

Basically, all the viruses have common characteristics. The virus has a coat protein or also known capsid. The main functions of viral capsids (coat protein) are to protect, transport and deliver their genome (Roos *et al*., 2007). From Table 4.6 shows that all the geminivirus encode the coat protein. For example, geminivirus manage the transport of their DNA within plants with the help of three proteins, the coat protein (CP), the nuclear shuttle protein, and the movement protein (MP) (Hehnle *et al*., 2004).

Besides that, the result also shows that all the geminivirus encodes the replication-associated protein. This is because the geminiviral replication-associated protein (Rep) is the only viral protein required for viral DNA replication (Behjatnia *et al.,* 1998). Compared with other reference genome, the gene 1 of an unknown geminivirus encodes the replication enhancement protein. The previous study shows that the C3 (Ren) of curtoviruses encodes a replication enhancer protein (Varsani *et al.,* 2014). Interestingly, only the gene 5 of an unknown geminivirus encodes the glyoxylate carboligase. Glyoxylate carboligase (GCL) (EC 4.1.1.b) is a thiamine diphosphate (ThDP)-dependent enzyme, which catalyzes the decarboxylation of glyoxylate and ligation to a second molecule of glyoxylate to form tartronate semialdehyde (TSA) (Nemeria *et al.,* 2012). However, the specific function of the glyoxylate carboligase in geminivirus is unknown.

5.7 The Phylogenetic Tree of an Unknown Geminivirus

In this study, the phylogenetic result shows that the geminiviruses can be classified into the East Asia (China, Taiwan, and Japan) and the Southeast Asia (Malaysia, Indonesia, Philippines and Vietnam) viruses. The unknown geminivirus (candidate virus) was located in the Southeast Asia group. This phylogenetic tree indicates that the unknown geminivirus share common ancestor with Tobacco leaf curl Indonesia virus C1, V2, V1 genes for replication-associated protein, putative V2 protein, coat protein, partial and complete coding sequence (cds). The result from phylogenetic tree suggests that the unknown geminivirus could be a Southeast Asia strain and it could be attack tobacco plants.

Tobacco is an important commercial crop. For example, in Colonial America, tobacco was a successful cash crop and it was used as legal tender (Scholthof, 2004) and in Cuba, tobacco (*Nicotiana tabacum*) is an important crop with 34 000 ha under cultivation (Moran *et al*., 2006). Nowadays, tobacco used not just for smoking but also can be for medical purpose. According to the article "Tobacco as Medicine, Indian Research Institute Wins Patent" on 02 December 2007, a medicine extracted from tobacco can be used in the manufacture of cancer and cardiac drugs known as 'solansole'.

The present of this virus in tobacco plants tend to cause disease. Such as the leaf curl disease of tobacco (TbLCD) which has been discovered for nearly a century (Singh *et al.*, 2012). Lucas (1975) quoted by Paximadis and Rey (2001) reported that TbLCD had a destructive effect on tobacco agriculture in East Africa, Zimbabwe and the Transvaal of South Africa. The virus could cause substantial yield losses in tobacco crops which will affect the economy of the producer country. Hina *et al*. (2012) stated that geminivirus are responsible for a considerable amount of crop damage worldwide.

Besides that, it was reported that tobacco leaf curl geminivirus (TLCV) is a serious pathogen that cause substantial yield losses in tobacco and tomato crops (Shimizu and Ikegami, 1999).

In addition, phylogenetic analysis of the sequence of the unknown geminivirus indicated that they are closely related to the ageratum yellow vein virus cluster, tomato leaf curl virus, soybean crinkle leaf virus and ageratum yellow vein virus isolate. Besides that, the phylogenetic tree also shows that papaya leaf curl China virus cluster are closely related to the cluster of tomato leaf curl Guangxi virus which includes ageratum yellow vein China virus.

Apart from the unknown geminivirus, there were only 5 type species of geminiviruses from the Southeast Asia strain such as malvastrum leaf curl Philippines virus isolate, tomato leaf curl Laos virus, tomato leaf curl Malaysia virus, soybean crinkle leaf virus and tobacco leaf curl Indonesia virus. Based on the previous report, soybean crinkle leaf disease occurred on soybean in many growing areas of Thailand (Iwaki *et al*., 1983). From the phylogenetic analysis, example of the geminiviruses from the East Asia strain is papaya leaf curl China virus, tomato yellow leaf curl Yunnan virus and ageratum yellow vein Taiwan virus isolate.

5.8 Future Study

This project can be improved in many ways for the future study. The parameter of SOAPdenovo in this study was default. Hence, one of the ways to improve this project is by changing the parameter of SOAPdenovo in order to get the higher N50. N50 is average sequence length from a list of sequence reads. Larger N50 values correlate to more complete assemblies (Ion Torrent, ND). In order to improve the result of gene prediction is by changing the parameter of GeneMark because the default parameter was used in this project. Besides that, there was slightly difference between the results from gene annotation with the previous research. For example, the InterPro result shows that the gene 1 encodes the geminivirus AL3 coat protein, while the previous study proposed that AL3 of begomovirus encodes the replication enhancer protein (Tiendrebeogo *et al*., 2008). This difference can be carried out for the further study.

The most important way to improve this project is we need to assess the assembled data from the assembler. The main point of this step is to assess the quality of the assembled data. In this study, we did not perform this step because we could not find the tool that specific to assess the quality of the assembled data from viruses. Mostly, the tool is used to assess the assembled data from eukaryotes and prokaryotes. For example, CEGMA method was used to identify the genes that highly conserved among all eukaryotes (Parra *et al*., 2007). Besides that, the short sequencing reads generated from Illumina Genome Analyzer can be used to identify the single nucleotide polymorphisms (SNPs) which will be important for the geminiviruses study. The previous report proposed that Illumina Genome Analyzer technology can be used to identify single nucleotide polymorphisms (SNPs) accurately by mapping the short reads onto the known reference genome (Wang *et al*., 2008).

CHAPTER 6

CONCLUSION

In conclusion, this study was carried out to show the process in identifying an unknown sequence reads generated from Illumina Genome Analyzer. The results of the phylogenetic tree suggest that the unknown geminivirus could be a Southeast Asia strain and it could be attack tobacco plants. Besides that, the *de novo* assembly indicates that only one scaffold that mapped to the geminivirus genomes. By using GeneMark, there were 5 predictive genes was predicted from the unknown geminivirus sequence. The annotation of the predictive genes describes the functions of the protein sequences where the predictive genes encode the replication-enhancement protein (gene 1), the coat protein (gene 2), the hypothetical protein (gene 3), the replication-associated protein (gene 4) and glyoxylate carboligase (gene 5). The results from this study can be used for the further study and also can add knowledge about the geminivirus.

REFERENCES

Alikhan,N.-F., Petty, N. K., Zakour, N. L. B. and Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics, 12: 1-10.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic Local Alignment Search Tool. J. Mol. Biol., 215: 403–410.

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N. and Yeh, L.-S. L. (2005). The Universal Protein Resource (UniProt). Nucleic Acids Research, 33: 154-159.

Baker, M. (2012). *De novo* genome assembly: what every biologist should know. Nature Methods, 9: 333-337.

Baliji, S., Black, M. C., French, R., Stenger, D. C. and Sunter, G. (2004). Spinach curly top virus: A Newly Described *Curtovirus* Species from Southwest Texas with Incongruent Gene Phylogenies. Virology. Pythopathology, 94: 772-779.

Behjatnia, S. A. A., Dry, I. B. and Rezaian, M. A. (1998). Identification of the replication-associated protein binding domain within the intergenic region of tomato leaf curl geminivirus. Nucleic Acids Research, 26: 925–931.

Besemer, J. and Borodovsky, M. (2005). GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Research, 33: 451–454.

Bisaro, D. M. (1996). Geminivirus DNA Replication. Plant Biotechnology Center and Department of Molecular Genetics. The Ohio State University. Available URL: http://dnareplication.cshl.edu/content/free/chapters/30_bisaro.pdf

Bolok Yazdi, H. R., Heydarnejad, J. and Massumi, H. (2008). Genome characterization and genetic diversity of beet curly top Iran virus: A geminivirus with a novel nonanucleotide. Virus Genes, 36: 539-545.

Boulton, M. I. (2002). Functions and interactions of mastrevirus gene products. Physiological and Molecular Plant Pathology, 60: 243-255.

Briddon, R. W., Mansoor, S., Bedford, I. D., Pinner, M. S., Saunders, K., Stanley, J., Zafar, Y., Malik, K. A. and Markham, P. G. (2001). Identification of DNA Components Required for Induction of Cotton Leaf Curl Disease. Virology, 285: 234-243.

Briddon, R. W., Patil, B. L., Bagewadi, B., Nawaz-ul Rehman, M. S., Fauquet, C. M. (2010). Distinct evolutionary histories of the DNA-A and DNA-B components of bipartite begomoviruses. BMC Evolutionary Biology, 10: 1-17.

Cann, A. J. (2001). Principles of Molecular Virology (Standard Edition). Academic Press. Available URL: https://books.google.com.my/books?id=3WvYMKmIvY8C

Chen, L.-F., Vivoda, E. and Gilbertson, R. L. (2011). Genetic diversity in curtoviruses: a highly divergent strain of *Beet mild curly top virus* associated with an outbreak of curly top disease in pepper in Mexico. Arch Virol, 156: 547–555.

Dunowska, M., Biggs, P. J., Zheng, T. and Perrott M. R. (2012). Identification of a novel nidovirus associated with a neurological disease of the Australian brushtail possum (*Trichosurus vulpecula*). Veterinary Microbiology, 156: 418-424.

Fauquet, C. M., Briddon, R. W., Brown, J. K., Moriones, E., Stanley, J., Zerbini, M. and Zhou, X. (2008). Geminivirus strain demarcation and nomenclature. Arch Virology, 153: 783-821.

Fondong, V. N. (2013). Geminivirus protein structure and function. Molecular Plant Pathology, 14: 635-649.

Edwards, D. J. and Holt, K. E. (2013). Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. BioMed Central, 3: 1-9.

Gaur, R. K., Hohn, T. and Sharma, P. (2013). Plant Virus-Host Interaction: Molecular Approaches and Viral Evolution. Elsevier Science. Available URL: https://books.google.com.my/books?id=Kfj8AAAAQBAJ

Ghanim, M. and Czosnek, H. (2000). Tomato Yellow Leaf Curl Geminivirus (TYLCV-Is) is transmitted among Whiteflies (*Bemisia tabaci*) in Sex-Related Manner. Journal of Virology, 74: 4378-4745.

Hanley-Bowdoin, L., Bejarano, E. R., Robertson, D. and Mansoor, S. (2013). Geminiviruses: masters at redirecting and reprogramming plant processes. Nature Reviews Microbiology, 11: 777–788.

Hefferon, K. L. and Dugdale, B. (2003). Independent expression of Rep and RepA and their roles in regulating bean yellow dwarf virus replication. Journal of General Virology, 84: 3465-72.

Hehnle, S., Wege, C. and Jeske, H. (2004). Interaction of DNA with the Movement Proteins of Geminiviruses Revisited. Journal of Virology, 78: 7698–7706.

Hina, S., Javed, M. A., Haider, S. and Saleem, M. (2012). Isolation and sequence analysis of cotton infecting begomoviruses. Pak. J. Bot., 44: 223-230.

Hughes, F. L., Rybicki, E. P., and von Wechmar, M. B. (1992). Genome Typing of Southern African Subgroup-1 Geminiviruses. Journal of General Virology, 73:1031-1040.

Hull, R. (2013). Plant Virology. Academic Press. Available URL: https://books.google.com.my/books?id=PYrZAAAAQBAJ

Hurst, C. J. (2011). Studies in Viral Ecology: Microbial and Botanical Host Systems. John Wiley & Sons. Available URL: http://books.google.com.my/books?id=IU6IPBHATAMC

Illumina, (ND). *De novo* Assembly Using Illumina Reads. Available URL: http://www.illumina.com/Documents/products/technotes/technote_denovo_assembly_ecoli.pdf

Illumina, (ND). Quality Scores for Next-Generation Sequencing. Available URL: http://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf

InterPro, (ND). Available URL: http://www.ebi.ac.uk/interpro/about.html

Ion Torrent, (ND). *De novo* Assembly Using Ion Semiconductor Sequencing. Available URL:https://tools.lifetechnologies.com/content/sfs/brochures/de_novo_assembly_Ion_CO23721_App_Note_V9.pdf

Iwaki, M., Thongmeearkom, P., Honda, Y. and Deema, N. (1983). Soybean crinkle leaf: A new whitefly-borne disease of soybean. Plant Disease, 67: 546-548.

Kircher, M., Stenzel, U. and Kelso, J. (2009). Improved base calling for the Illumina Genome Analyzer using machine learning strategies. Genome Biology, 10:1-9.

Krupovic, M., Ravantti, J. J. and Bamford, D. H. (2009). Geminiviruses: a tale of a plasmid becoming a virus. BMC Evolutionary Biology, 9: 1-11.

Kumar, J.,Kumar, J., Singh, S. P. and Tuli, R. (2014). Association of satellites with a mastrevirus in natural infection: complexity of *Wheat dwarf India virus* disease. Journal Virology, 1-35.

Langmead, B. (2011). Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics, 1-24.

Lapierre, H. and Signoret, P. A. (2004). Viruses and Virus Diseases of Poaceae (Gramineae). Institut National De La Recherche Agronomique (France). Editions Quae. Available URL: https://books.google.com.my/books?id=K6jAYYMvElcC

Leggett, R. M., Ramirez-Gonzalez, R. H., Clavijo, B. J., Waite D. and Davey, R. P. (2013). Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. Frontiers in Genetics, 4: 1-5.

Li, R., Li, Y., Kristiansen, K. and Wang, J. (2008). SOAP: short oligonucleotide alignment program. Bioinformatics, 24: 713-714.

Lin, Y., Li, J., Shen, H., Zhang, L., Papasian, C. J. and Deng, H.-W. (2011). Comparative studies of *de novo* assembly tools for next-generation sequencing technologies. Bioinformatics, 27: 2031-2037.

Liu, L., Saunders, K., Thomas, C. L., Davies, J. W. and Stanley, J. (1999). Bean yellow dwarf virus RepA, but not Rep, binds to maize retinoblastoma protein, and the virus tolerates mutations in the consensus binding motif. Virology, 256: 270-279.

Lukashin, A. V. and Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. Nucleic Acids Research, 26: 1107-1115.

Moran, Y. M., Ramos, P. L., Dominguez, M., Fuentes, A. D., Sanchez, Y. and Crespo, J. A. (2006). Tobacco leaf curl Cuba virus, a new begomovirus infecting tobacco (*Nicotiana tabacum*) in Cuba. Plant Pathology, 55: 570.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., *et al*. (2002). InterPro: An integrated documentation resource for protein families, domains and functional sites. Briefings in Bioinformatics, 3: 225-235.

Navot, N., Pichersky, E., Zeidan, M., Zamir, D. and Czosnek, H. (1991). Tomato yellow leaf curl virus: A whitefly-transmitted geminivirus with a single genomic component. Virology, 185: 151–161.

NCBI News, (2006). New TreeView Display Option in NCBI BLAST. Available URL: file:///C:/Users/USER/SkyDrive/Documents/tree/NCBI.html

Nemeria, N., Binshtein, E., Patel, H., Balakrishnan, A., Vered, I., Shaanan, B., Barak, Z., Chipman, D. and Jordan, F. (2012). Glyoxylate carboligase: a unique thiamin diphosphatedependent enzyme that can cycle between the 4′-aminopyrimidinium and the 1′,4′-iminopyrimidine tautomeric forms in the absence of the conserved glutamate. Biochemistry, 51: 7940–7952.

Nielsen, R., Paul, J. S., Albrechtsen, A. and Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics, 12: 443-451.

Padidam, M., Beachy, R. N. and Fauquet, C. M. (1995). Classification and identification of geminiviruses using sequence comparisons. Journal of General Virology, 76: 249-263.

Parra, G., Bradnam, K. and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics, 23: 1061-1067.

Paximadis, M. and Rey, M. E. C. (2001). Genome organization of Tobacco leaf curl Zimbabwe virus, a new, distinct monopartite begomovirus associated with subgenomic defective DNA molecules. Journal of General Virology, 82: 3091-3097.

Peng, Y., Lai, Z., Lane, T., Nageswara-Rao, M., Okada, M., *et al*. (2014). *De Novo* Genome Assembly of the Economically Important Weed Horseweed Using Integrated Data from Multiple Sequencing Platforms. Plant Physiology, 166: 1241–1254.

Renteria-Canett, I., Xoconostle-Cazares, B., Ruiz-Medrano, R. and Rivera-Bustamante R. F. (2011). Geminivirus mixed infection on pepper plants: Synergistic interaction between PHYVV and PepGMV. Virology Journal, 8: 1-13.

Rhee, S. Y., Wood, V., Dolinski, K. and Draghici, S. (2008). Use and misuse of the gene ontology annotations. Nature Reviews Genetics, 9:509-515.

Roos, W. H., Ivanovska, I. L., Evilevitch, A. and Wuite, G. J. L. (2007). Viral capsids: Mechanical characteristics, genome packaging and delivery mechanisms. Cellular and Molecular Life Sciences, 64: 1484-1497.

Saitou, N. and Nei, M. (1987). The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees. Mol. Biol. Evol. 4: 406-425.

Saripalli, C. (2008). Annotation of Whitefly Expressed Sequence Tags and Validation of Genes with Potential Significance to Begomovirus Transmission. The University of Arizona. Plant Science. ProQuest. Available URL: https://books.google.com.my/books?id=H2SuhRbLziIC

Saunders, K., Briddon, R. W. and Stanley, J. (2008). Replication promiscuity of DNA-β satellites associated with monopartite Begomoviruses; deletion mutagenesis of the *Ageratum yellow vein virus* DNA-β satellite localizes sequences involved in replication. Journal of General Virology, 89: 3165-3172.

Scholthof, K.-B.G. (2004). Tobacco mosaic virus: A model system for plant biology. Annu. Rev. Phytopath, 42: 13-34.

Shimizu, S. and Ikegami, M. (1999). Complete Nucleotide Sequence and the Genome Organization of Tobacco Leaf Curl Geminivirus from Japan. Microbiol. Immunol., 43: 989-992.

Singh, M. K., Haq, Q. M. R. and Mandal, B. (2012). Evidence of the Association of Radish leaf curl virus with Tobacco Yellow Leaf Curl Disease in Bihar, India. Indian J. Virol, 23: 64-69.

The Gene Ontology Consortium, (2008). The Gene Ontology project in 2008. Nucleic Acids Research, 36: 440-444.

Tiendrebeogo, F., Traore, V. S. E., Baro, N., Traore, A. S., Konate, G. and Traore, O. (2008). Characterization of *Pepper yellow vein mali virus* in *Capsicum* sp. in Burkina Faso. Plant Pathology Journal, 7:155-161.

Timchenko, T., Katul, L., Aronson, M., Vega-Arreguı´n, J. C., Ramirez, B. C., Vetten, H. J. and Gronenborn, B. (2006). Infectivity of nanovirus DNAs: induction of disease by cloned genome components of *Faba bean necrotic yellows virus.* Journal of General Virology, 87: 1735–1743.

Tobacco as Medicine, Indian Research Institute Wins Patent (2007). Medindia: Drug News. Available URL: http://www.medindia.net/news/Tobacco-as-Medicine-Indian-Research-Institute-Wins-Patent-30130-2.htm

Varsani, A., Martin, D. P., Navas-Castillo, J., Moriones, E., Herna´ndez-Zepeda, C., Idris, A., Zerbini, F. M and Brown, J. K. (2014). Revisiting the classification of curtoviruses based on genome-wide pairwise identity. Arch Virol, 159: 1873-1882.

Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Zhang, J., *et al*. (2008). The diploid genome sequence of an Asian individual. Nature, 456: 60–65.

Yadava, P., Suyal, G. and Mukherjee, S. K. (2010). Begomovirus DNA replication and pathogenicity. Current Science, 98: 360-368.

Zhou, X. P., Xie, Y. and Zhang, Z. K. (2001). Molecular characterization of a distinct begomovirus infecting tobacco in Yunnan, China. Arch Virol, 146: 1599-1606.

APPENDICES

APPENDIX A

Table 8.1: Scaffold output blast against the geminivirus reference genomes.

Query ID	Scaffold ID	Percent Identity (%)	Length	X	Y	Start pos (Query)	End pos (Query)	Start pos (Scaffold)	End pos (Scaffold)	E-value	Bitscore
AB079689.1	C11095	78.2	1381	263	31	1256	2617	2743	1382	0	848
AB055009.1	C11095	79.66	1367	234	34	1240	2584	2749	1405	0	944
NC_014596.1	C11095	85.91	1526	184	25	1202	2707	2749	1235	0	1598
GU001879.1	C11095	79.38	1508	259	44	1266	2746	2717	1235	0	1014
NC_009553.1	C11095	83.9	1491	212	22	1251	2727	2737	1261	0	1399
HM164547.2	C11095	78.79	1386	261	30	1239	2609	2749	1382	0	900
AJ566744.1	C11095	92.23	1325	101	2	1246	2569	2737	1414	0	1875
NC_002817.1	C11095	79.21	1371	251	31	1261	2614	2735	1382	0	922

APPENDIX B

Table 8.2: Full name of geminivirus reference genomes from NCBI database.

Query ID	Full Name of Geminivirus
AB079689.1	Tobacco leaf curl Japan virus-[JP3] DNA, complete genome, isolate: TLCV-Jp3
AB055009.1	Tobacco leaf curl Kochi virus-[KK] DNA, complete genome, isolate: TLCV-KK
NC_014596.1	Tobacco leaf curl Pusa virus DNA-A, complete genome
GU001879.1	Tobacco curly shoot virus - [SC118], complete genome
NC_009553.1	Tobacco leaf curl Thailand virus, complete genome
HM164547.2	Tobacco leaf curl virus isolate TLCV-Korea-KJ, complete genome
AJ566744.1	Tobacco leaf curl Yunnan virus - [Y161] complete genome, isolate Y161
NC_002817.1	Tobacco leaf curl Zimbabwe virus, complete genome

APPENDIX C

Table 8.3: Quality scores and base calling accuracy (Illumina, ND).

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%