Evaluation and Annotation of *Bemisia tabaci* Transcriptome Data using Paired-end RNA Sequencing

NIKUNJ SATISHCHANDRA NAIK

A thesis submitted in partial fulfilment

of the requirements of University of Greenwich

for the degree of Master of Philosophy

March 2017

DECLARATION

"I certify that this work has not been accepted in substance for any degree, and is not concurrently being submitted for any degree other than that of (Master of Philosophy) being studied at the University of Greenwich. I also declare that this work is the result of my own investigations except where otherwise identified by references and that I have not plagiarised the work of others".

Nikunj Satishchandra Naik _____ Date _____

(Student)

Professor Susan Seal _____ Date _____

(First supervisor)

Professor John Colvin _____ Date _____

(Second supervisor)

ACKNOWLEDGEMENTS

Undertaking this research has been a truly life-changing experience for me and it would not have been possible to do without the support and guidance that I received at every step of this journey.

I would like to express my sincere gratitude to my supervisor Professor Susan Seal for the continuous support and encouragement she gave me throughout the programme. I benefitted greatly from her experience and suggestions while carrying out the research. I would also like to thank my second supervisor, Professor John Colvin for his encouragement and valuable input.

I gratefully appreciate the funding received from Natural Resources Institute (NRI), University of Greenwich for supporting me with tuition fees. Many thanks to NRI for providing facilities and support during my research.

I greatly appreciate the computing support received through the collaboration between the European Bioinformatics Institute and NRI. Thank you to Dr. Paul Kersey for making this possible.

I am indebted to all my friends and fellow researchers for their help and support during this work. Their encouragement and mental support was imperative for completion of this work.

I would like to thank my mom, dad and all family members for their belief in me, although my mom is not with us but her blessings keep me motivated to finish this work. This journey would have not been possible without the support of my wife Zankhana, her love, faith and encouragement during this work helped a lot to finish what I started four years ago.

ABSTRACT

Bemisia tabaci (*B. tabaci*), a major crop pest of important food security crops, transmits more than 100 plant viruses. Despite its global importance as one of the world's top 100 invasive species, limited genomic resources are available in public domains. In this study, transcriptomic datasets from cassava and non-cassava colonizing *B. tabaci* populations were used to compare sequence divergence between populations.

We sequenced transcriptomes of three cassava colonizing populations and three populations which do not colonize cassava as a host plant to generate a large catalogue of transcripts for comparative analysis. Twenty four *de novo* assemblies using multi k-mer values were produced using four transcriptome assemblers, Trinity, Velvet/Oases, SOAPdenovo-Trans and CLC Bio to maximize the diversity and completeness of assembled transcripts. Trinity assemblies were chosen based on evaluation statistics using Transrate, DETONATE, BUSCO and CRB-BLAST. The resulting clustered assembly consisted of 185,895-287,559 contigs, ~38% (for non-cassava) and ~34% (for cassava populations) of which could be functionally annotated when compared against NCBI non redundant database using BLASTX.

The assembled transcriptome data will serve as a genomic resource for future *B. tabaci* studies. Furthermore, our results also showed the performance of publicly available transcriptome assemblers as well as important factors affecting *de novo* assembly.

LIST OF ABBREVIATIONS

The following table describes the abbreviations and acronyms used in this thesis.

Abbreviation	Explanation
BGMV	*Bean golden mosaic virus*
BLAST	Basic local alignment search tool
BUSCO	Benchmarking universal single-copy orthologs
CA	Clustered assembly
CBSD	Cassava brown streak disease
CBSV	*Cassava brown streak virus*
cDNA	Complementary deoxyribonucleic acid
CLCrV	*Cotton leaf crumple virus*
CMB	Cassava mosaic begomovirus
CMD	Cassava mosaic disease
CRB-BLAST	Conditional reciprocal basic alignment search tool
DETONATE	*De novo* transcriptome RNA-seq assembly with or without the truth evaluation
dNTP	Deoxyribonucleotide triphosphate
EC	Enzyme code
EST	Expressed sequence tag
FAO	Food and Agriculture Organization
GO	Gene ontology
HMM	Hidden Markov model
KEGG	Kyoto encyclopaedia of genes and genomes
MEAM1	Middle East-Asia Minor 1
MED	Mediterranean
mtCOI	Mitochondrial cytochrome oxidase subunit one
NGS	Next generation sequencing
ORF	Open reading frame
PCG	Protein coding gene
PCR	Polymerase chain reaction
rRNA	Ribosomal ribonucleic acid
RSEM-EVAL	RNA-seq by expectation-maximization
SLCV	*Squash leaf curl virus*
SMRT	Single molecule real time
SSR	Simple sequence repeats
TLCV	*Tobacco leaf curl virus*
ToMoV	*Tomato mottle virus*
TYLCV	*Tomato yellow leaf curl virus*
UCBSV	*Uganda cassava brown streak virus*

CONTENTS

FIGURES

TABLES

Chapter 1: Introduction

Cassava is one of the most important crops of the tropical world and is consumed by more than 500 million people (Dutt *et al*., 2005). It is important for its high calorie content, low production cost and its ability to adapt to most soil and environmental conditions (Herrera Campo *et al*., 2011). Given these characteristics, cassava holds significant promise for improving food security in tropical regions where climate, soils and societal stresses constrain production (Bellotti and Arias, 2001). Major factors responsible for cassava production losses are pests and diseases that have emerged in the past few decades, and have caused multi-billion-dollar crop losses.

The whitefly, *Bemisia tabaci* (Gennadius) (Hemiptera: Aleyrodidae), is an important pest of various crops, weeds and ornamental plants at all growth stages. The pest causes direct damage by feeding upon the phloem sap and indirect damage by honey dew secretion and vectoring plant viruses (Martin 1999). It causes damage to more than 600 plants, resulting sometimes in complete yield loss in certain crops. *B. tabaci* is responsible for substantial damage to crops by transmitting diseases like cassava mosaic disease (CMD) and cassava brown streak disease (CBSD) (Herrera Campo *et al*., 2011). Cassava mosaic and cassava brown streak diseases are two important constraints to cassava production. Economic damage from CMD has been substantial causing an average of 50% yield loss in infected areas in both Africa and India where cassava plays an important role in food security. CBSD has caused annual losses of $50 million with up to 80% yield loss in East Africa alone affecting more than 20 million people (Herrera Campo *et al*., 2011).

The main objective of this study was to perform comparative bioinformatic analyses on cassava and non-cassava populations of *B. tabaci* to identify differences that could be related to the ability of the certain *B. tabaci* populations to feed on cassava. This research involved transcriptome analysis of different *B. tabaci* populations using a sequence comparison approach to understand the roles of potential genes involved in cassava utilization.

Chapter 2: Literature review

2.1 Cassava

Cassava, *Manihot esculenta* Crantz, is the only species in the genus *Manihot* (*Euphorbiaceae*) grown as a food crop (Fauquet and Fargette, 1990; Blagbrough *et al*., 2010). Cassava is an important food security crop in the tropics and subtropics, where it is cultivated mainly for human consumption (Ariyo *et al*., 2006; Herrera Campo *et al*., 2011). Cassava is considered as an important source of dietary food energy for many African countries where food production is the main concern to match the population growth (Legg *et al*., 2011).

2.1.1 Origin

Cassava is widely grown in the tropics and sub-tropics and believed to have originated from South America, but was introduced to Africa by the Portuguese in the late 16th century (Ariyo *et al*., 2006). Later in the 18th century, cassava was introduced to the east coast of Africa and the Indian Ocean islands of Madagascar, Reunion and Zanzibar, and in India and Sri Lanka in the mid-18th century (Fauquet and Fargette, 1990).

2.1.2 Importance of cassava

Cassava is the main carbohydrate supply and source of income for many African countries where an increase in production is important to meet the demand as a human food (Wydra and Verdier, 2002). It is also considered an important food security crop, able to be grow under marginal environmental conditions and low soil fertility (Legg 1999). Cassava has several advantages over other food crops, such as rice and maize, as it is easy to grow and yields well in poor conditions. Because of the perceived agriculture advantages of cassava production and its role in food security and biofuel production, a threat to cassava production is of serious concern in developing countries (Cardoso *et al*., 2005).

2.1.3 Cassava utilization

Cassava is grown primarily for its ability to accumulate and store starch within large swollen root structures (Taylor *et al*., 2004). It is used for on-farm consumption, as well as a cash crop for smallholder farmers and for commercial operations on large scale farms to feed livestock and also for processing into starch used in food and chemical industries (Taylor *et al*., 2004). In food processing, cassava starch is used mainly for paper, textile, and adhesive

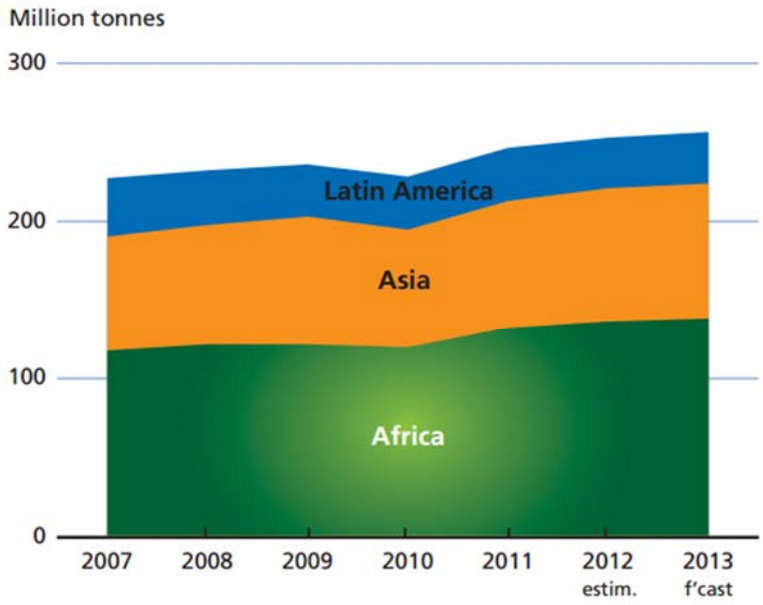
manufacturing and in the oil drilling industry. It is also an important agent for many derived sugar products like glucose, maltodextrins and mannitol (Nassar and Ortiz, 2007).

Within the past few decades, an increased interest in the potential of bio-ethanol to replace conventional fossil fuels has stimulated research into cassava as a possible source of energy. Cassava is increasingly an attractive energy crop due to its high CO_2 fixation ability, high water-use efficiency, high carbohydrate content, and greater starch: ethanol conversion ratio compared to other crops (Kristensen *et al*., 2014).

2.1.4 Cassava production

Global cassava production has roughly doubled in the past 30 years from 124 million to 252 million tonnes in 2012 of which over 130 million tonnes was grown in Africa with about 84 million tonnes in Asia and about 31 million tonnes in Latin America (FAO, 2013). Cassava's importance has dramatically changed between 1980 and 2011, due to the global harvested area expanding from 13.6 million to 19.6 million hectares with an increased industrial demand for cassava in East and Southeast Asia, especially for ethanol and in sub-Saharan Africa as an important food security crop (FAO, 2013).

In recent years cassava production in sub-Saharan Africa has increased from 118 million tonnes in 2010 to over 130 million tonnes in 2012 (FAO, 2013). The majority of this increase has occurred in Nigeria, Africa's largest producer, which grew over 54 million tonnes in 2012 compared to only 42 million tonnes in 2010. Among the other sub-Saharan producers, Ghana's output also increased (~4%) with ~15 million tonnes in 2013. In Asia, cassava production is set to increase to over 85 million tonnes in 2013 due to industrial utilization of cassava in form of alcohol and ethanol (FAO, 2013).



World production of cassava (FAOSTAT, 2013)

Figure 2.1: FAO biannual report on global food markets (November 2013), accessed on 17th April 2014

2.1.5 Cassava pests and diseases

Cassava production is constrained by many pests and diseases. The major pests affecting cassava production are cassava green mite, cassava whitefly (*B. tabaci*) and cassava mealybug (Night *et al*., 2011). The important virus diseases affecting cassava production are CMD and CBSD. CMD can be caused by many distinct species and strains of begomoviruses (genus *Begomovirus*, family *Geminiviridae*), and these viruses are transmitted through infected cuttings or by the whitefly vector *B. tabaci* (Ariyo *et al*., 2006). CBSD is caused by cassava brown streak viruses, of which two different species have been identified namely *Cassava brown streak virus* (CBSV) and *Uganda cassava brown streak virus* (UCBSV). Viruses causing CMD and CBSD are transmitted through stem cuttings and also by the whitefly vector, *B. tabaci*.

2.2 Whitefly *Bemisia tabaci*

Bemisia tabaci is the most economically important species complex of interest out of over 1200 known whitefly species in the world (Mound & Halsey, 1978). The *B. tabaci* complex represents a genetically diverse yet morphologically indistinguishable pests of many ornamental and greenhouse crops (Perring *et al*., 1993; Boykin *et al*., 2007).

Members of the *B. tabaci* species complex are known for their ability to cause damage by direct feeding on phloem sap. They extract large amounts of sap from host plants and this plays a major role in honey dew excretion onto the surface of leaves and fruit, which results in sooty mould development (Brunt, 1986; Cohen *et al*., 1988; Brown and Bird, 1992; Boykin *et al*., 2007; Herrera Campo *et al*., 2011). Sooty mould growth on honey dew reduces photosynthetic capabilities and also affects plant quality, which makes produce have a poor market value (Byrne and Bellows, 1991). The polyphagous nature of the *B. tabaci* complex has been observed on more than 600 different plant species (Mound and Halsey, 1978) including some of the important food and industrial crops such as tobacco, tomato, peppers, squash, cucumber, beans, pumpkin, dew flower, watermelon, cabbage, sweetpotato, peanut, soybean, eggplant, okra and cotton.

B. tabaci is also capable of acting as a vector for over 100 plant viruses especially geminiviruses (genus *Begomovirus*) in tropical and subtropical regions of the world (Liu *et al*., 2013). It has been the subject of significant interest during the past 30 years by transmitting plant viruses such as cassava mosaic begomoviruses (CMBs), cassava brown

streak viruses (CBSVs), *Tomato mottle virus* (ToMoV) and *Tomato yellow leaf curl virus* (TYLCVs), which have produced devastating results in their respective hosts (Mugerwa *et al*., 2012).

B. tabaci has gained worldwide exposure due to yield losses in global food production estimated to exceed US$5 billion annually (Czosnek and Brown, 2010). The worldwide importance of *B. tabaci* is mainly supported by the nature of host adaptability within new geographical environments, current status of resistance to a range of pesticides and increased global trade of plant materials (Basit *et al*., 2013).

2.2.1 History

B. tabaci was first described in Greece in 1889 as a tobacco pest and named as the tobacco whitefly (Gennadius, 1889). In 1897, the first New World *B. tabaci* were observed on sweet potato which was originally described as *Aleyrodes inconspicua* Quaintance and was known as the sweet potato whitefly (Quaintance, 1900). Major *B. tabaci* occurrences were reported on cotton between 1920 and 1930 in India (Horowitz, 1986). Subsequently, major outbreaks occurred, in Sudan and Iran in the 1950s, El Salvador (1961), Mexico (1962), Brazil (1968), Turkey (1974), Israel (1976), Thailand (1978) and Arizona and California (1981) (Horowitz, 1986). Outbreaks in soybean took place in Brazil in 1972-73 (Kogan and Turnipseed, 1987) and in Indonesia in 1981-82 (Samudra and Naito, 1991). Major *B. tabaci* invasions were recorded in Florida on ornamental plants, especially poinsettia in 1986 (Price *et al*., 1987). They were also reported from northern European countries, including the United Kingdom, Netherlands, France and Germany when a highly destructive strain was imported on ornamentals in 1987 (Cheek and MacDonald, 1993). The species is now globally distributed and its presence is reported commonly in Africa, southern Europe, the Middle East, Asia, Australia, Pacific and Americas (Hu *et al*., 2011).

2.2.2 Biotypes

The classification of whiteflies has long been a topic of debate with the gross morphology of all adult *B. tabaci* populations showing similar characteristics. Due to these indistinguishable morphological characteristics, a range of other techniques have been used to classify different populations within the *B. tabaci* complex. Biotypes of *B. tabaci* were first characterised in 1950s by biochemical variations within whiteflies, their host ranges, ability to adapt to different geographical environments and capability to transmit different

plant viruses (Bird, 1957). Through analysis based on esterase profiles, 19 distinct *B. tabaci* biotypes were identified and named as letter codes A to S (Bedford *et al.*, 1992; Brown *et al.*, 1995a; Banks *et al.*, 1999).

2.2.3 Genetic classification methods

The classification of *B. tabaci* populations has also been based on phylogenetic analysis of sequences of the 18S rRNA gene, mitochondrial cytochrome oxidase subunit one (mtCOI) gene and nuclear DNA ribosomal ITS1 regions (Chowda-Reddy *et al*., 2012). The species clusters identified by Dinsdale *et al*. (2010) based on phylogenetic analysis and pairwise comparisons (including 79 whitefly samples taken from cassava growing areas of Kenya, Tanzania, and Uganda) were named as Mediterranean (Q, J, L, sub-Saharan Africa silverleaf); Middle East–Asia Minor 1 (B, B2); Middle East–Asia Minor 2; Indian Ocean (MS); Asia I (H, M, NA); Australia/Indonesia; Australia (AN); China 1 (ZHJ3); China 2; Asia II 1 (K, P, ZHJ2); Asia II 2; Asia II 3 (ZHJ1); Asia II 4; Asia II 5 (G); Asia II 6; Asia II 7 (Cv); Asia II 8; Italy (T); sub-Saharan Africa 1; sub-Saharan Africa 2 (S); sub-Saharan Africa 3; sub-Saharan Africa 4; New World (A, C, D, F, Jatropha, N, R, Sida); and Uganda.

Based on the method proposed by Dinsdale *et al*. (2010), Hu *et al*. (2011) added four new species clusters to *B. tabaci* cryptic complex based on phylogenetic analysis of samples with > 3.5% divergence with respect to their partial mtCOI gene sequence. The new clusters were named as Asia II 9, Asia II 10, Asia III and China 3 (Hu *et al*. 2011).

2.2.4 Economic importance of the *B. tabaci* species and the viruses they transmit

A major outbreak caused by viruses transmitted by *B. tabaci* was reported in the form of cassava mosaic disease (CMD) in Uganda (1990s) and this resulted in up to 100% yield loss to cassava crops (Otim-Nape *et al.*, 1997). Subsequent results of CMDs were reported in East and Central Africa, showing major crop losses in cassava which affected over 200 million people (Legg *et al*., 2006).

Tomato yellow leaf curl virus (TYLCV) was first discovered in Israel on tomato crops. TYLCV mainly discovered in parts of Europe and in eastern Mediterranean countries and parts of Africa, Asia, Australia, and the Caribbean where losses were up to 100% (Cohen and Harpaz, 1964). Another type of tomato virus spread by *B. tabaci* was first observed in 1989, known as *Tomato mottle virus* (ToMoV). The virus was detected in regions of Florida

in the 1990 to 1991, with estimated losses of $125 million on tomato crops (Polston *et al*., 1993). Several other plant viruses transmitted by *B. tabaci* such as *Tobacco leaf curl virus* (TLCV), *Squash leaf curl virus* (SLCV), *Cotton leaf crumple virus* (CLCrV) and *Bean golden mosaic virus* (BGMV) cause heavy yield losses to host plants (Bedford *et al*., 1994).

2.3 Next Generation sequencing technologies

2.3.1 Overview

Determining the genomic information of any living organism is the most crucial step in molecular biology for understanding evolutionary, functional and structural relationships (Shokralla *et al*., 2012). DNA sequencing technologies has played an important role in understanding biological phenomena (Liu *et al*., 2012). Over the past two decades, advances in traditional 'Sanger' DNA sequencing have altered genomics research and allowed researchers to conduct experiments that were previously not possible.

The next-generation sequencing (NGS) technologies introduced in 2005 are high throughput in nature compared to Sanger DNA sequencing and are more affordable. NGS technologies can sequence several human genomes in a single run within days compared to Sanger sequencing, which originally took more than 10 years and US$2.7 billion to complete (Berglund *et al*., 2011). NGS has the potential to accelerate biological research by creating new opportunities in genome sequencing projects including *de novo* genome sequencing, resequencing genomes for improvements, mRNA profiling, and can be applied to a broad range of applications such as molecular cloning, breeding, finding pathogenic genes, and comparative and evolution studies (Zhou *et al*., 2010; Liu *et al*., 2012). New sequencing technologies have also introduced new areas of genomics research by pairing with computational, statistical and mathematical programs and algorithms to analyse vast sequencing datasets.

2.3.2 History and Fundamentals of NGS technologies

The traditional DNA sequencing method was first introduced by Sanger *et al.* (1977). The DNA sequencing technology developed by Frederick Sanger was based on a chain-termination method (known as Sanger sequencing) and was adopted as a "first generation" laboratory and commercial sequencing technology due to its high efficiency and low radioactivity. A decade later, the company Applied Biosystems introduced the first automated sequencing machine which incorporated advances and improvements made to the

Sanger sequencing technology; the machine, was based on capillary electrophoresis which was much faster and more accurate than previous versions (Smith *et al*., 1986).

A series of other next-generation sequencing machines including Roche 454, Applied Biosystems SOLiD system, Illumina genome analyser, Ion Torrent, HeliScope, SMART DNA sequencer have been commercially introduced in recent years based on their sequencing and detection techniques. These machines possess the ability to generate tens of millions of sequencing reads in parallel (Shokralla *et al*., 2012).

NGS technologies do not require a conventional cloning based procedure to amplify and separate DNA samples and can be used directly from a pool of cDNA library fragments generated through reverse transcription or PCR amplified molecules (Mardis, 2008). These technologies have faced many challenges since their first launch in 2005.

2.3.3 Roche 454 genome sequencer

Roche 454 was the first next-generation genome sequencing system introduced commercially by 454 Life Sciences in 2005 based on sequencing-by-synthesis pyrosequencing technology. The 454 pyrosequencing uses a pyrophosphate detection technology released during nucleotide incorporation by DNA polymerase. This reaction produces light for each incorporated base by the action of enzyme luciferase which is proportional to the number of nucleotides incorporated (Shokralla *et al*., 2012). Single dNTPs are added one by one to the reaction and the sequencing-by-synthesis pyrosequencing process repeated (Liu *et al*., 2012).

The initial sequencing output produced 100-150 bp in terms of average read lengths with more than 200,000 reads making the file size about 20 Mb per run in 2005. Roche 454 launched a new improved system in 2008 called the 454 GS FLX Titanium (Liu *et al*., 2012). The main features of this system were that it could produce up to 700 bp read lengths with 99.9% accuracy. The advantages of Roche systems are that they only take about 10 hours to complete each sequencing run, with read lengths (up to 700 bp) and do not require manpower for library construction. The major limitation with these systems are that the reagent cost is high compared to other NGS systems and it produces errors for poly-bases longer than 6 bp (Liu *et al*., 2012).

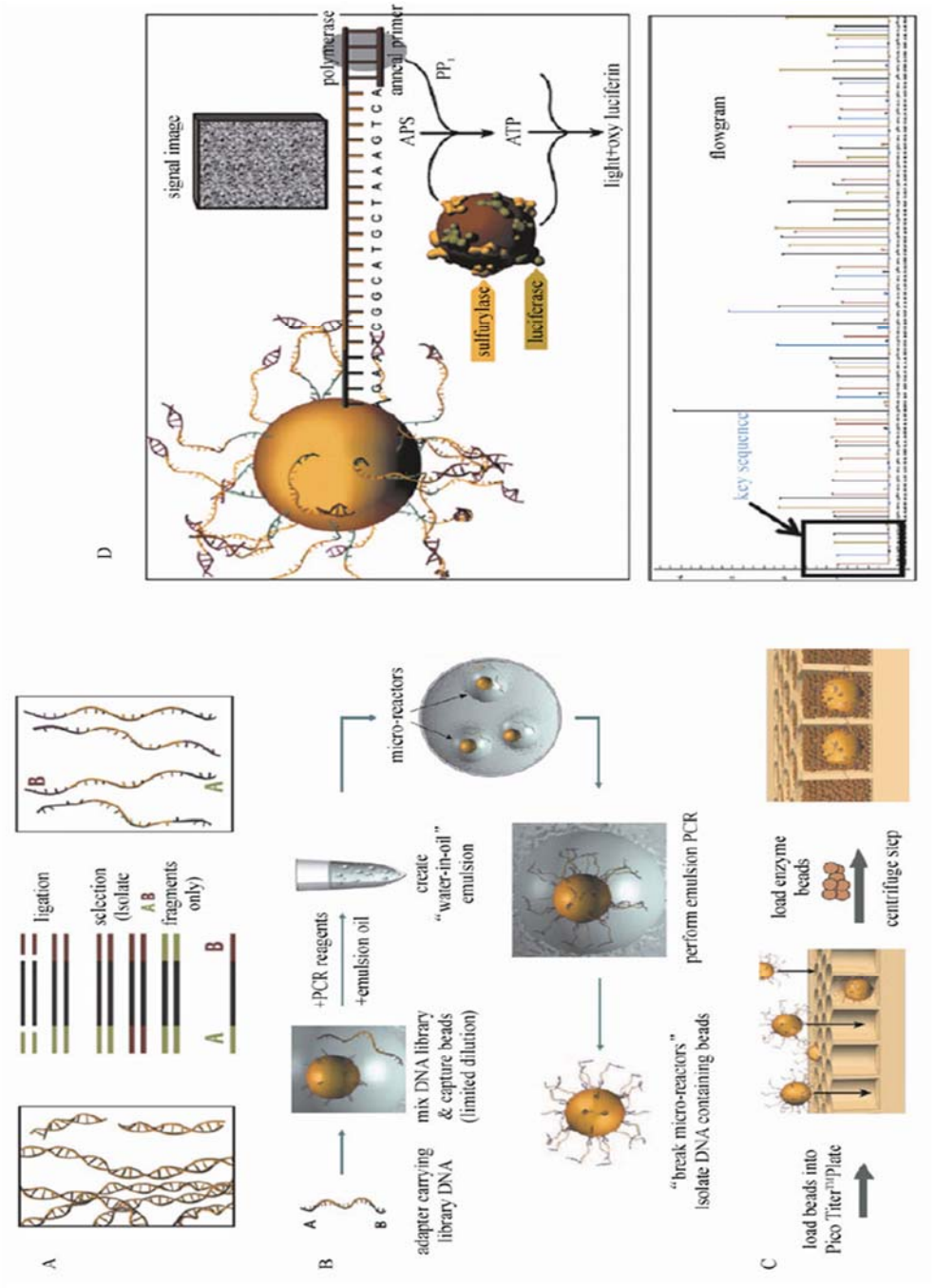


Figure 2.2: The working principle of Roche 454 system. (A) Prepare DNA fragments using adapter. (B) Emulsion based PCR amplification. (C) Load amplified beads into PicoTiter^{TM} plate. (D) Sequencing-by-synthesis reaction and imaging. (http://www.454.com)

2.3.4 Applied Biosystems SOLiD system

Applied Biosystems introduced the SOLiD system in 2007 based on sequencing-by-ligation technology. In this technique, adapter-linked DNA fragments are immobilized with complementary oligos on to the surface of 1 mm magnetic beads. These DNA fragments are mixed with emulsion oil which creates a water-in-oil emulsion that are individually amplified using emulsion PCR containing amplification reagents and capture beads. The amplified beads are attached to a glass slide and then placed inside a fluidics cassette of the sequencer (Shokralla *et al*., 2012). Before the sequencing cycle, a primer is hybridised to the adapter in the first step and then a mixture of oligonucleotide octamers are also hybridised to the ligation mixture. The fluorescent labelled octamer detects corresponding four oligo bases followed by deletion of fluorescent oligos and the ligation cycle is repeated (Zhou *et al*., 2010).

Each step of octamer ligation determines five bases (Zhou *et al.*, 2010). The continued sequencing process can generate more base pairs similarly and different base positions can be identified using different size primers. The current read length of the SOLiD system is about 35 bases and because of the nature of the process, a low error rate can be achieved. The SOLiD system is capable of detecting more than 50 million bead clusters and therefore it can produce gigabytes of data per run (Ansorge, 2009).

In late 2010, the SOLiD 5500xl system was launched by Applied Biosystems which has up to 250 Gb sequencing capacity with improved read lengths (85 bp), accuracy (99.99%) and speed (complete run within 7 days). The major drawback of this system was its short read length compared to other NGS systems, which generate read lengths in the region of 700~900 bp (Liu *et al*., 2012).

2.3.5 The Illumina (Solexa) Genome Analyser / HiSeq system

The Genome analyser was commercialised in 2006 by Solexa, and the company was later purchased by Illumina in 2007 (Liu *et al*., 2012). The Illumina system adopted the sequencing by synthesis approach. Illumina genome analysers use three steps for sequencing. In the first step, DNA is fragmented in to hundreds of millions of copies and sheared ends are repaired and ligated with oligo adapters at both ends. In the second step, an 8-channel genome analyser flow cell is used for cluster generation (Buermans and Dunnen, 2014). This flow cell contains densely distributed mono oligos for high throughput

sequencing. These oligos bind to adapters which are ligated DNA fragments which undergo amplification and generate millions of clusters. In the last step, each cluster is supplied with polymerase and fluorescent labelled nucleotides with their 3' ends inactivated. This specialized method of blocking the 3' ends ensures single base pair sequencing. Each incorporated nucleotide is identified using an imaging technology. The fluorescent labelled 3' ends are then removed and the sequencing by synthesis cycle is repeated (Zhou *et al*., 2010).

Four versions of Illumina sequencers are available in the market today known as Genome analyzer IIx, HiSeq 2000, HiSeq 1000 and MiSeq. Due to their high throughput outputs, up to 600 Gb (HiSeq 2000), 300 Gb (HiSeq 1000), 95 Gb (Genome analyzer IIx) and 1.5 to 2 Gb (MiSeq) makes them popular amongst sequencing projects. Illumina introduced an upgraded version of HiSeq 2000 in 2012 and this was named HiSeq 2500. This new system can produce about 120 Gb of data within 27 hours, and makes the system capable of sequencing whole genomes in 24 hours (Shokralla *et al*., 2012).

2.3.6 Ion Torrent

Life Technologies introduced a real-time detection system in 2010 to create a direct connection between the chemical and digital methodology to enable fast, accurate and high throughput sequencing. The system was based on Watson's nucleic acid chemistry and named as PostLight sequencing technology (Shokralla *et al*., 2012). The Ion Torrent works on the principle of a biochemical reaction that occurs during nucleotide incorporation into a DNA strand using polymerase action, resulting in the release of a hydrogen ion as a by-product of the reaction. The system uses high-density arrays of micro-machine wells to run biochemical process in parallel. Each well on the plate contains a library prepared DNA template and ion-sensitive layer at the base. The unique ion sensor technology detects charges from hydrogen ions during nucleotide incorporation without any other imaging or lighting technology (Buermans and Dunnen, 2014).

The ion torrent sequencing system uses three different ion chips 314 (10 Mb); 316 (100 Mb); and 318 (1 Gb) depending on sequence data required. Recently, Life Technologies added another advanced version; the Ion Proton bench top sequencer. The most important aspect of Ion Proton is that it can sequence human genome within few hours which took much more time to complete with other sequencing systems (Shokralla *et al*., 2012).

2.3.7 HeliScope

HeliScope was the first commercial single-molecule sequencing system introduced by Helicos Biosciences in 2008 (Shokralla *et al*., 2012). It is based on sequencing-by-synthesis on a single DNA molecule using DNA polymerase and fluorescently labelled nucleotides to identify incorporated bases. Specially designed fluorescent nucleotides stop the sequencing process until the added nucleotide's presence is detected and captured by a highly sensitive CCD camera attached to a microscope. Unincorporated nucleotides are then washed away with by-products remaining from the previous cycle followed by a chemical procedure to remove fluorescent labels from the extended DNA strand. Repeated cycles of single-base extension, label-cleaving and detection procedure can generate up to 1 billion sequence reads (Shokralla *et al*., 2012).

2.3.8 SMART DNA sequencer

Pacific Biosciences introduced their single-molecule real-time (SMRT) DNA sequencing system in 2010 (Shokralla *et al*., 2012). It is based on a real-time fluorescent based technique that requires no amplification for sample preparation as it involves single molecule sequencing-by-synthesis approach. The SMRT system utilizes a 'zero mode waveguide identification system' made of nano particles to detect DNA polymerase activity. During a sequencing cycle, the DNA polymerase adds complimentary nucleotides to the template single stranded DNA strand. It uses fluorescence labelled phosphor-linked nucleotides with modified terminal phosphate groups which are released during nucleotide incorporation. This sequencing cycle does not require a washing step after each cycle, which improves the nucleotide incorporation frequency as well as sequence quality. The SMRT platform utilizes natural DNA polymerase property to incorporate 10 or more nucleotides per second in parallel (Shokralla *et al*., 2012).

Chapter 3: Evaluation of *de novo* transcriptome assemblies using paired-end RNA-Seq data of the whitefly, *Bemisia tabaci*

3.1 Introduction

High-throughput sequencing has revolutionized research to understand and annotate genetic profiles of model and non-model organisms using RNA sequencing (O'Neil *et al*., 2013). RNA sequencing is a powerful and cost-effective way of studying transcriptome profiles of organisms in the absence of a reference genome (Zhao *et al*., 2011). To study such transcriptomes, many *de novo* transcriptome assembly programs are available for assembling short sequencing reads generated using Illumina and Roche platforms. However, in the absence of a reference transcript set, assembly evaluation can be difficult especially when the transcript expression levels are different or the number of gene isoforms present due to alternative splicing (Zhao *et al*., 2011). In addition to the complexity in assembling transcripts, many *de novo* assembly programs take a different approach for assembling raw sequencing reads, as well as provide multiple input parameters to allow variations in transcriptome assembly generation using the same data set (Smith-Unna *et al*., 2015). Commonly used quality parameters for evaluating such assemblies are contig length distribution, number of contigs, mean contig length (N50) etc. but these are primitive and often do not provide accurate measures of assembly quality (Li *et al*., 2014).

While the transcriptome assembly is difficult, another important factor to consider is sequencing coverage. In genome assembly, sequencing coverage is generally uniform, whereas the transcriptome sequencing coverage is highly variable and depends on gene expression levels (Surget-Groba and Montoya-Burgos, 2010). It is therefore important to understand the overlap length between two sequences and assign a value to consider them as contiguous also referred to as k-mer length (Surget-Groba and Montoya-Burgos, 2010). The optimal k-mer value for any transcriptome assembly depends on the sequencing depth, read quality and complexity of the transcriptome to be assembled. Transcriptome assembly using higher k-mer lengths can recover longer and contiguous fragments while losing poorly expressed transcripts or with lower k-mer lengths to recover poorly expressed transcripts which will result in numerous fragmented transcripts (Surget-Groba and Montoya-Burgos, 2010). Therefore, an approach with various k-mer lengths is highly desirable for *de novo* transcriptome assembly without compromising between these two extremes.

We designed this study to evaluate the transcriptome assembly of three *B. tabaci* populations, Israel B (IB), Israel Q (IQ) and Montpellier Q (MQ), which do not colonize cassava and three cassava colonizing populations, East African cassava Nam (AfC), East African cassava Ssanje (Ss) and West African cassava Nigeria (NigC) using the Illumina HiSeq 2000 platform. Here we used four *de novo* assemblers Trinity (Grabherr *et al*., 2011), SOAPdenovo-Trans (Xie *et al*., 2014), CLC Genomics Workbench (CLC bio) and Velvet (Zerbino and Birney, 2008) followed by Oases (Schulz *et al*., 2012) with different k-mer values to compare and evaluate the assemblies generated for these six *B. tabaci* transcriptome data sets. All *de novo* assemblers used in this study use *de Bruijn* graph for computational and memory efficiency, which in turn produces complex sub-graph structure of connecting nodes. Each node is connected with a series of edges and if this connection overlaps by k-1 nucleotides, then this connection can be considered as possible transcript. As each path represents the possible transcript, a scoring method applied to each graph relies on the original read sequences and the sequencing information to identify best possible transcripts (Grabherr *et al*., 2011).

The Trinity assembler uses a novel error removal model for detecting variations in gene expression levels within each graph to score possible transcripts using dynamic programming while Oases uses a sub-graph structure to estimate all possible transcripts and these are scored based on a heuristic algorithm (Xie *et al*., 2014). SOAPdenovo-Trans incorporates the error removal model of Trinity and the robust heuristic approach of Oases for scoring using its own transitive reduction method to improve scaffolding graphs for more accurate results (Xie *et al*., 2014). The commercial CLC Bio *de novo* assembler was also used to measure the differences between both open source and commercial algorithms which are based on the same *de Bruijn* theory. In this study, we compared six *B. tabaci* transcriptome data sets using primary metrics as well as metrics based on read mapping and annotation to evaluate the assemblies. In addition, we applied multi k-mer strategy to examine how various k-mer length affected assembly outcomes.

3.2 Materials and Methods

3.2.1 cDNA library preparation and Illumina sequencing

Total RNAs were used to construct cDNA libraries for samples IB, IQ, MQ, AfC, Ss and NigC. Sequencing libraries for all six samples were generated using Illumina HiSeq 2000 platform for both ends with read lengths of 100bp.

3.2.2 Quality control

Raw sequencing reads produced using Illumina HiSeq 2000 platform were first assessed using Cutadapt software (Martin, 2011) to remove adaptor contamination and were then quality checked using the FastQC program. The results showed base ambiguity for first ~13bp in all samples which may be due to hexamer contamination in sequencing data and were subsequently trimmed using a custom Practical Extraction and Reporting Language (PERL) script to facilitate sequence assembly.

3.2.3 *De novo* assembly

Cleaned reads were *de novo* assembled using the Trinity (v2.0.6), SOAPdenovo-Trans (release 1.03), CLC Genomics Workbench (v7.4) and Velvet (v1.2.10) followed by Oases (v0.2.8). All the assemblies were performed using the same assembly parameters to keep the same condition for comparing and evaluating results. The parameters used with Trinity were –left (forward reads), --right (reverse reads), --seqType (FASTQ file type), --max_memory (memory used to run the program), --KMER_SIZE (25, 27 and 29) and --SS_lib_type (direction of reads). The CLC de novo assembly was run using k=25 (word size), k=27 and k=29 with automatic bubble size. The parameters used with SOAPdenovo-Trans were avg_ins = 260 (average insert size), reverse_seq = 0, asm_flags = 3, q1 = forward reads and q2 = reverse reads for –K (k-mer values) = 25, 27 and 29. Oases assembly pipeline was used for assembling quality reads using parameters –m 25 (initial k-mer value), -M 29 (last k-mer value), -fastq (input file type), -shortPaired (short sequencing reads), -separate (separate read files) using -ins_length (insert size) 260.

3.2.4 Assembly statistics computation

Assembly statistics for each data set were calculated using reference-free quality assessment tool TransRate (Smith-Unna *et al*., 2015). TransRate tool was run with parameters –assembly (assembly file), --left (forward reads) and --right (reverse reads).

In addition, all assemblies were scored using DETONATE (*DE novo* TranscriptOme rNA-seq Assembly with or without the Truth Evaluation) (Li *et al*., 2014). DETONATE consists of two scoring methods: RSEM-EVAL (RNA-Seq by Expectation-Maximization) and REF-EVAL. RSEM-EVAL is a reference free evaluation method based on assembly and the sequencing reads used for that assembly while REF-EVAL requires a reference to score the assembly. As the genome information for *B. tabaci* is not available, we used RSEM-EVAL to score all assemblies using parameters --paired-end, upstream_read_file (forward reads), downstream_read_file (reverse reads) assembly_fasta_file and L (average fragment size).

3.2.5 Annotation

Assembled transcripts are often compared against protein databases of same or related species using BLAST to assess the assembly completeness. In the absence of full *B. tabaci* genome sequence, reference data set of *Acyrthosiphon pisum* and *Diaphorina citri* were used. Both *A. pisum* and *D. citri* are the members of the Hemiptera Group for which full genome information is available. Assembled sequences for all six samples were annotated against the *A. pisum* protein dataset (AphidBase, release v2.1b) and *D. citri* protein dataset using CRB-BLAST (Conditional Reciprocal BLAST) (Aubry *et al*., 2014) with parameters -query (query fasta file), -target (target fasta file as nucleotide or protein) and -evalue (default 1e-05).

3.2.6 Evaluating *de novo* assembly by re-aligning reads to assembled contigs

To assess the representation of reads in assembled contigs, we used a Trinity bundled script bowtie_PE_separate_then_join.pl using the parameters --target (assembly file), --left (left reads), --right (right reads), --aligner bowtie, --p 8 (number of processors). Bowtie is a short read aligner mainly used to align large sets of short DNA sequences (reads) against a reference creating alignments scored properly and improperly paired alignments based on paired read alignment information such as orientation of pair and insert size.

3.2.7 Assessment of assembly completeness

The TransDecoder tool was used to identify likely coding regions against the set of assembled contigs. TransDecoder identifies long open reading frames (ORFs) and scores them according to their sequence composition. By default, it reports ORFs that are at least 100 amino acids long. The completeness of assembled contigs were also evaluated using the Benchmarking Universal Single-Copy Orthologs (BUSCO) software (Simao *et al*., 2015).

BUSCO uses longest ORFs and Hidden Markov Model (HMM) amino acid profiles to align transcripts against the set of conserved ortholog gene sets. We used 2675 genes from arthropods to assess the completeness of transcriptome.

3.2.8 Redundancy removal and generation of cluster assembly

To generate meta assembly, all transcripts from different k-mer assemblies were clustered and redundancy removal were performed using the CD-HIT-EST (v4.6.4) (Fu *et al*., 2012) program with parameters –n 10, -T 12, -r 1 and –c 1.0. CD-HIT-EST was used to remove transcripts which are entirely covered by other transcripts with 100% identity as the merging different k-mer assemblies will introduce redundancy. The non-redundant transcripts with no less than 200bp were used to produce optimized cluster assembly.

3.3 Results

3.3.1 *B. tabaci* transcriptome data

A total of 46,892,996 quality trimmed reads for IB, 81,867,506 reads for IQ, 82,289,190 reads for MQ, 30,310,262 reads with AfC, 38,438,672 reads for Ss and 43,590,886 reads for NigC were obtained with a read length of 87 bp (Figure 3.1). The reads obtained for all samples were of good quality with >35 Phred quality score (40 is the highest effective score that a single base can receive under normal conditions which indicates that there is a 1 in 10,000 chance that the called base is incorrect) (Ewing and Green, 1998).

3.3.2 Assembly statistics

In the absence of a reference genome for *B. tabaci*, quality trimmed reads were *de novo* assembled using the Trinity, CLC Genomics, SOAPdenovo-Trans and Velvet followed by Oases using recommended transcriptome assembly parameters. We compared each of the assemblies using:

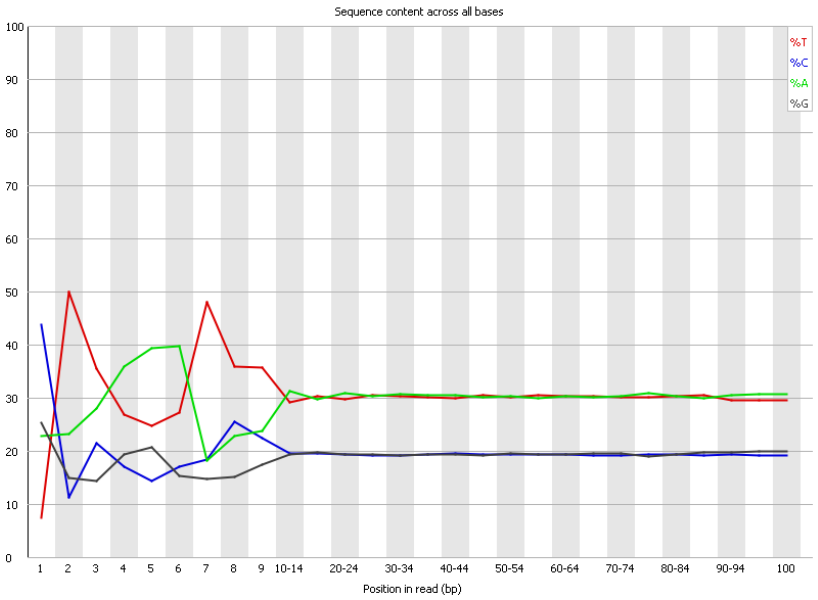
(1) number of contigs assembled
(2) the minimum and maximum length of assembled contig
(3) mean contig length
(4) number of contigs with length >1 kb
(5) N50 length of contigs

In addition to these primary assembly measures, we also compared the percentage of reads mapped along with the proportion of reads mapped to generate good assemblies, assembly score and an optimal assembly score.

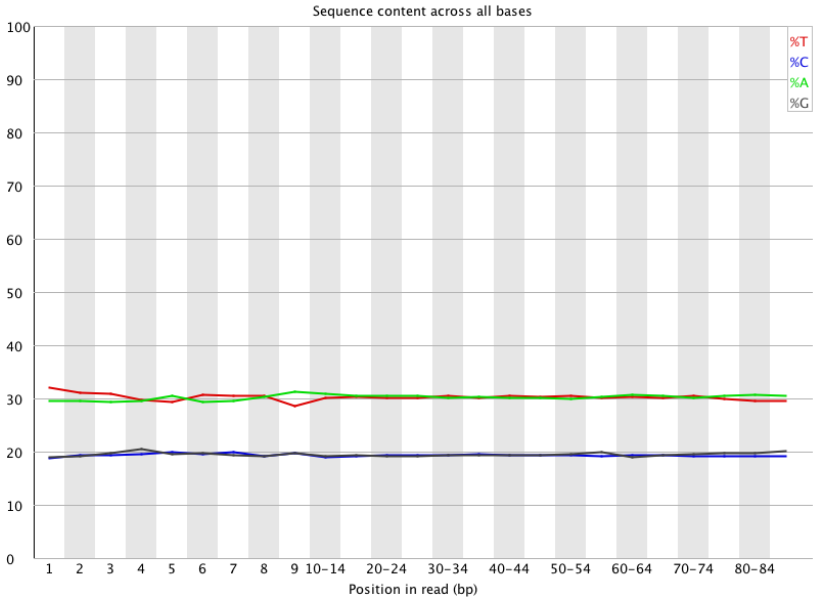
The total number of contigs assembled using k-25, k-27 and k-29 varied across four assemblers (Figure 3.2). The highest number of contigs produced per assembly was 153,888 for IQ k-25 assembly. The contig number steadily decreased as the k-mer value increased from 25 to 29 for assemblies generated using the Trinity and Velvet/Oases, whereas the contig numbers increased for assemblies generated using the CLC and SOAPdenovo-Trans. The shortest contig reported by all assemblies was 75 bp for Velvet/Oases (k-25 Ss assembly). Though a number of contigs with shortest length were assembled in the Velvet/Oases assembly, it assembled the longest contig amongst all four assemblers with length of 36,897 bp for MQ k-25 assembly. The contigs from SOAPdenovo-Trans assembly

were relatively short with an average contig length between 430 bp and 675 bp compared to 1060 bp to 1512 bp from Velvet/Oases assembly (Figure 3). The N50 contig length (Figure 3.4) for SOAPdenovo-Trans and Velvet/Oases assemblies was approximately two times longer than that of Trinity and CLC. Here, N50 is the length of the longest contig where all the contigs of that length compose at least 50% of the bases of the assembly (Li *et al*., 2014).

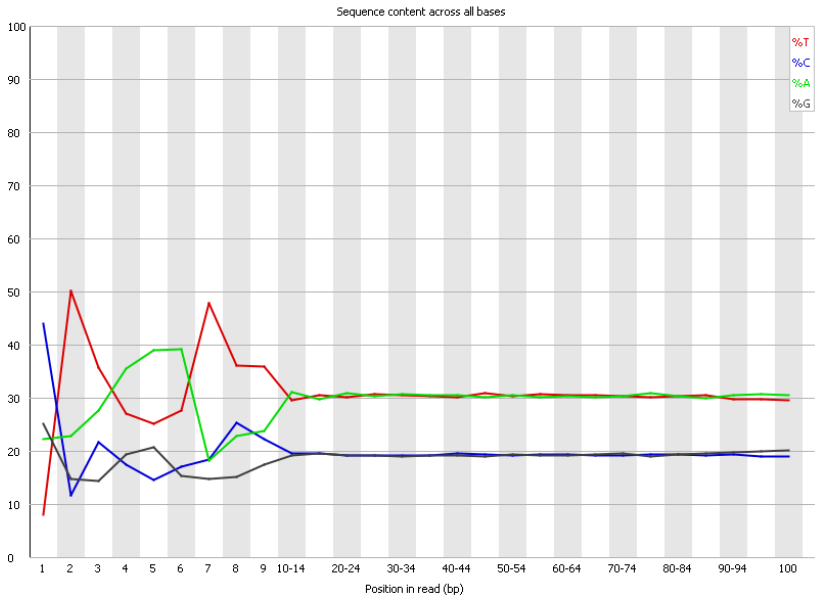
Though a greater number of contigs with length of >200 bp were assembled in the Trinity assemblies, the highest percentage of contigs with length >1 kb were assembled for Velvet/Oases assemblies (Figure 3.5).



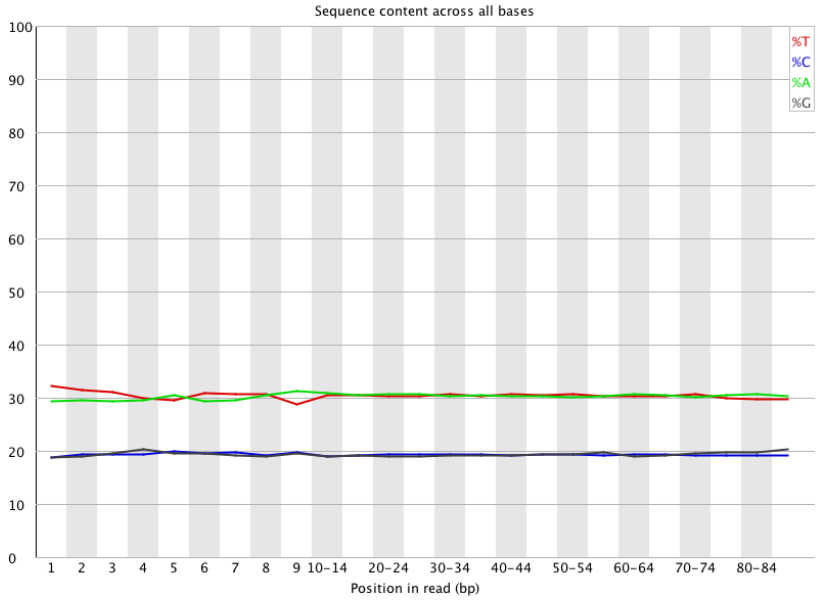
(A) Raw forward primer derived reads



(B) Trimmed forward primer derived reads



(C) Raw reverse primer derived reads



(D) Trimmed raw reverse primer derived reads

Figure 3.1: Quality of sequence reads determined by FastQC showed variation in base calling for first ~13bp and were trimmed to facilitate sequence assembly. A) Result of raw forward sequence reads for sample AfC where Y-axis represents percentage of base calling for each position in read and the X-axis shows position in read (bp). B) The result shows accurate base calling ratio after trimming first ~13bp. C) Result of raw reverse sequence reads for sample AfC. D) Trimmed reverse sequence reads. Similarly, all other samples were processed and trimmed for first ~13bp.

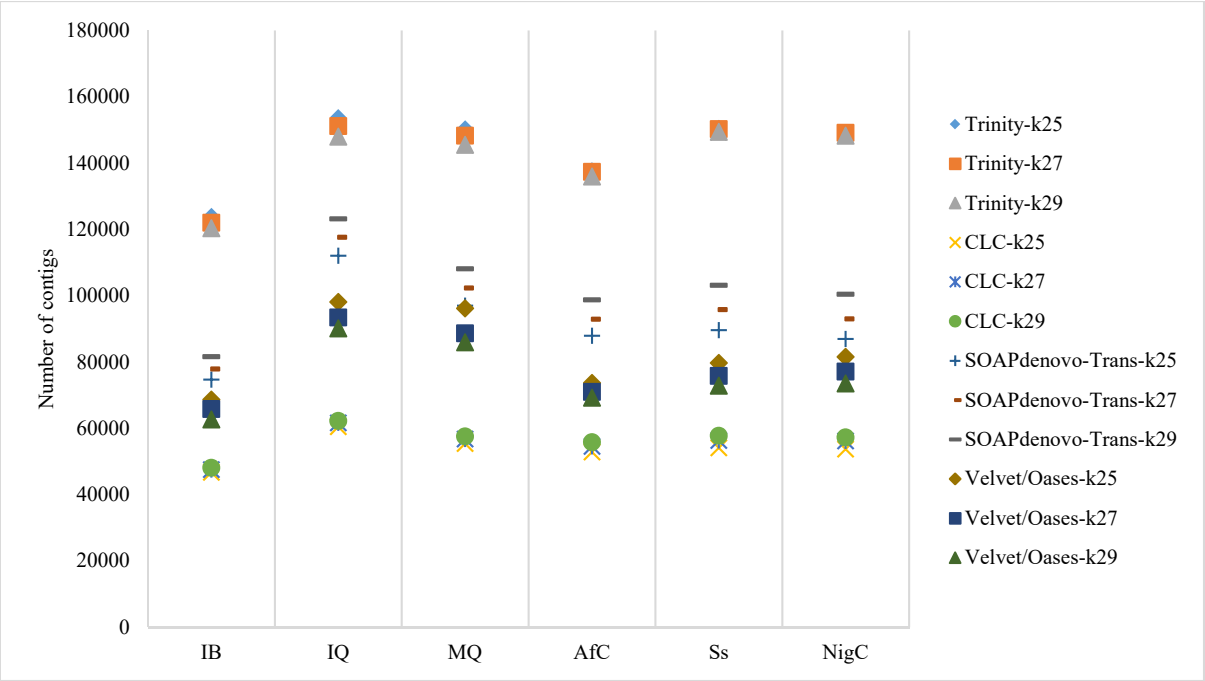


Figure 3.2: Number of contigs assembled using various assemblers and k-mer sizes.

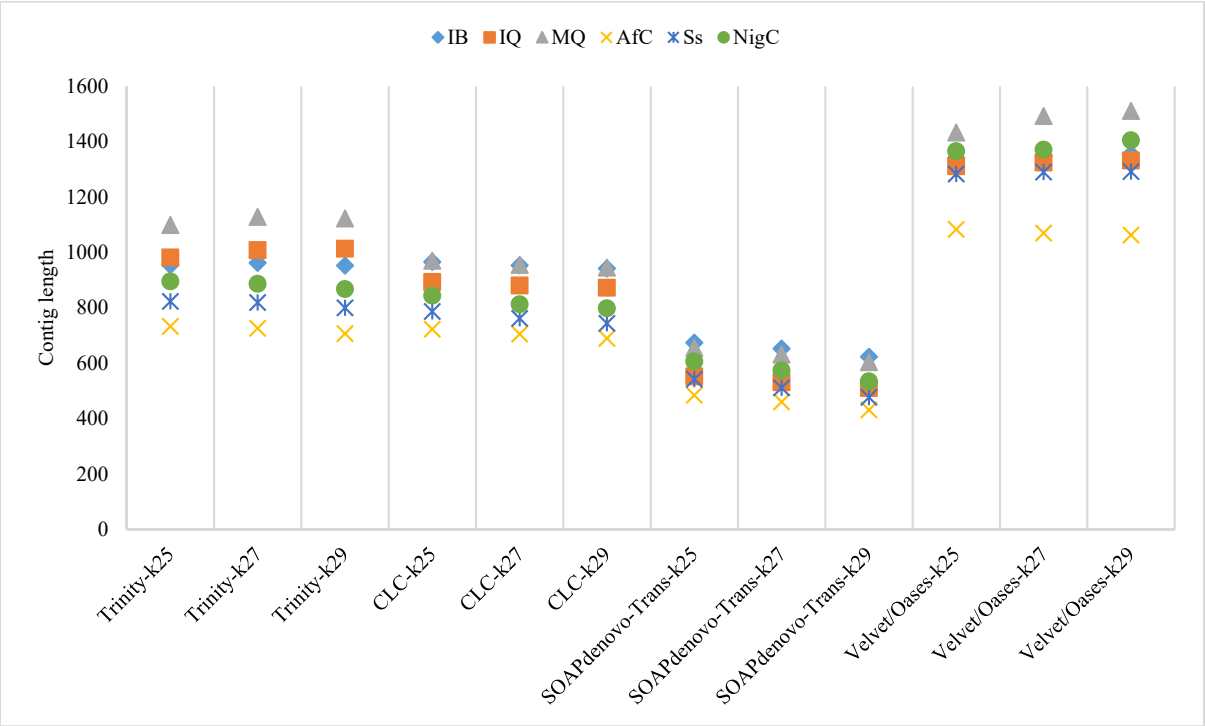


Figure 3.3: Average contig length of contigs assembled using Trinity, CLC, SOAPdenovo-Trans and Velvet/Oases with k-mer size of 25, 27 and 29.

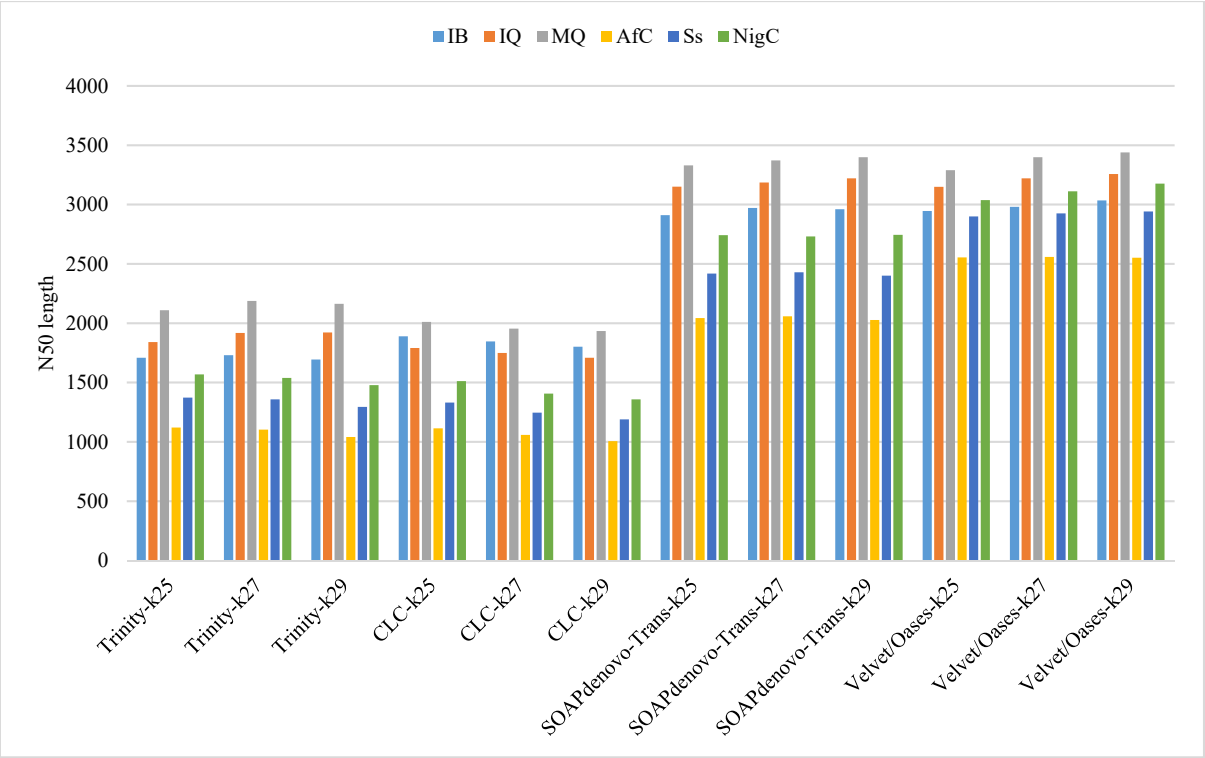


Figure 3.4: N50 values of contigs assembled using Trinity, CLC, SOAPdenovo-Trans and Velvet/Oases with k-mer size of 25, 27 and 29.

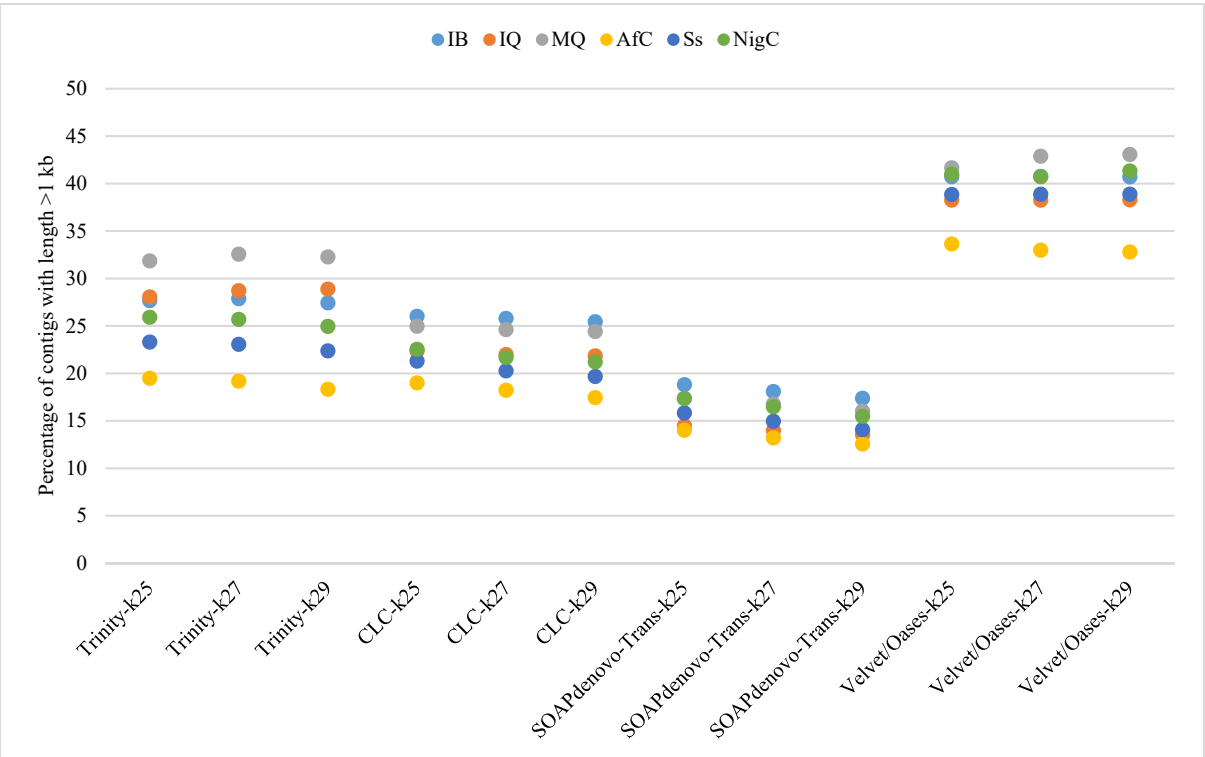


Figure 3.5: Percentage of contigs assembled using Trinity, CLC, SOAPdenovo-Trans and Velvet/Oases with length greater than 1kb.

In addition to above primary quality metrics, all assemblies were scored using TransRate. TransRate uses reads as input to map contigs against reads to report quality scores for contigs and assemblies. TransRate reports two key reference-free statistics: the contig score and the assembly score. The contig score measures accuracy and completeness of assembly for each individual contig that are assembled from the read, whereas the assembly score measures accuracy and completeness of the whole assembly by combining the proportion of the reads used to calculate individual contig scores. An increased assembly score (maximum 1.0) correspond to an assembly that is more biologically accurate. The optimal assembly score (maximum 1.0) can be used to filter out bad contigs based on individual contig scores from an assembly, resulting only well-assembled contigs.

Figure 3.6 shows that the assemblies generated using the Trinity mapped successfully compared to other three assembly programs. However, the percentage of reads mapped to assemblies ranged from 55-85% and thus shows that a significant amount of reads failed to be assembled by all four *de novo* assembly programs. The results also indicate that despite the assemblies generated by Velvet/Oases produced contigs with highest average length and N50 length; only about ~60% of total reads were mapped successfully. The optimal assembly score results in Figure 3.7 also showed that the Trinity outperformed other three *de novo* assembly programs. The assemblies generated using Trinity with k 27 for sample NigC produced the highest assembly score amongst all samples. Unfortunately the CLC assemblies for sample IQ with k-27 and 29 failed due to memory limitation even when running on Linux servers with 24 cores and 1 TB of memory and were not included in Figures 3.6 and 3.7.

All assemblies were also scored using DETONATE to further evaluate the assemblies based on the percentage of reads that mapped successfully. The RSEM-EVAL method of DETONATE also uses reads to score each assembly using RSEM probabilistic model. In addition, RSEM-EVAL also provides a contig score based on how well each contig is supported by reads it is assembled from. As the RSEM-EVAL score is used for evaluating best assembly, the contig score can be used to optimize the assemblies by filtering low scoring contigs from assembly. The results in Figure 3.8 indicated that the Trinity produced most accurate assemblies for k-25, 27 and 29 based on RSEM-EVAL score as the higher scores are better than lower scores. The assemblies produced highest score were generated using k-29. The results also showed that despite the assemblies generated using Velvet/Oases

scored better in contig statistics, it scored worst amongst all four assemblers when compared using RSEM-EVAL score.

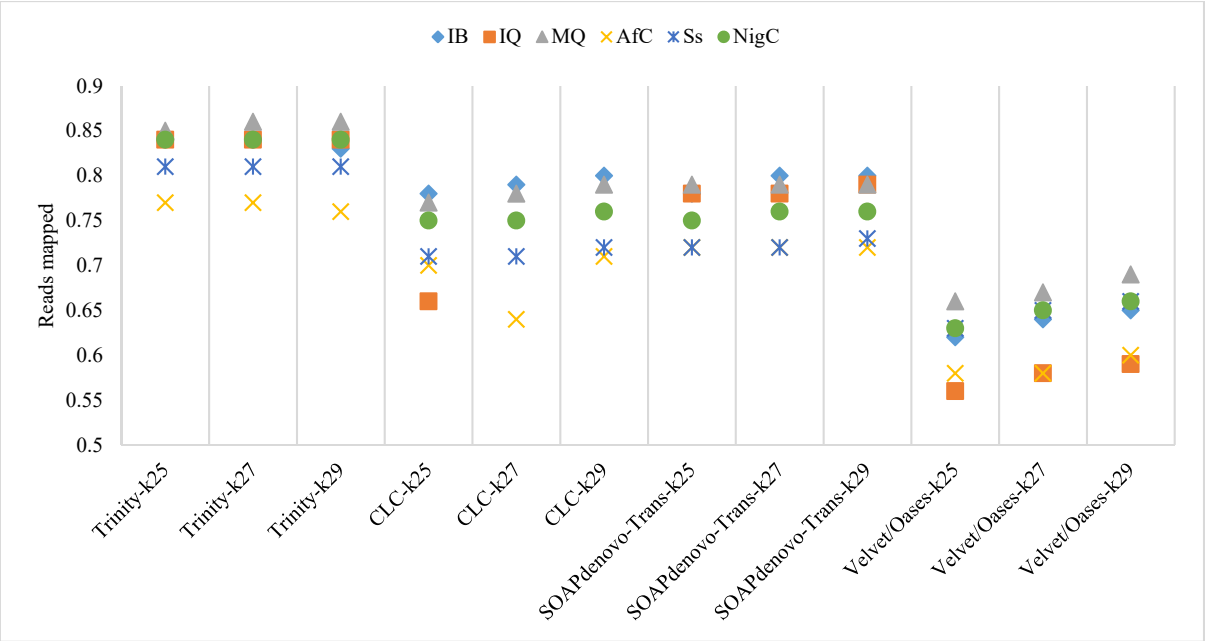


Figure 3.6: Percentage of good mappings where both paired end reads aligned in a correct orientation on the same contig without overlapping either end of the contig.

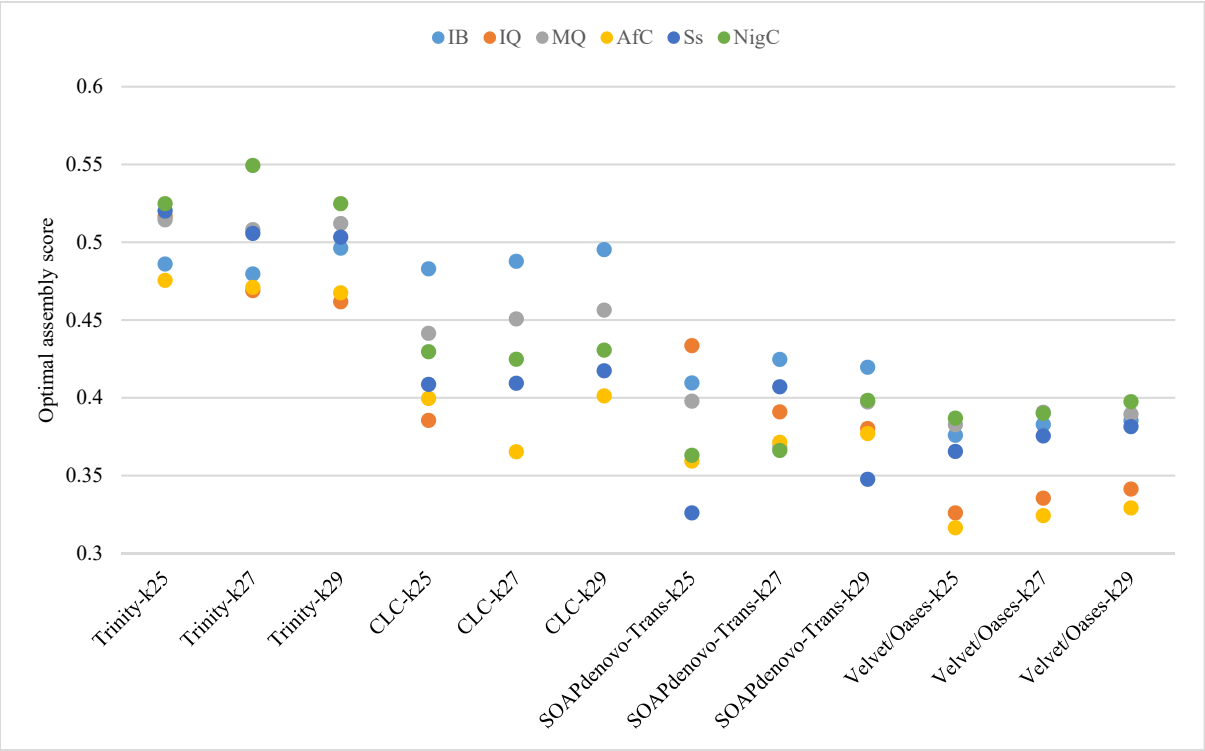
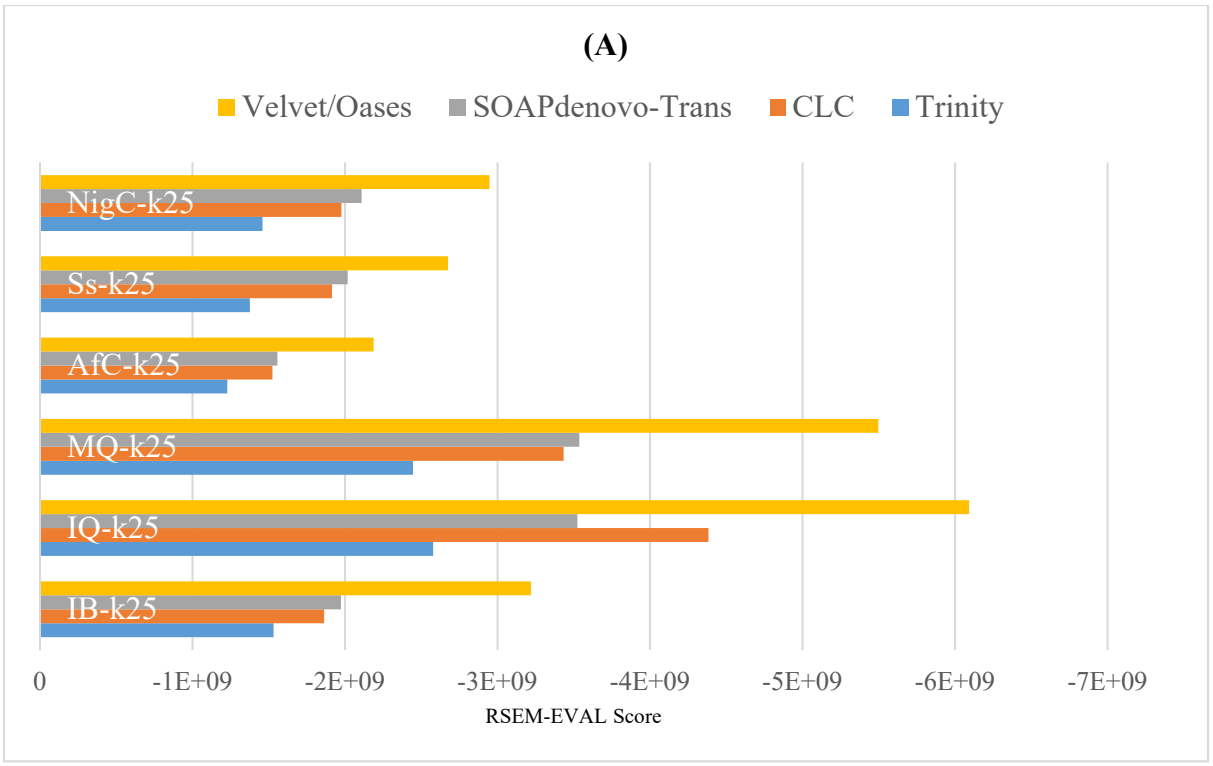
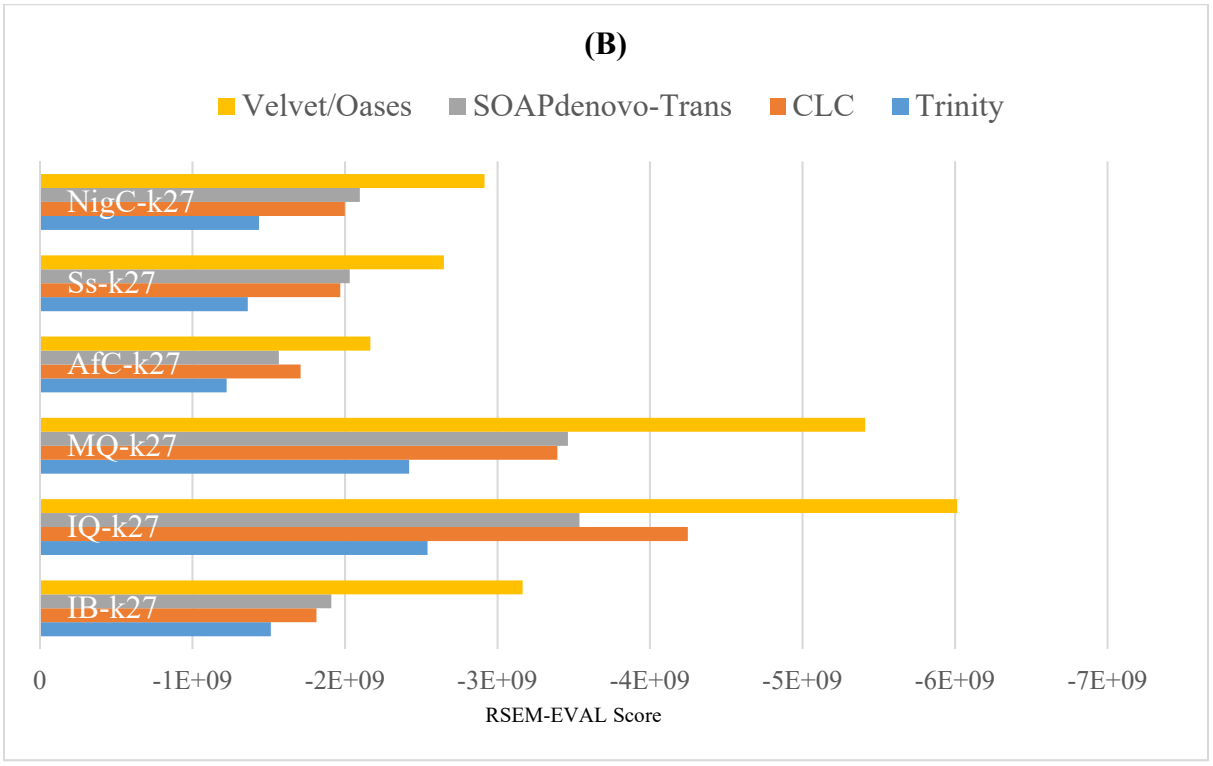


Figure 3.7: Optimal assembly score obtained using TransRate by measuring accuracy and completeness of each assembly.


(A)
Velvet/Oases
SOAPdenovo-Trans
CLC
Trinity
NigC-k25
Ss-k25
AfC-k25
MQ-k25
IQ-k25
IB-k25
0
-1E+09
-2E+09
-3E+09
-4E+09
-5E+09
-6E+09
-7E+09
RSEM-EVAL Score


(B)
Velvet/Oases
SOAPdenovo-Trans
CLC
Trinity
NigC-k27
Ss-k27
AfC-k27
MQ-k27
IQ-k27
IB-k27
0
-1E+09
-2E+09
-3E+09
-4E+09
-5E+09
-6E+09
-7E+09
RSEM-EVAL Score

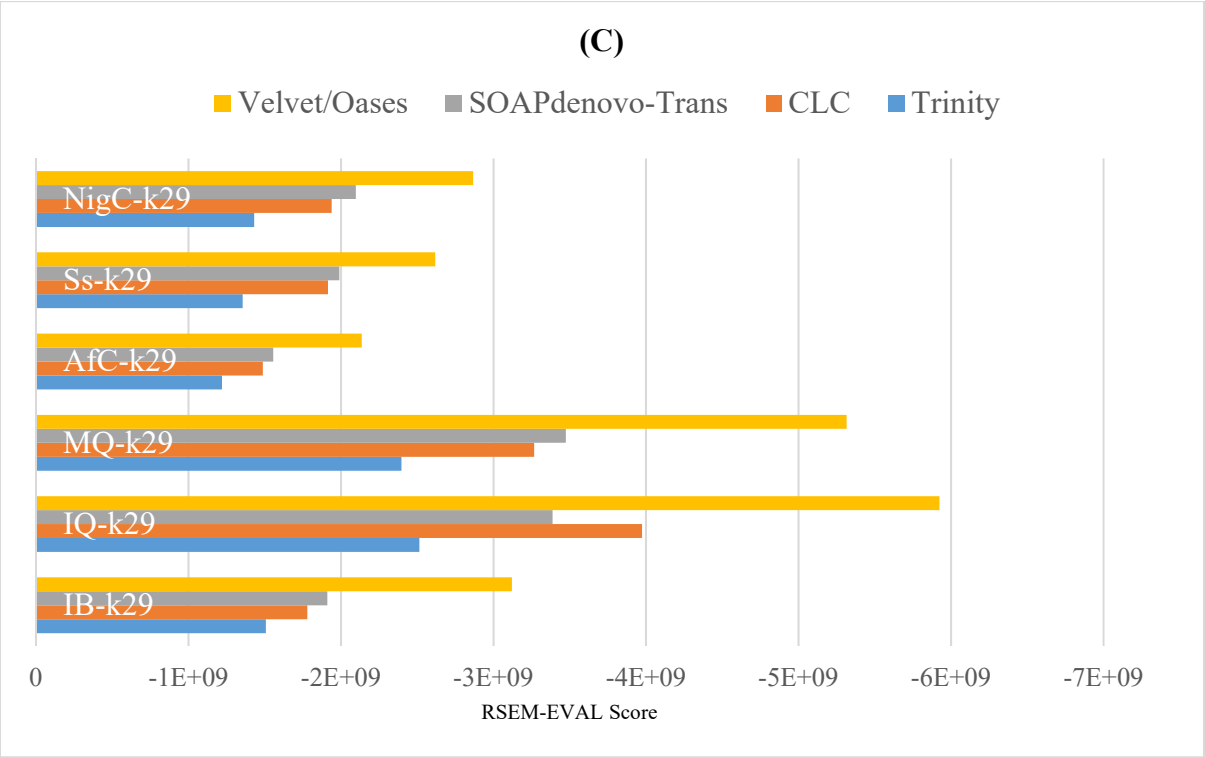


Figure 3.8: RSEM-EVAL scores for the *B. tabaci* transcriptome assemblies. The x-axis represents the RSEM-EVAL score for each assembly, with blue representing the highest RSEM-EVAL score. A) RSEM-EVAL scores calculated for assemblies generated using k-mer 25. Here scores near to zero indicates more accurate assemblies. B) RSEM-EVAL scores calculated for assemblies generated using k-mer 27. C) RSEM-EVAL scores calculated for assemblies generated using k-mer 29.

3.3.3 Annotation Statistics

Another approach using annotation metrics may be more appropriate for assessing assembly quality. For this approach, a closely related species with full genome information can be used as the closely related species are more likely to share common genes than distantly related species. In the absence of full genome information for *B. tabaci*, assembled contigs were annotated against a protein dataset of *A. pisum* and *D. citri* using CRB-BLAST. CRB-BLAST uses bi-directional BLAST alignments using BLASTX (assembly to reference) and TBLASTN (reference to assembly). The main reason for annotating in both directions is to identify sequence and a protein that have a best match with each other known as reciprocal best hits (RBH), that the two sequences are orthologues and are derived from the same ancestral locus.

The results in Figures 3.9 and 3.10 indicated that the assemblies generated using the Trinity produced highest number of reciprocal best hits whereas the assemblies generated using CLC produced lowest number of reciprocal best hits for all six samples.

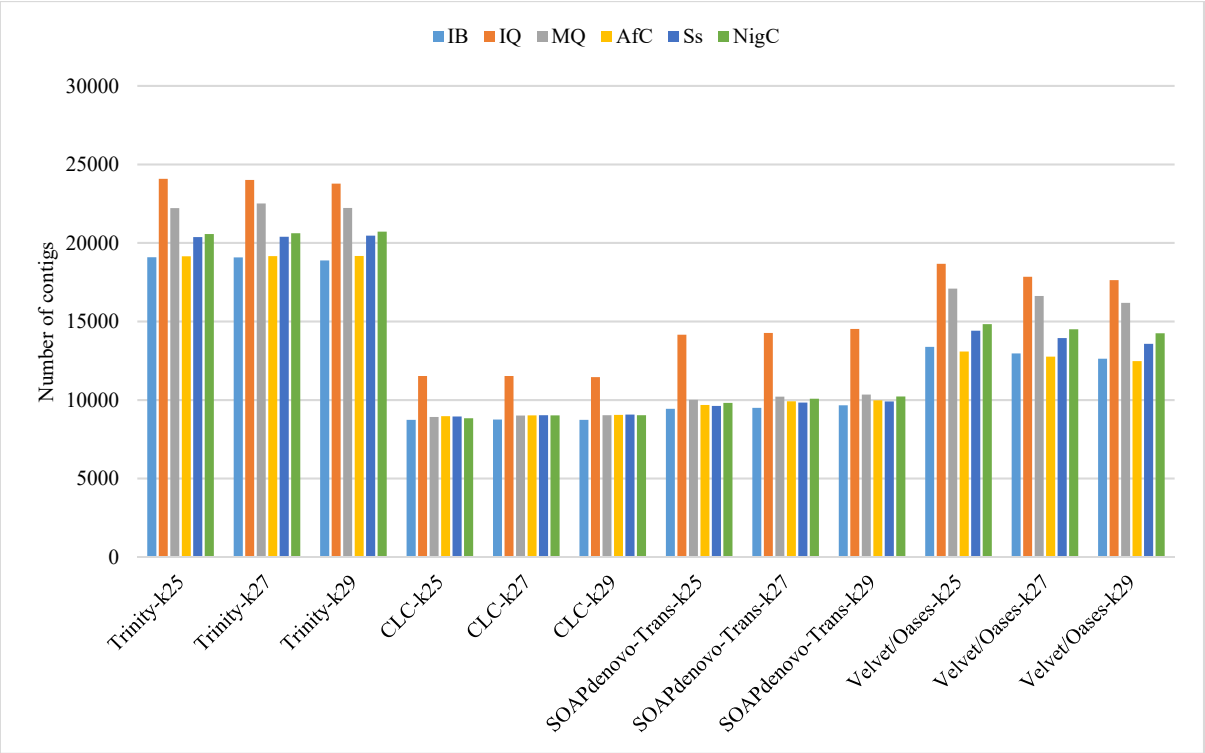


Figure 3.9: Number of reciprocal best hits against *Acyrthosiphon pisum* protein dataset using CRB-BLAST.

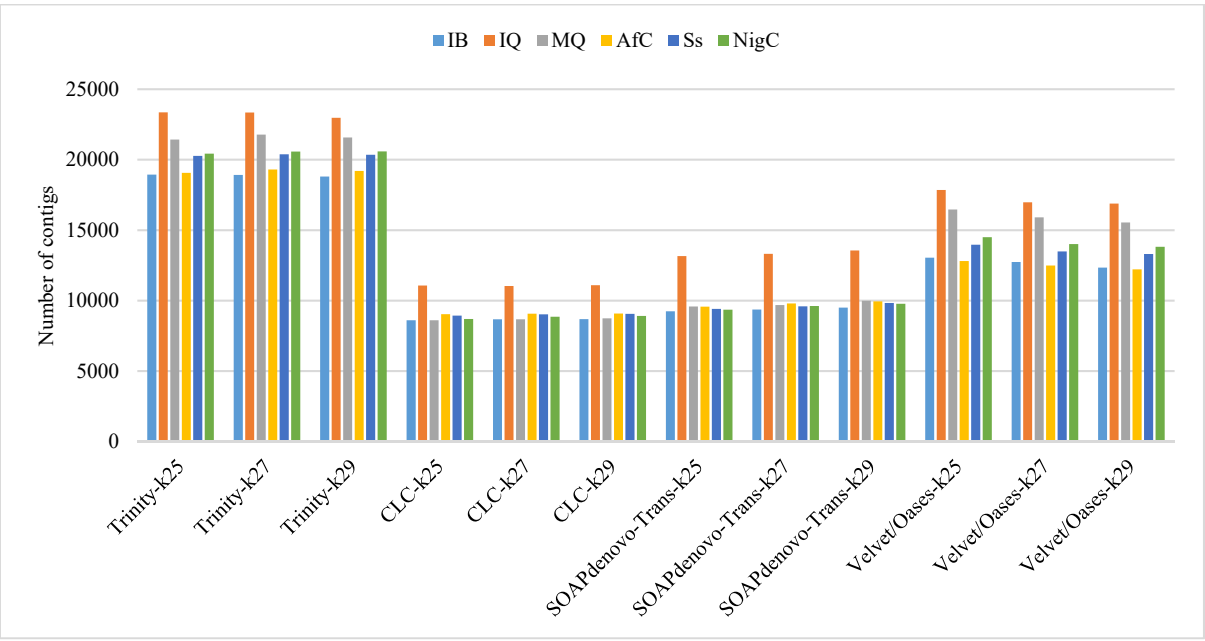
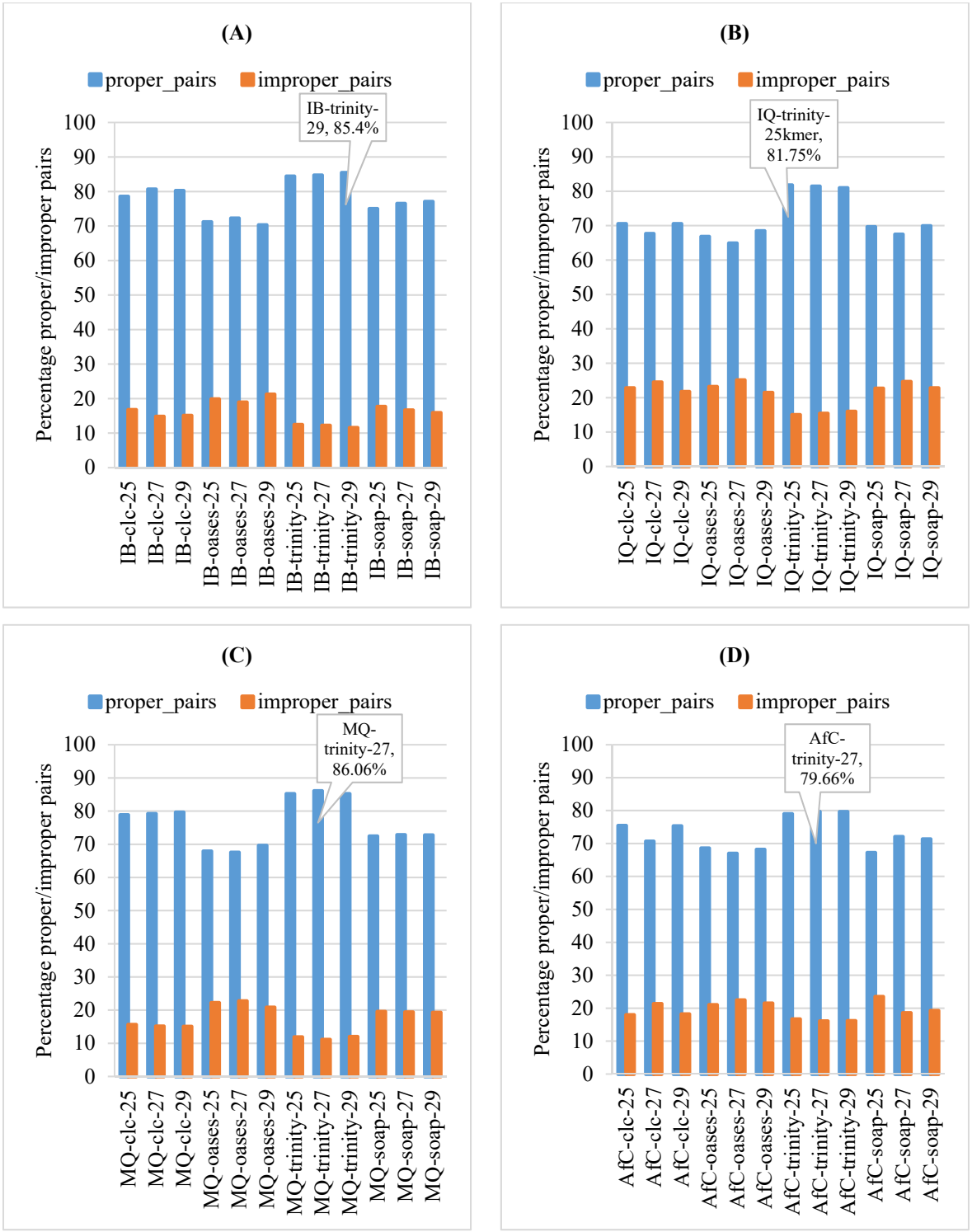


Figure 3.10: Number of reciprocal best hits against *Diaphorina citri* protein dataset using CRB-BLAST.

3.3.4 Evaluating the quality of assembly

To assess the completeness of assemblies, we have aligned raw RNA-Seq reads from each sample back to assemblies generated using the Trinity, CLC, SOAPdenovo-Trans and Velvet/Oases with k-mer 25, 27 and 29 to quantify read representation. The script aligns each read file to assembly separately, and then links up the pairs to count number of reads that are found as properly paired in addition to those that align to separate contigs also referred to as improper pairs. The program also provides indication of only left and right aligning pairs to a contig.

Generally, in a high-quality assembly, the higher the percentage of reads exist as proper pairs less the possibility of fragmented assembly. Here, in Figure 3.11, the average percentage of paired end reads classed as proper pairs were ~84% (IB), ~81% (IQ), ~85% (MQ), ~79% (AfC), ~89% (Ss) and 83% (NigC) for assemblies generated using the Trinity. The results showed that the assembled contigs of Trinity aligns well with the reads in comparison with other assemblers. The results also indicate that only ~10-15% contigs do not match with the reads and thus can be classified as improperly paired. There could be biological differences within the samples that may influence the mapping orientation, but these could not be quantified or accounted for in the analysis due to the lack of a reference genome. Such differences include gene duplication, gene deletion and insertions.


(A)
proper_pairs
improper_pairs
IB-trinity-29, 85.4%
Percentage proper/improper pairs
100
90
80
70
60
50
40
30
20
10
0
IB-clc-25
IB-clc-27
IB-clc-29
IB-oases-25
IB-oases-27
IB-oases-29
IB-trinity-25
IB-trinity-27
IB-trinity-29
IB-soap-25
IB-soap-27
IB-soap-29
(B)
proper_pairs
improper_pairs
IQ-trinity-25kmer, 81.75%
Percentage proper/improper pairs
IQ-clc-25
IQ-clc-27
IQ-clc-29
IQ-oases-25
IQ-oases-27
IQ-oases-29
IQ-trinity-25
IQ-trinity-27
IQ-trinity-29
IQ-soap-25
IQ-soap-27
IQ-soap-29
(C)
proper_pairs
improper_pairs
MQ-trinity-27, 86.06%
Percentage proper/improper pairs
MQ-clc-25
MQ-clc-27
MQ-clc-29
MQ-oases-25
MQ-oases-27
MQ-oases-29
MQ-trinity-25
MQ-trinity-27
MQ-trinity-29
MQ-soap-25
MQ-soap-27
MQ-soap-29
(D)
proper_pairs
improper_pairs
AfC-trinity-27, 79.66%
Percentage proper/improper pairs
AfC-clc-25
AfC-clc-27
AfC-clc-29
AfC-oases-25
AfC-oases-27
AfC-oases-29
AfC-trinity-25
AfC-trinity-27
AfC-trinity-29
AfC-soap-25
AfC-soap-27
AfC-soap-29

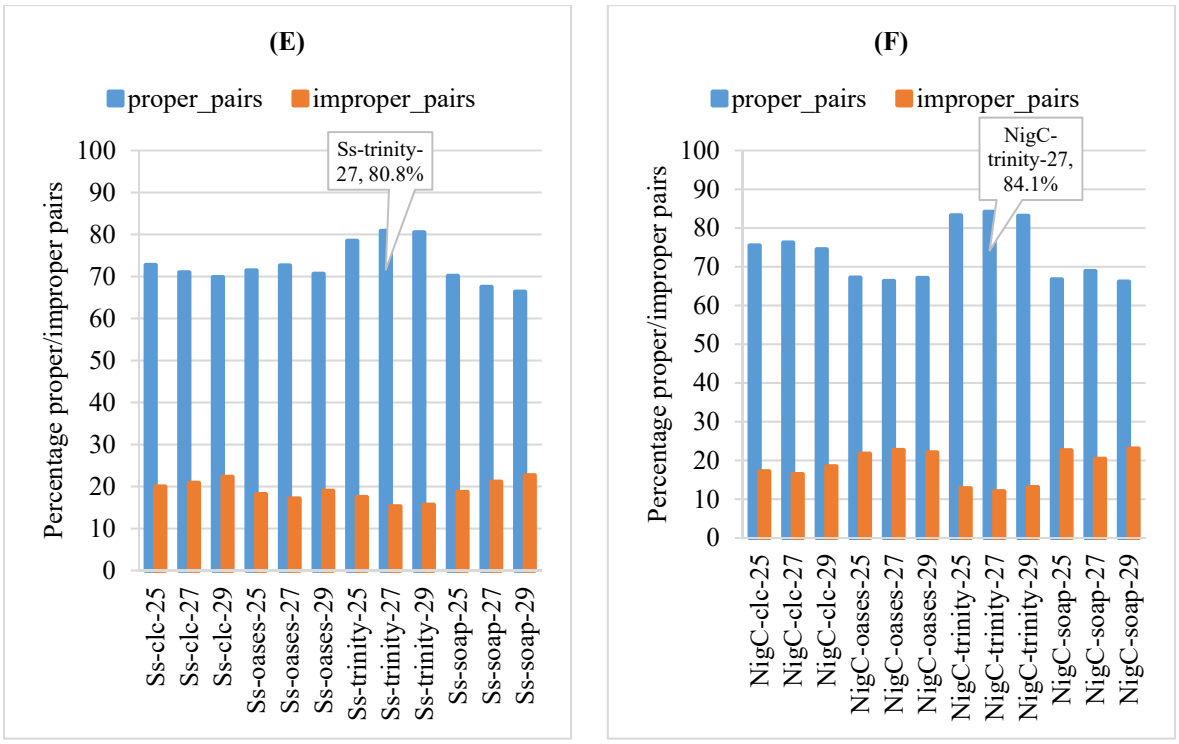


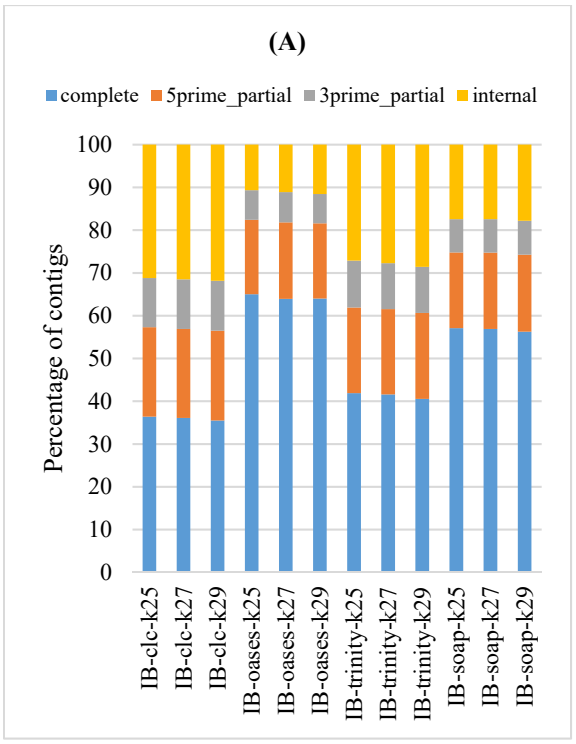
Figure 3.11: Percentage of properly and improperly paired reads for assemblies generated using Trinity, CLC, SOAPdenovo-Trans and Velvet/Oases with k-mer 25, 27 and 29. A) Representation of reads against assemblies of sample IB. B) Representation of reads against assemblies of IQ. C) Representation of reads against assemblies of MQ. D) Representation of reads against assemblies of AfC. E) Representation of reads against assemblies of Ss. F) Representation of reads against assemblies of NigC.

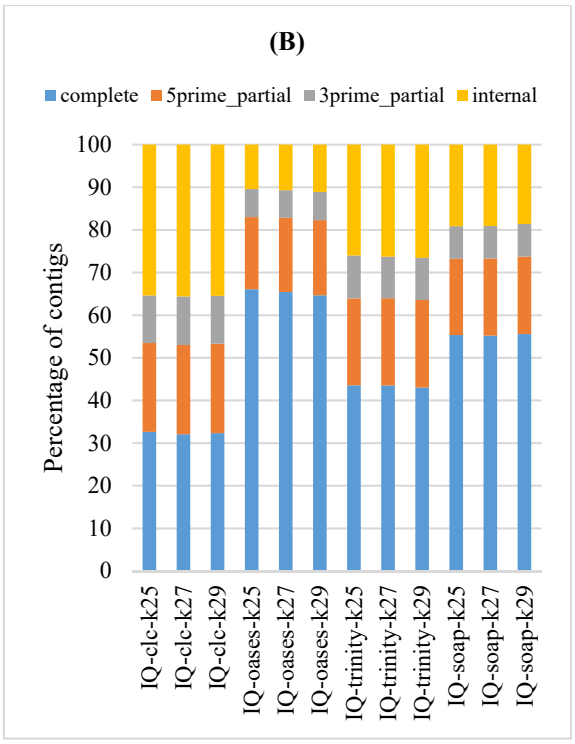
3.3.5 Assessment of transcriptome completeness

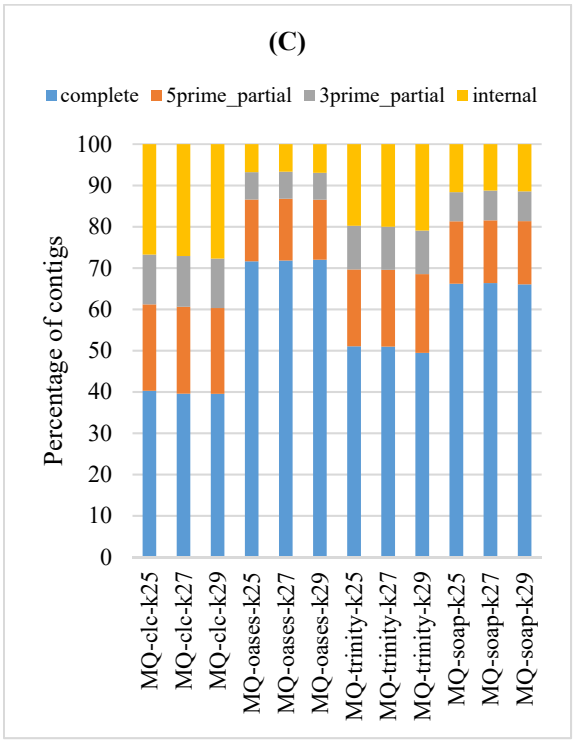
Transcriptome completeness is an important measure of quality of assembly as well an important measure for downstream analysis. All assembled transcripts were assessed using TransDecoder to identify full length, partial and internal contigs based on open reading frames (ORFs). The contigs assembled using the Oases were found to encode highest percentage of full length ORFs compared to assemble using the CLC, Trinity and SOAPDenovo (Figure 3.12). Also the total percentage of internal ORFs were comparatively less within Oases assemblies. The CLC assemblies were found to be highly fragmented with highest percentage of internal ORFs; missing both start and stop codons.

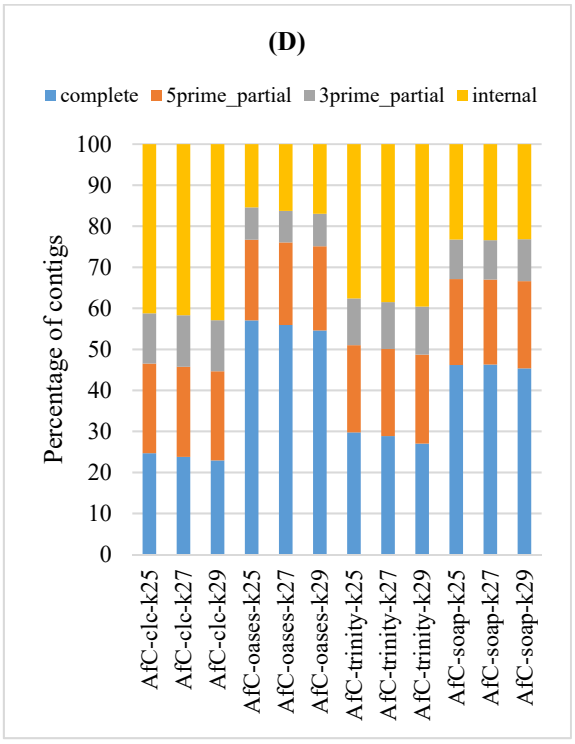
BUSCO uses homology modelling to search single copy orthologs curated based on sequence uniqueness and conservations levels derived from hundreds of genomes (Simao *et al*., 2015). We measured the completeness of assembled transcripts using 2,675 conserved genes from available arthropod genomes. We found that transcripts assembled using k-25 had highest number of hits compared to assembled using K-27 and 29. Here, we have only selected Trinity assemblies as all assembly evaluation statistics suggested that the Trinity outperformed other three assemblers.

We found that 73% (IB), 79% (IQ), 82% (MQ), 59% (AfC), 70% (Ss) and 72% (NigC) of genes had a hit in our assembled transcriptome, choosing only assemblies using k-25. Although k-25 assemblies were better, other assemblies using k-27 and 29 also resulted in a relatively similar number of hits (Figure 3.13).


(A)
complete
5prime_partial
3prime_partial
internal
Percentage of contigs
100
90
80
70
60
50
40
30
20
10
0
IB-clc-k25
IB-clc-k27
IB-clc-k29
IB-oases-k25
IB-oases-k27
IB-oases-k29
IB-trinity-k25
IB-trinity-k27
IB-trinity-k29
IB-soap-k25
IB-soap-k27
IB-soap-k29


(B)
complete
5prime_partial
3prime_partial
internal
Percentage of contigs
100
90
80
70
60
50
40
30
20
10
0
IQ-clc-k25
IQ-clc-k27
IQ-clc-k29
IQ-oases-k25
IQ-oases-k27
IQ-oases-k29
IQ-trinity-k25
IQ-trinity-k27
IQ-trinity-k29
IQ-soap-k25
IQ-soap-k27
IQ-soap-k29


(C)
complete
5prime_partial
3prime_partial
internal
Percentage of contigs
100
90
80
70
60
50
40
30
20
10
0
MQ-clc-k25
MQ-clc-k27
MQ-clc-k29
MQ-oases-k25
MQ-oases-k27
MQ-oases-k29
MQ-trinity-k25
MQ-trinity-k27
MQ-trinity-k29
MQ-soap-k25
MQ-soap-k27
MQ-soap-k29


(D)
complete
5prime_partial
3prime_partial
internal
Percentage of contigs
100
90
80
70
60
50
40
30
20
10
0
AfC-clc-k25
AfC-clc-k27
AfC-clc-k29
AfC-oases-k25
AfC-oases-k27
AfC-oases-k29
AfC-trinity-k25
AfC-trinity-k27
AfC-trinity-k29
AfC-soap-k25
AfC-soap-k27
AfC-soap-k29

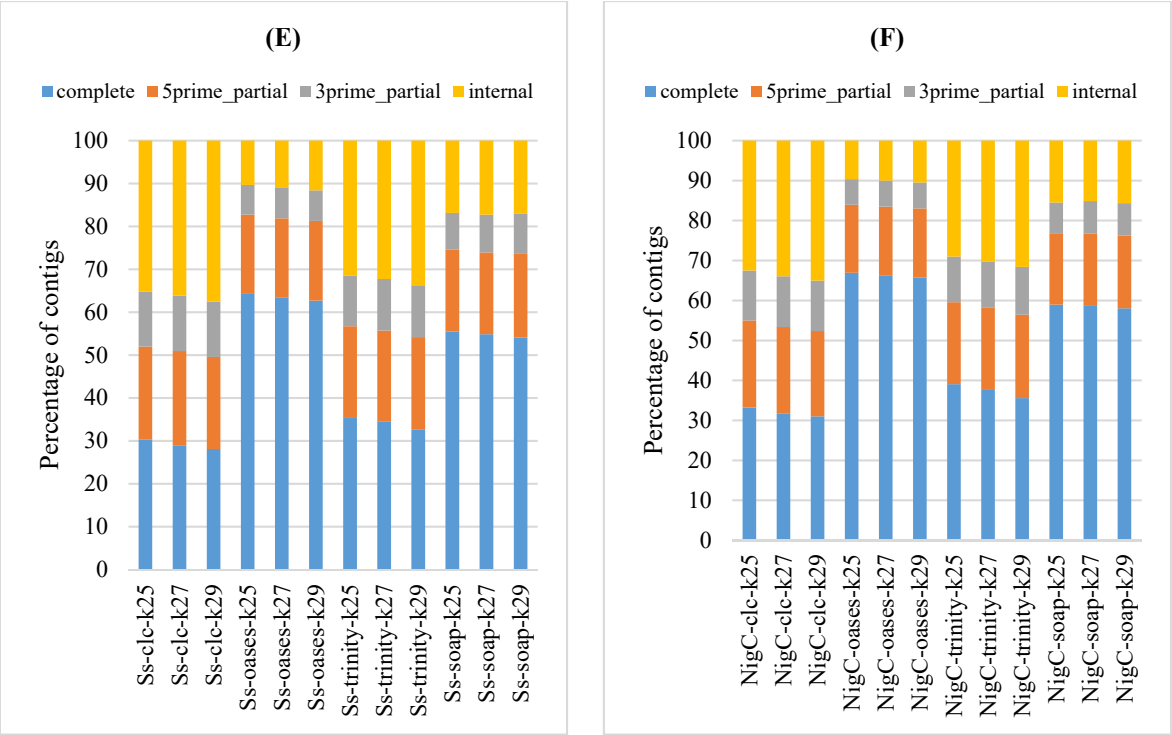


Figure 3.12: Total percentage of contigs identified as complete, partial and internal. Here, contigs containing start and stop codon referred as complete; 5prime-partial, missing start codon; 3prime-partial, missing stop codon and internal, missing both start and stop codons. A) Number of protein coding contigs predicted in sample IB. B) Number of protein coding contigs predicted in IQ. C) Number of protein coding contigs predicted in MQ. D) Number of protein coding contigs predicted in AfC. E) Number of protein coding contigs predicted in Ss. F) Number of protein coding contigs predicted in NigC.

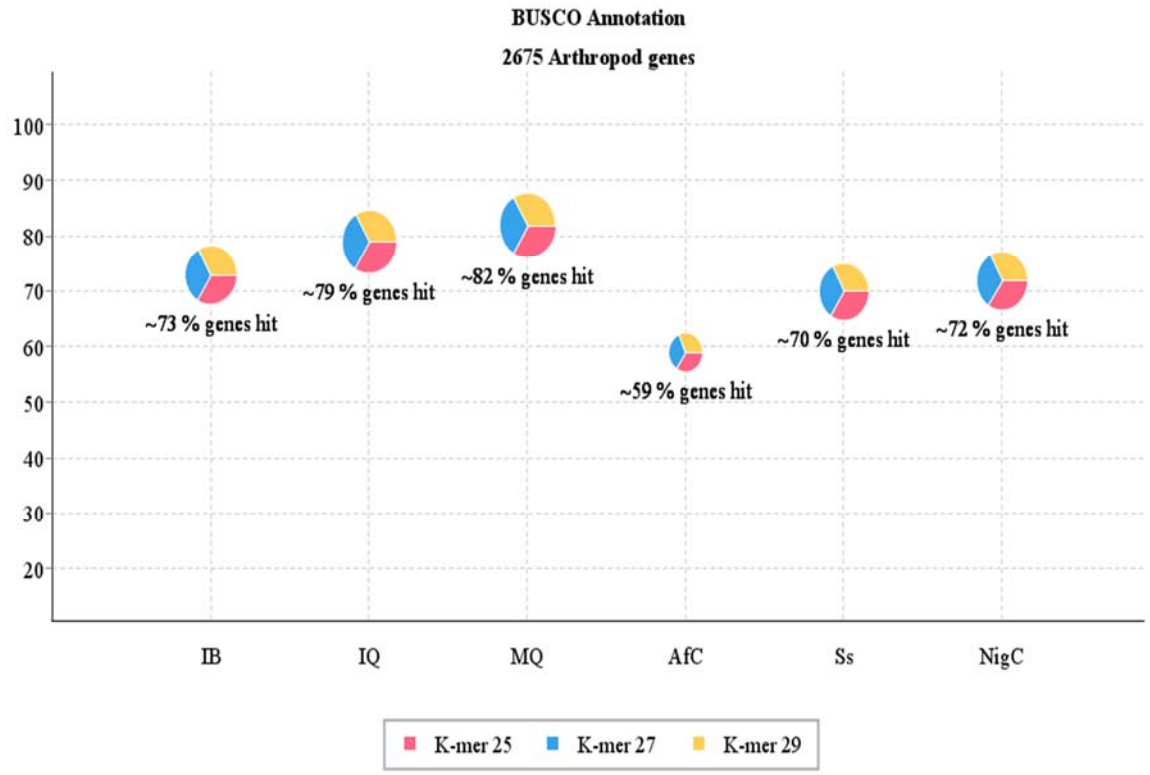


Figure 3.13: BUSCO hits against assembled transcripts showing similarity between all three k-mer values.

3.3.6 Assembly clustering and optimization

The generation of clustered assemblies (CAs) was performed using CD-HIT-EST with 100% identity cut-off to ensure all duplicate transcripts are removed for downstream analysis. Alongside CD-HIT-EST, we also used custom Perl script to remove those which were 100% identical both in length and in similarity to save computing time while performing CD-HIT-EST. Figure 3.14 shows the CAs obtained after processing all Trinity transcripts with varying k-mer.

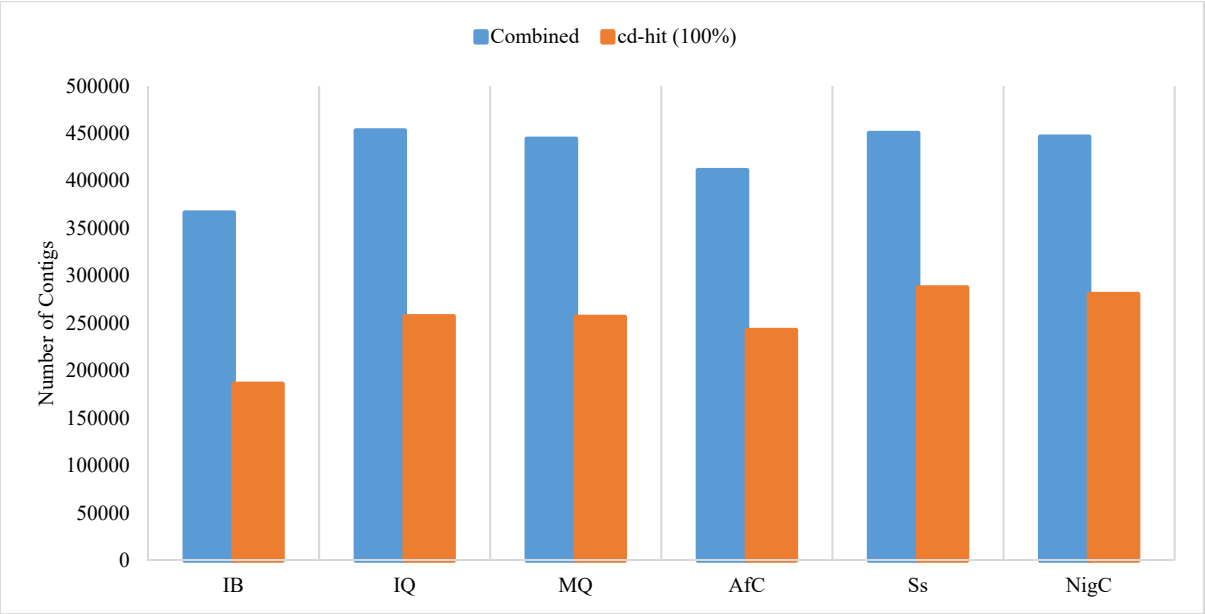


Figure 3.14: Number of contigs decreased after removing transcripts which were entirely covered by other transcripts with 100% identity.

3.4 Discussion

With developments in the past decade in next generating sequencing, short read transcriptome analysis has become more commonly used for studying non-model organisms. However, the transcriptome assemblies generated using short reads need to be validated as accurate *de novo* assembly is crucial for downstream analysis (Amin *et al*., 2014). In this study, we compared and evaluated the performance of four *de novo* transcriptome assemblers: Trinity, CLC, SOAPdenovo-Trans and Velvet followed by Oases using *B. tabaci* transcriptome data sets. In order to reveal the important factors for selecting the best assembly, we measured results in terms of contig statistics, assembly statistics and annotation statistics using different reference-free quality measures. Of particular interest in finding the best assembly parameters was to set some guidelines for assembling short Illumina reads in the absence of a reference genome.

Comparison of *de novo* transcriptome assemblies showed that each assembly program results in different outcomes even though the assemblers used in this study are all based on *de-Bruijn* graph. The multi k-mer approach used in this study showed that the number of contigs and length of contigs varied across assemblies using the same data set. In our comparison of four *de novo* assemblers using *B. tabaci* short read Illumina data, we found that the Trinity outperformed CLC, SOAPdenovo-Trans and Velvet/Oases in the percentage of reads mapped successfully, overall assembly score and annotation score. Here, we dismissed other reference free measures such as average contig length, number of contigs and N50 as Trinity produced the highest number of contigs with relatively small contig average length and N50 lengths compared to Velvet/Oases. These suggested that the primary measures for evaluating assemblies are unclear and requires other measures to assess the quality of assemblies. Overall, these results are similar to those obtained by O'Neil and Emrich (2013) when they assessed the quality of *Drosophila melanogaster* data set. They also found that the primary assembly measures such as contig count, mean length and N50 length repurposed from genome assembly evaluation metrics may not be useful for assessing quality of transcriptome assemblies.

When scored using TransRate, Trinity produced the best optimal assembly score based on high scoring contigs obtained for each assembly. Out of three k values, assembly using k-25 produced best assemblies, while k-27 and k-29 assemblies were comparable with k-25. Similarly, the highest percentage of contigs were successfully mapped back to reads for

assemblies generated using the Trinity as shown in Figure 3.6 compared to CLC, SOAPdenovo-Trans and Velvet/Oases. These results suggested that the contigs assembled using CLC, SOAPdenovo-Trans and Velvet/Oases did not find a good match within reads or that the reads aligned to multiple contigs due to sequence similarity between contigs. This shows redundancy particularly in Velvet/Oases assemblies as only about ~60% of the contigs were successfully mapped to reads. Another reason could be that the reads originated from different contigs were merged into same contigs in the form of chimeric contigs. These chimeric sequences can cause major problems in downstream analysis, especially when working on expression analysis studies (Mundry *et al*., 2012).

We also evaluated assemblies based RSEM-EVAL probability model. RSEM-EVAL assembly score generated for all assemblies using k-25, 27 and 29 suggested that the assemblies produced using Trinity with k-29 performed better. This result was contrary to results produced by an algorithmically similar program TransRate. Of the evaluation metrics using mapping, no definite results were obtained to score best assembly with optimal k-mer size, annotation based statistics were used to further support these findings. However, in the absence of reference genome, assembled transcriptome data sets are often compared against closely related, well annotated species to understand assembly completeness (O'Neil and Emrich, 2013). We annotated assembled contigs of each assembly against *Acyrthosiphon pisum* and *Diaphorina citri* protein data set using CRB-BLAST to produce best possible match using best e-value cutoff scores, which are important for finding best possible hits. When searched in both directions using CRB-BLAST (i.e. BLASTX and TBLASTN), assemblies using Trinity with k-29 produced marginally higher hits, 19,173 for sample AfC, 20,464 for Ss and 20,713 for NigC compared to 19,161 (AfC), 20,387 (Ss), 20,615 (NigC) with k-27 and 19,145 (AfC), 20,371 (Ss) and 20,565 (NigC) with k-25, whereas assemblies generated using Trinity with k-25 produced highest hits for samples IB (19,088) and IQ (24,085) against *Acyrthosiphon pisum* (Figure 3.9). Whereas, the CRB-BLAST results in Figure 3.10 showed that the assemblies using Trinity with k-27 produced slightly higher hits for samples MQ (21,771), AfC (19,295) and Ss (20,373) against *Diaphorina citri* than assemblies with k-25 and k-29.

While evaluating the assembly completeness by calculating the read representation, we found that the Trinity assemblies aligns well with raw reads in comparison to other assemblies but again as observed in all other evaluation statistics, we found no single

dominant k-mer across all samples as the best assembly. Sample IB, assembled best using k-mer 29, whereas k-mer 25 was best for IQ and k-mer 27 for all other samples MQ, AfC, Ss and NigC. We used TransDecoder to predict full length protein coding contigs within assemblies and found that the Trinity assemblies had highest number of complete ORFs. While this suggested that Trinity assemblies were best but when used total percentage of contigs with complete ORFs, Oases assemblies were found to be better than the Trinity. Here, we used percentage instead counts as the number of contigs were varying. But when we compared assemblies for completeness using BUSCO, we found that the Trinity assemblies with k-mer 25 had highest hits against referenced arthropod genes which suggest that the default k-mer 25 set as default for earlier version of the Trinity assembler is found to be optimal in this case. But then when we compared the same results with k-mer 27 and 29, the difference was relatively small, which we have observed in all our evaluation statistics.

In summary, our analysis indicated that Trinity performed best in assembling full length transcripts based on percentage of reads successfully mapped back to contigs, number of CRB BLAST hits obtained against *A. pisum* and *D. citri* and based on BUSCO analysis. However, the in-depth analysis using different k-mer sizes for evaluating assemblies suggested that no single k-mer value exhibited as optimal for data used in this study and therefore we used clustered assembly using the Trinity transcripts by merging all k-mer assemblies into one. But because the redundancy in transcripts while generating CAs can cause bias in downstream analysis, we used CD-HIT-EST to remove all redundant transcripts with 100% similarity. While optimizing and analysing de novo assembly of C. *sinensis* using SOAPdenovo, ABySS, trans-ABySS, Oases and Trinity, Zhao *et al*. 2011 found that Trinity consistently performed better in terms of transcript accuracy, integrity, completeness and sensitivity of assembled transcripts under single k-mer strategy as the older version of Trinity (release 20110519) did not support multi k-mer values and therefore suggested that assembly results can be further improved if MK strategy is applied to Trinity. Zhao *et al*. 2011 also found that Trinity reduces number of fused transcripts by using strand specific assembly parameters when paired-end reads were used. Furthermore, by taking the advantage of multi k-mer properties, we incorporated all contigs into one final assembly to incorporate different quantities of unique contigs produced in each single k-mer assembly to improve biological information retrieval from transcriptome (Haznedaroglu *et al*., 2012).

Chapter 4: Annotation of the *Bemisia tabaci* transcriptome derived from *de novo* clustered assembly

4.1 Introduction

The whitefly *B. tabaci* is a species complex of more than 34 morphologically indistinguishable cryptic species (Boykin and De Barro, 2014). Two species of the complex, Middle East-Asia Minor 1 (MEAM1) and Mediterranean (MED), referred to previously as the B and Q biotypes have risen to international importance due to their damage potential (Wang *et al*., 2011). Despite *B. tabaci's* global importance as one of the world's top 100 invasive species, limited genomic sequence resources are available in public domains (Wang *et al*., 2010). Currently (October 3, 2016), there are about 12,094 EST, 14,359 protein and 260,065 nucleotide sequences available for *B. tabaci* on NCBI (source: http://www.ncbi.nlm.nih.gov). While efforts to produce complete genome sequence of *B. tabaci* remain unpublished, cDNA sequencing to study candidate genes are ongoing.

Over the past years, next generation sequencing has significantly improved the efficiency and speed of gene discovery and also accelerated research in areas of gene-expression profiling and comparative genomics studies (Wang *et. al*., 2010). Furthermore, next generation sequencing, especially RNA sequencing has been widely used to provide information on transcript profiles of organisms, and also to understand biological processes in both well-studied model organisms and non-model organisms (Zhao *et al*., 2011). Recent transcriptome studies on *B. tabaci* populations MED and MEAM1 revealed the sequence divergence between these two species of whiteflies and also provided a useful resource to study evolutionary relationships between species, insecticide resistance and genes responsible for host plant utilization (Wang *et al*., 2011).

The aim of this study was to generate transcriptome profiles of cassava and non-cassava colonizing *B. tabaci* populations. These transcriptome datasets will provide useful molecular and functional resource to understand and compare sequence divergence between cassava and non-cassava colonizing *B. tabaci* populations as well as to shed light on possible mechanisms that enable *B. tabaci* to utilize cassava as a host plant. The comparison between samples used in this study and published *B. tabaci* transcriptome of MED (Wang *et al*.,

2010), MEAM1 (Wang *et al*., 2011; Xie *et al*., 2012) and Asia II 3 (Wang *et al*., 2012) is essential to understand the roles of conserved genes and their global sequence divergence.

4.2 Methods

4.2.1 Annotation

For functional annotation, Trinity assembled RNA-seq contigs obtained from clustered-assembly for all samples were searched using BLASTx (Altschul *et al*., 1990) against the non-redundant (nr) NCBI nucleotide database using a cut-off E-value of 10^{-3}. The top-hits for BLASTx results were retrieved and stored in a separate file for statistical analysis. The publicly available version of Blast2GO V.2.8.0 (Gotz et. al., 2008) was then used to perform Gene Ontology (GO), Enzyme Code (EC) mapping and to identify KEGG (Kyoto Encyclopaedia of Genes and Genomes) metabolic pathways helped by BLASTx results. Amino acid sequences of all samples were obtained using the TransDecoder tool with default parameters. The Pfam database was searched using hmmscan (Finn *et al*., 2011) to identify conserved domains in protein sequences.

4.2.2 Secretome identification

The presence of signal peptide motif in protein sequences of *B. tabaci* was predicted using the SignalP (Nielsen, 2017) with default parameters. TMHMM (Krogh et. al., 2001) program was used to predict transmembrane regions in amino acid sequences using default parameters.

4.2.3 Molecular marker identification

To identify Simple Sequence Repeats (SSR), contigs were searched for identification and localization of microsatellites using MIcroSAtellite identification tool (MISA PERL script) (Thiel *et al*., 2003). Microsatellites were identified with parameters: mononucleotide repeats with >=10 repeats, di-nucleotide repeats with >=6 repeats and tri-, tetra-, penta- and hexa-nucleotide repeats with >=5 repeats. Compound microsatellites were examined with an interval of <=100 bp of the total motif length. Statistical analysis was performed to summarize the number of SSR with type of motif and the length distribution of motif.

4.3 Results

4.3.1 Overview of assemblies

After eliminating redundant contigs and clustering all k-mer assemblies for each sample, final clustered assemblies were generated using the Trinity based on evaluation statistics discussed in Chapter 3. The assembled contigs for all samples were varying in length from 224 bp to 32,043 bp. The most number of contigs (287,559) were assembled for sample Ss with a mean length (N50) of 986.92bp and total size of 109,930,193bp. The number of contigs assembled for sample Ss (287,559) and sample NigC (280,616) were higher than the samples IB, IQ, MQ and AfC with 185,895, 257,163, 256,401 and 242,664 respectively as shown in Table 4.1.

Most of the assembled contigs (64.74% (IB), 63.06% (IQ), 59.47% (MQ), 73.16% (AfC), 68.91% (Ss) and 65.50% (NigC)) were 200 to 1000 bp in length, followed by 1000 to 2000 bp and, 2000 to 3000 bp in length as shown in Figure 4.1. Although the majority of contigs fall between 200 to 1000 bp, a total of 4194 (IB), 7245 (IQ), 9654 (MQ), 1704 (AfC), 3131 (Ss) and 4293 (NigC) contigs with length >5,000 bp were obtained.

	IB	IQ	MQ	AfC	Ss	NigC
Total number of reads	46,892,996	81,867,506	82,289,190	30,310,262	38,438,672	43,590,886
Total base pairs (bp)	115,514,683	150,689,812	175,787,885	90,710,656	109,930,193	121,086,445
Read length (bp)	87	87	87	87	87	87
Total number of contigs	185,895	257,163	256,401	242,664	287,559	280,616
Mean length of contigs	1154.6	1221.26	1354.83	881.4	986.92	1080.84
Min length of contigs (bp)	224	224	224	224	224	224
Max length of contigs (bp)	21,769	26,405	26,876	22,130	20,496	32,043
GC content (%)	39	39	39	38	39	38

Table 4.1: Total number of sequencing reads obtained from Illumina paired-end sequencing and the number of contigs obtained by clustering contigs assembled using the Trinity software.

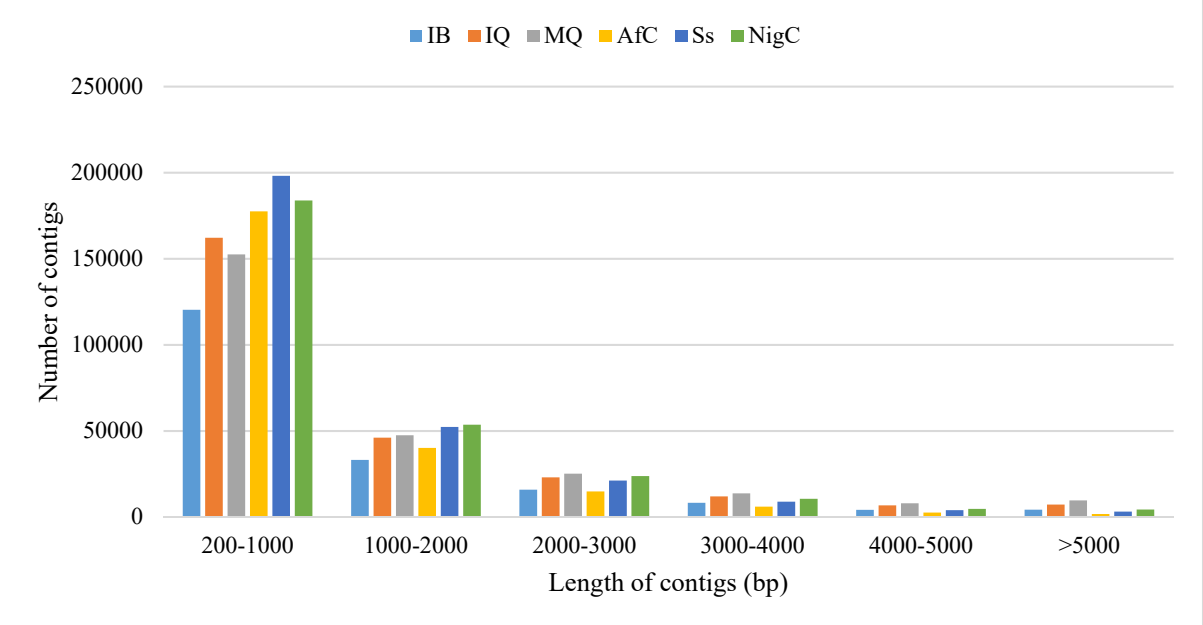


Figure 4.1: Contig length distribution of six samples based on Trinity assembly. All assembled contigs were used to see the differences in length and the number of contigs assembled for each sample. Y-axis represents the length of each assembled contig for each sample. X-axis indicates the total number of contigs assembled for each sample.

4.3.2 Functional annotation

To assess the putative function, all assembled contigs were submitted to BLASTx similarity search against NCBI non-redundant protein database with an E-value cut-off of 10^{-3}. Using this approach, 70,648 (IB), 98,777 (IQ), 95,895 (MQ), 81,023 (AfC), 97,157 (Ss) and 96,569 (NigC) contigs with significant BLASTx hits were found. Of these, a total of 22,122 (IB), 27,894 (IQ), 23,535 (MQ), 25,290 (AfC), 26,276 (Ss) and 25,558 (NigC) contigs corresponded to unique protein accessions in the NR protein database.

The E-value distribution (Figure 4.2) of the BLASTx top hits revealed that 61% of the mapped sequences showed strong homology (smaller than $1.0E^{-40}$) for sample IB, 62% for IQ, 65% for MQ, 57% for AfC, 60% for Ss and 62% for NigC, and other 39% (IB), 38% (IQ), 35% (MQ), 43% (AfC), 40% (Ss) and 38% (NigC) of the contigs ranged between $1.0E^{-4}$ to $1.0E^{-40}$. The sequence similarity distribution graph (Figure 4.3) of the predicted proteins showed that 40.86% (IB), 39.55% (IQ), 41.98% (MQ), 40.14% (AfC), 40.93% (Ss) and 41.57% (NigC) of the contigs have a similarity ranging from 40% to 60%, while 12.66% (IB), 14.73% (IQ), 10.89% (MQ), 13.51% (AfC), 12.20% (Ss) and 11.70% (NigC) of the contigs showed similarity higher than 80%.

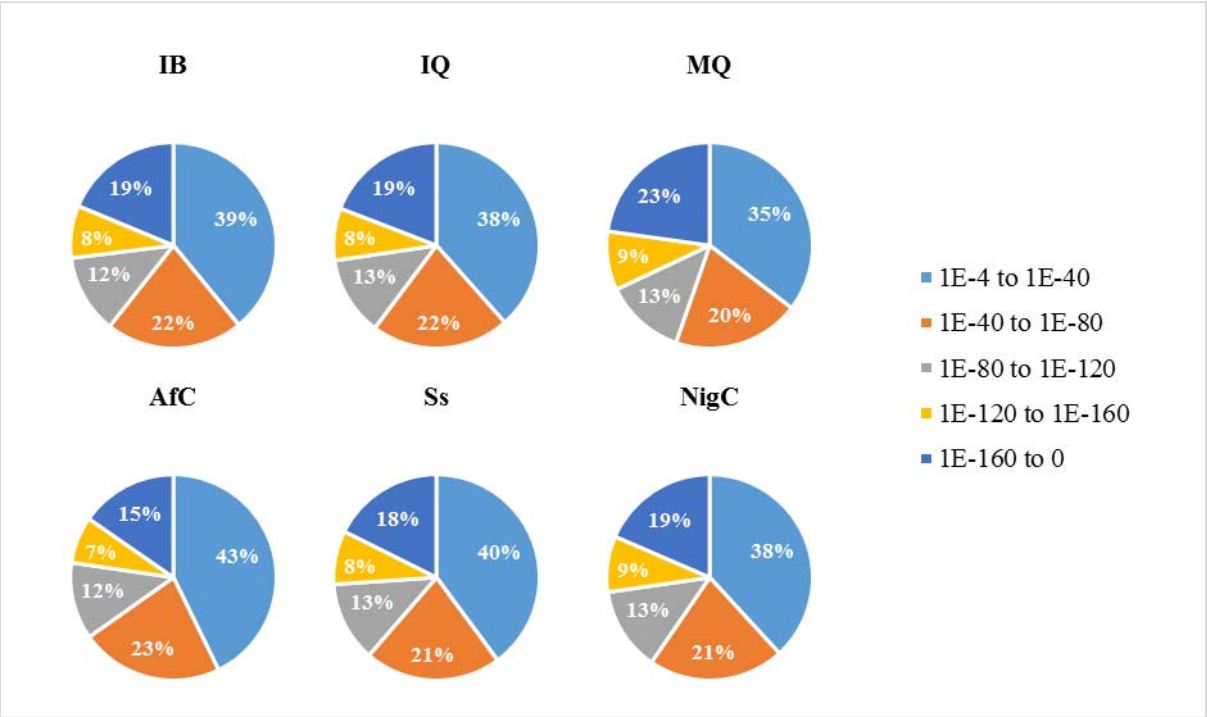


Figure 4.2: E-value distribution of top BLASTx hits for all samples showed similar distribution patterns amongst samples.

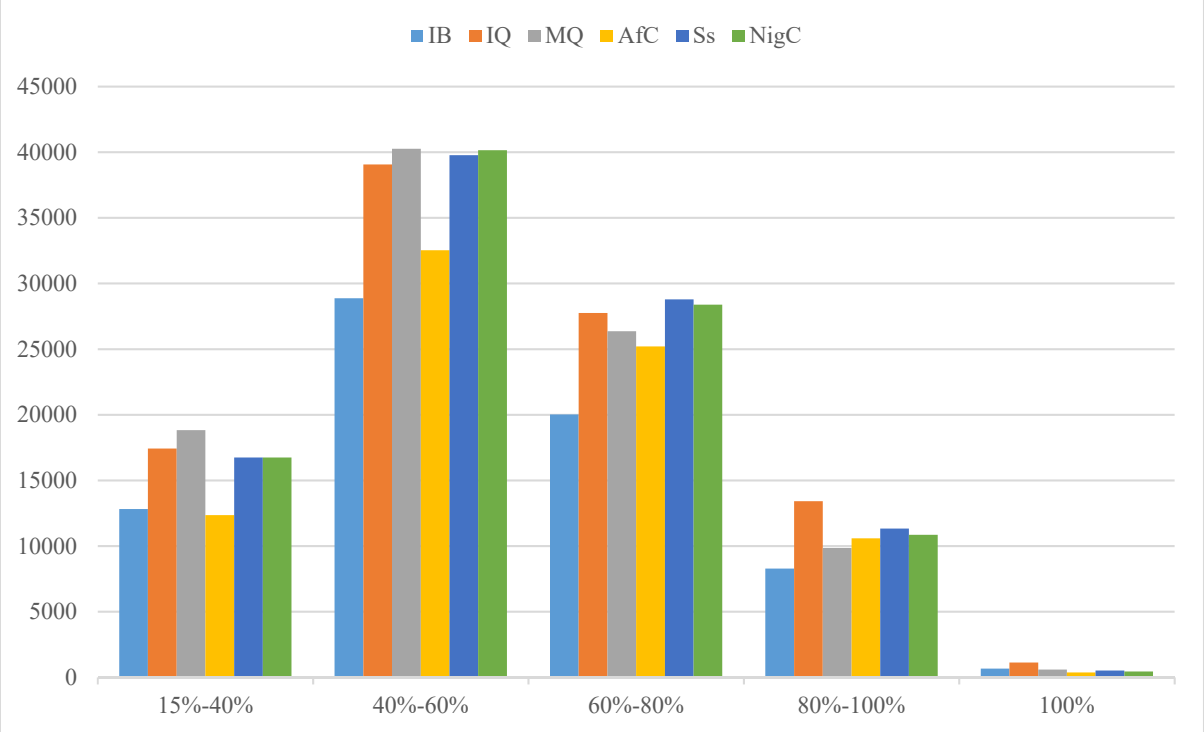


Figure 4.3: Sequence similarity distribution of top BLASTx hits for all samples. Y-axis represents the number of BLASTx top hits and the X-axis represents the percentage identities.

Species distribution of BLASTx top hits for all samples is shown in Figure 4.4. The highest percentage (23.06% (IB), 20.40% (IQ), 22.56% (MQ), 23.82% (AfC), 23.15% (Ss), 23.48% (NigC) of the contigs were matched to termite (*Zootermopsis nevadensis*). Besides the *Z. nevadensis*, the greater number of contigs matched to other insect species were *Acyrthosiphon pisum*, *Diaphorina citri*, *Tribolium castaneum*, *Pediculus humanus* and *Athalia rosae.*

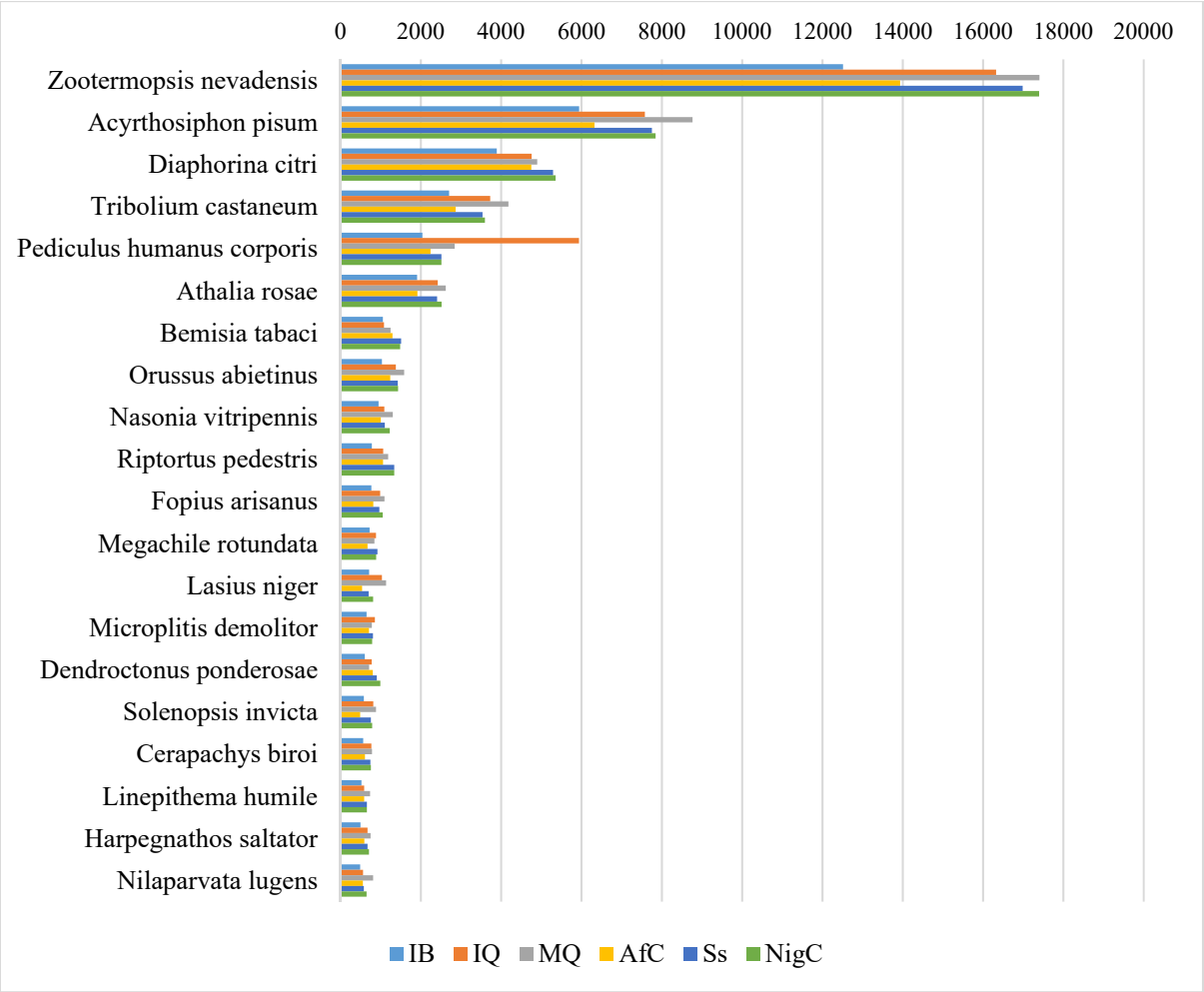
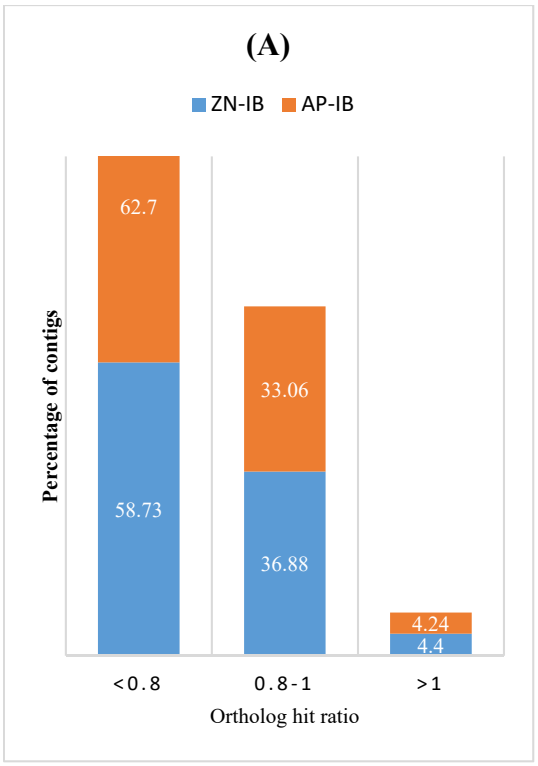
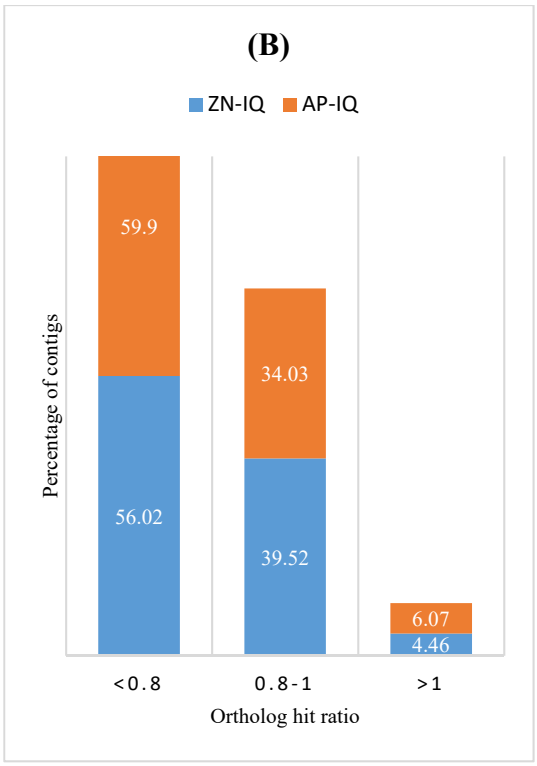
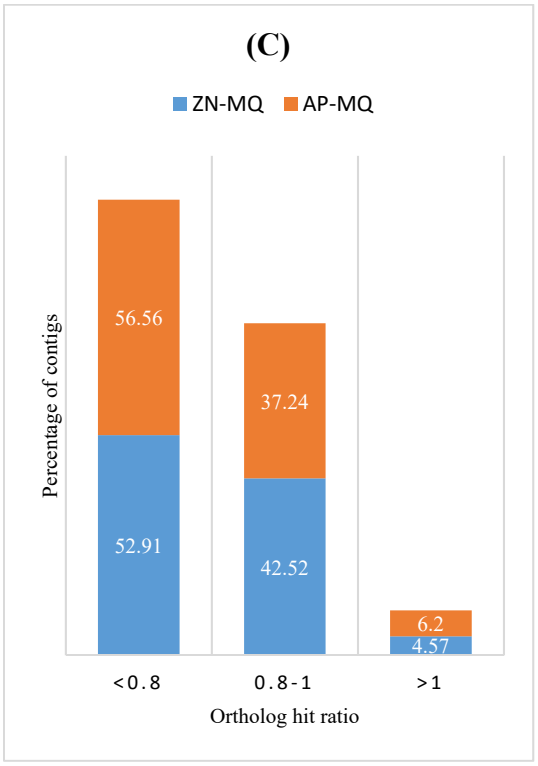
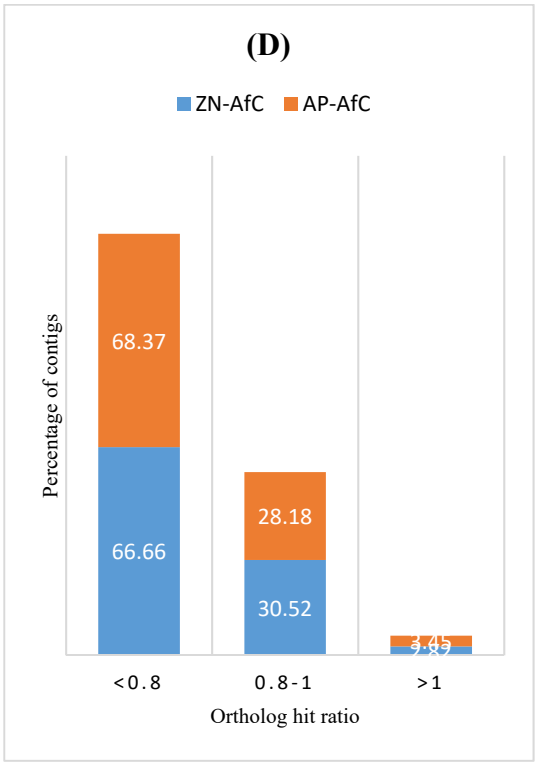


Figure 4.4: Top hit species distribution of the total number of homologous sequences matched with an e-value cut-off of 1.0E-3 against NR database. The horizontal bar represents the total number of hits found in each species.


(A)
ZN-IB
AP-IB
62.7
58.73
33.06
36.88
4.24
4.4
Percentage of contigs
<0.8
0.8-1
>1
Ortholog hit ratio


(B)
ZN-IQ
AP-IQ
59.9
56.02
34.03
39.52
6.07
4.46
Percentage of contigs
<0.8
0.8-1
>1
Ortholog hit ratio


(C)
ZN-MQ
AP-MQ
56.56
52.91
37.24
42.52
6.2
4.57
Percentage of contigs
<0.8
0.8-1
>1
Ortholog hit ratio


(D)
ZN-AfC
AP-AfC
68.37
66.66
28.18
30.52
3.45
Percentage of contigs
<0.8
0.8-1
>1
Ortholog hit ratio

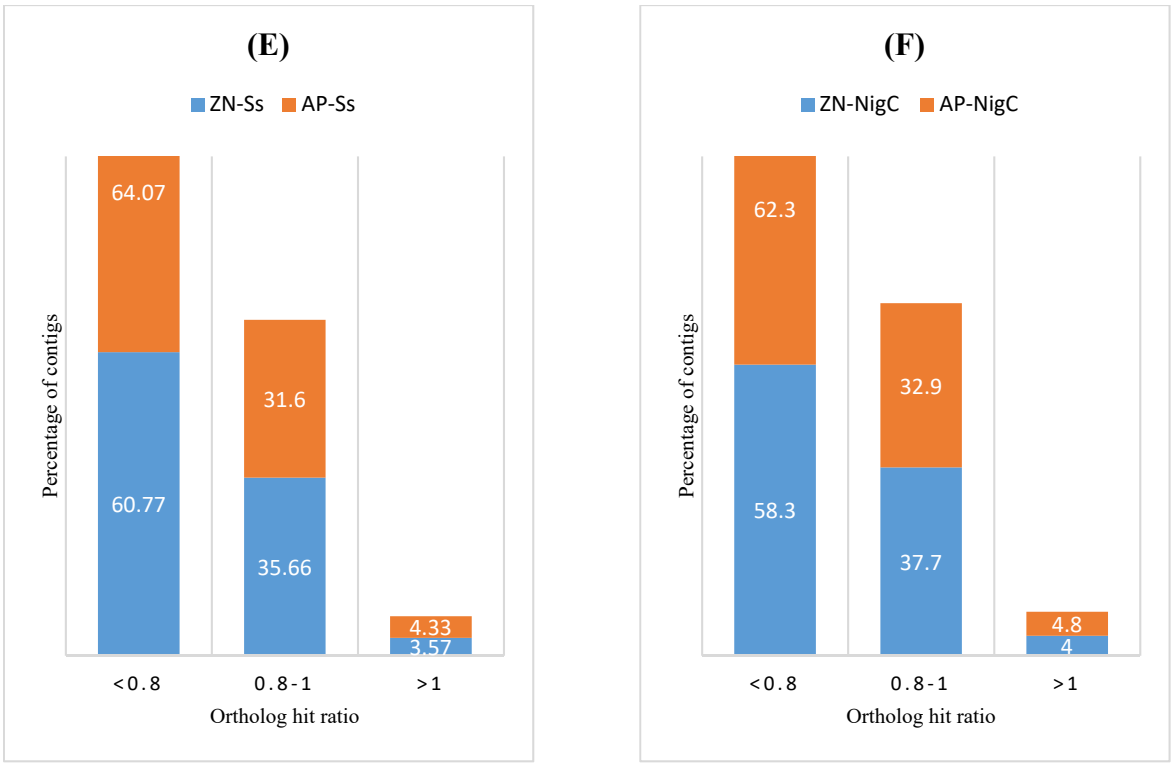
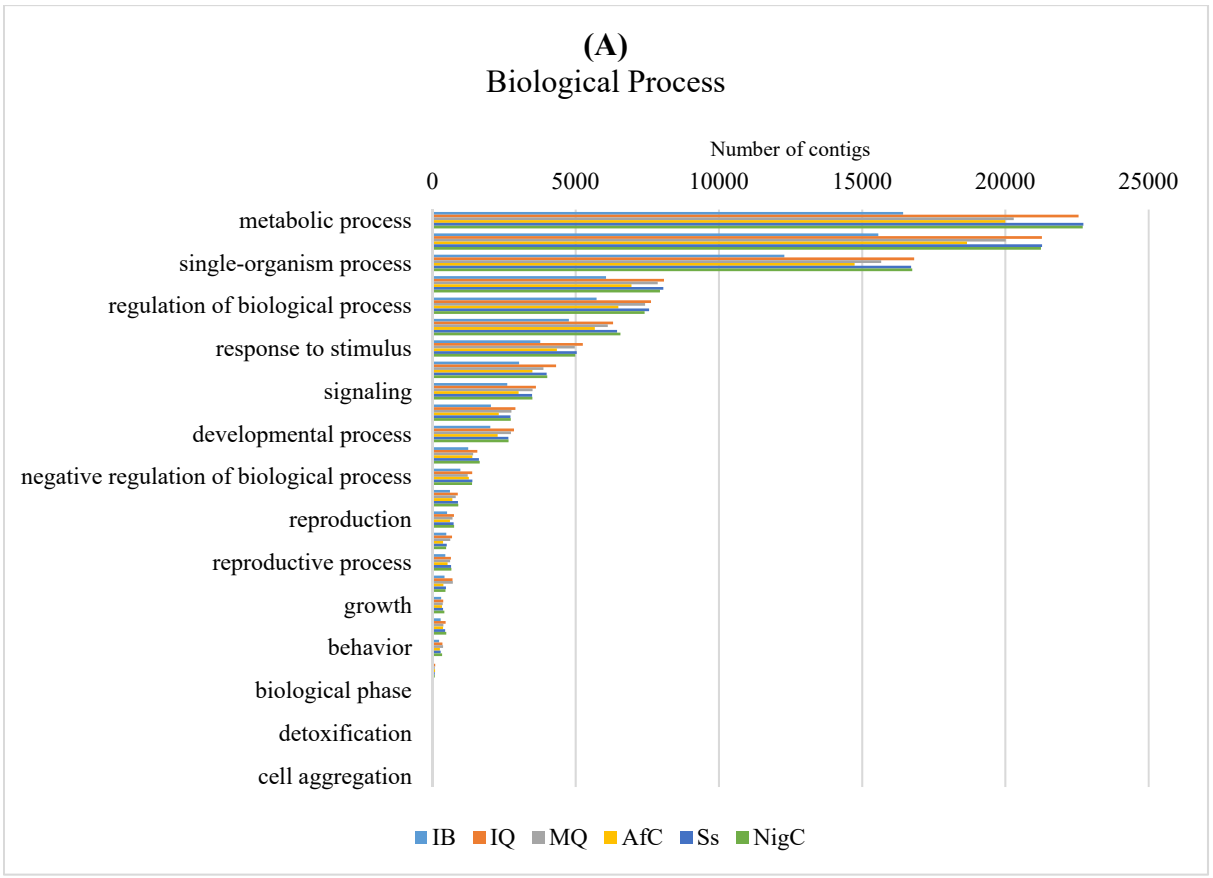


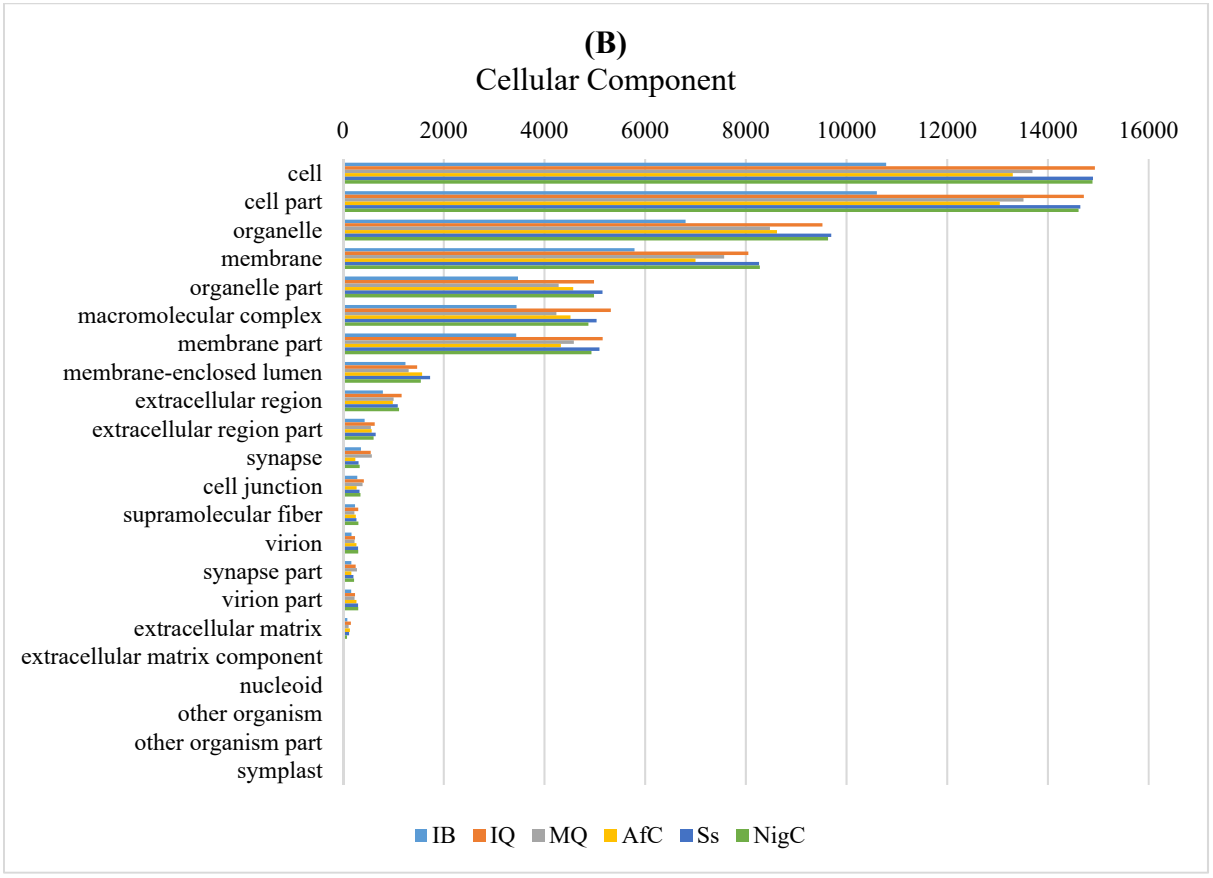
Figure 4.5: Comparison between *Z. nevadensis* and *A. pisum* using ortholog hit ratio shows the percentage of contigs that are likely to be fully assembled. Here, we have calculated the ortholog hit ratio of total hits obtained for *Z. nevadensis* (ZN) and *A. pisum* (AP) to identify what percentage of contigs are matching against each organism. Ratio 1.0 indicates fully assembled whereas ratios <1.0 indicates partially assembled contigs. Ratios >1.0 indicates insertions within contigs. A) Ortholog hit ratio scores for sample IB. B) Ortholog hit ratio scores for sample IQ. C) Ortholog hit ratio scores for sample MQ. D) Ortholog hit ratio scores for sample AfC. E) Ortholog hit ratio scores for sample Ss. F) Ortholog hit ratio scores for sample Ss.

4.3.3 Gene Ontology (GO) classification and pathway analysis

The *B. tabaci* transcriptome was annotated using GO terms based on BLASTx hits. GO is an international classification system to describe the functions and properties of genes in any organism. A total of 164,599 (IB), 228,062 (IQ), 210,474 (MQ), 197,546 (AfC), 225,881 (Ss) and 225,058 (NigC) GO terms were used to categorized the functions of predicted *B. tabaci* proteins into three main categories and 63 level-2 sub-categories, i.e., biological process (27 sub-categories), molecular function (14 sub-categories) and cellular component (22 sub-categories).

A frequency distribution of the number of contigs mapped to biological process, molecular function and cellular component is depicted in Figure 4.6. The majority of the GO terms were assigned to biological process, followed by cellular component and molecular function. Of these, the highly represented terms for biological process were metabolic process (20.54% (IB), 20.56% (IQ), 19.76% (MQ), 21.18% (AfC), 20.98% (Ss), and 20.97% (NigC)) and cellular process (19.46% (IB), 19.40% (IQ), 19.49% (MQ), 19.76% (AfC), 19.65% (Ss), and 19.62% (NigC)). Within molecular function category, catalytic activity (42.70% (IB), 42.64% (IQ), 42.74% (MQ), 43.70% (AfC), 43.57% (Ss), and 43.55% (NigC)) and binding (42.66% (IB), 42.46% (IQ), 42.68% (MQ), 42.13% (AfC), 42.29% (Ss), and 42.48% (NigC)) prominently represented GO terms, while the dominant GO terms for cellular component were cell (22.34% (IB), 21.92% (IQ), 22.34% (MQ), 22.11% (AfC), 21.87% (Ss), and 22.09% (NigC)), cell part, organelle and membrane. Figure 4.7 showed that contigs with smaller length were annotated with most number of GO terms compared to longer contigs.


(A)
Biological Process
Number of contigs
0
5000
10000
15000
20000
25000
metabolic process
single-organism process
regulation of biological process
response to stimulus
signaling
developmental process
negative regulation of biological process
reproduction
reproductive process
growth
behavior
biological phase
detoxification
cell aggregation
IB
IQ
MQ
AfC
Ss
NigC


(B)
Cellular Component
0
2000
4000
6000
8000
10000
12000
14000
16000
cell
cell part
organelle
membrane
organelle part
macromolecular complex
membrane part
membrane-enclosed lumen
extracellular region
extracellular region part
synapse
cell junction
supramolecular fiber
virion
synapse part
virion part
extracellular matrix
extracellular matrix component
nucleoid
other organism
other organism part
symplast
IB
IQ
MQ
AfC
Ss
NigC

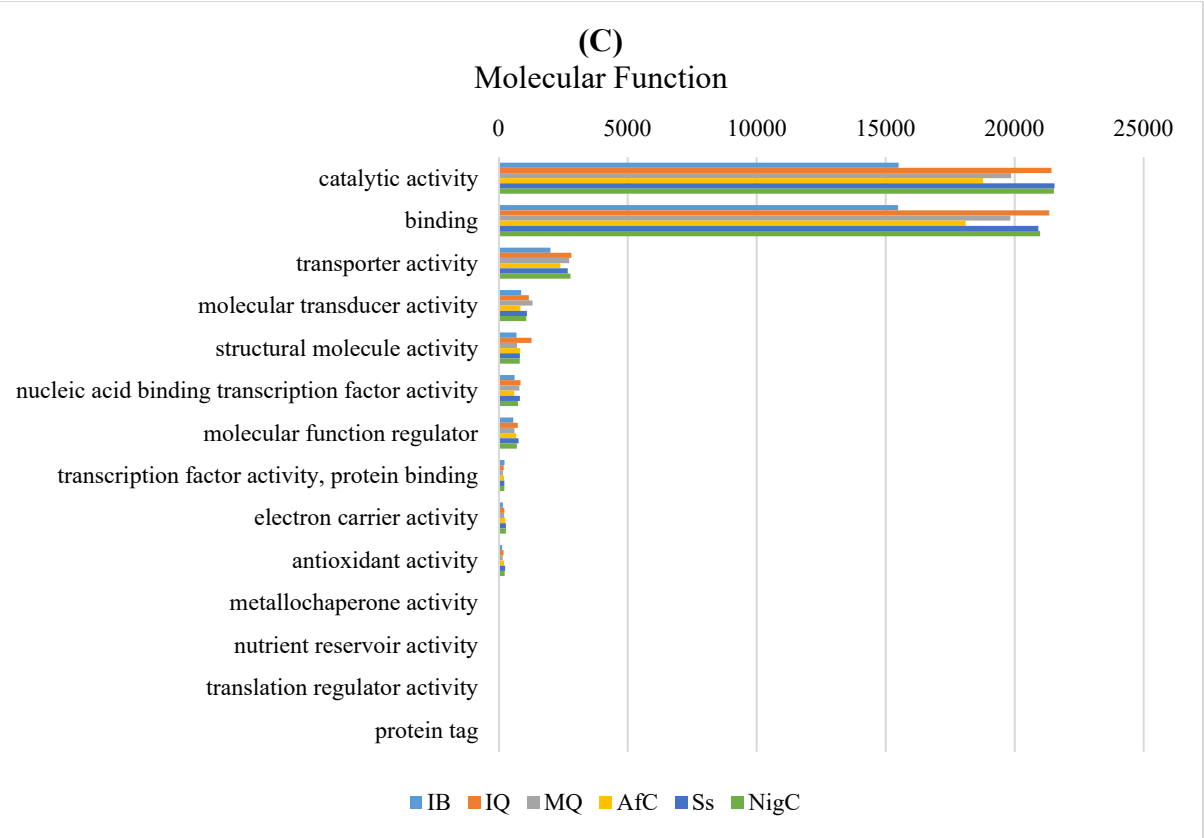
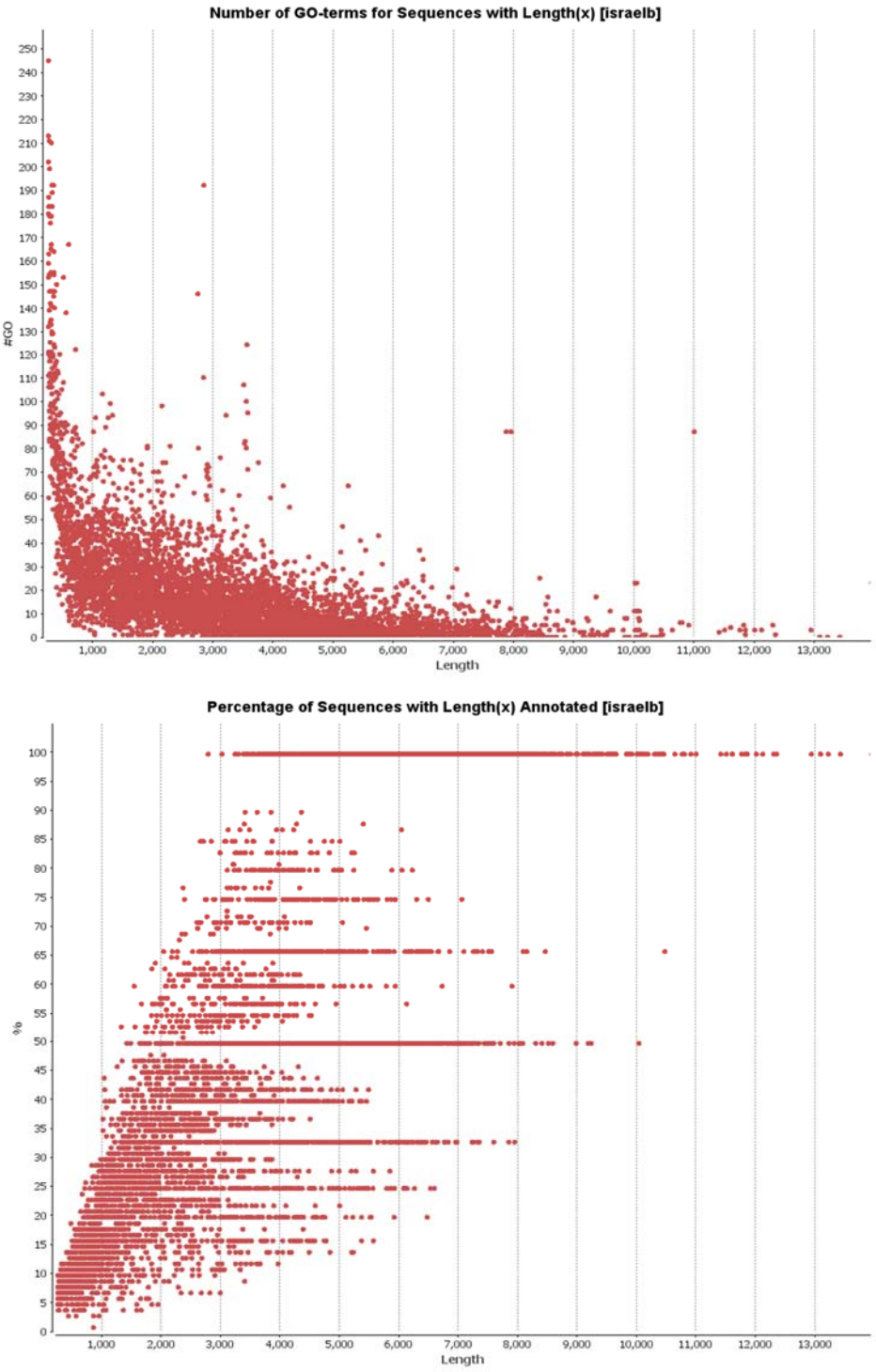
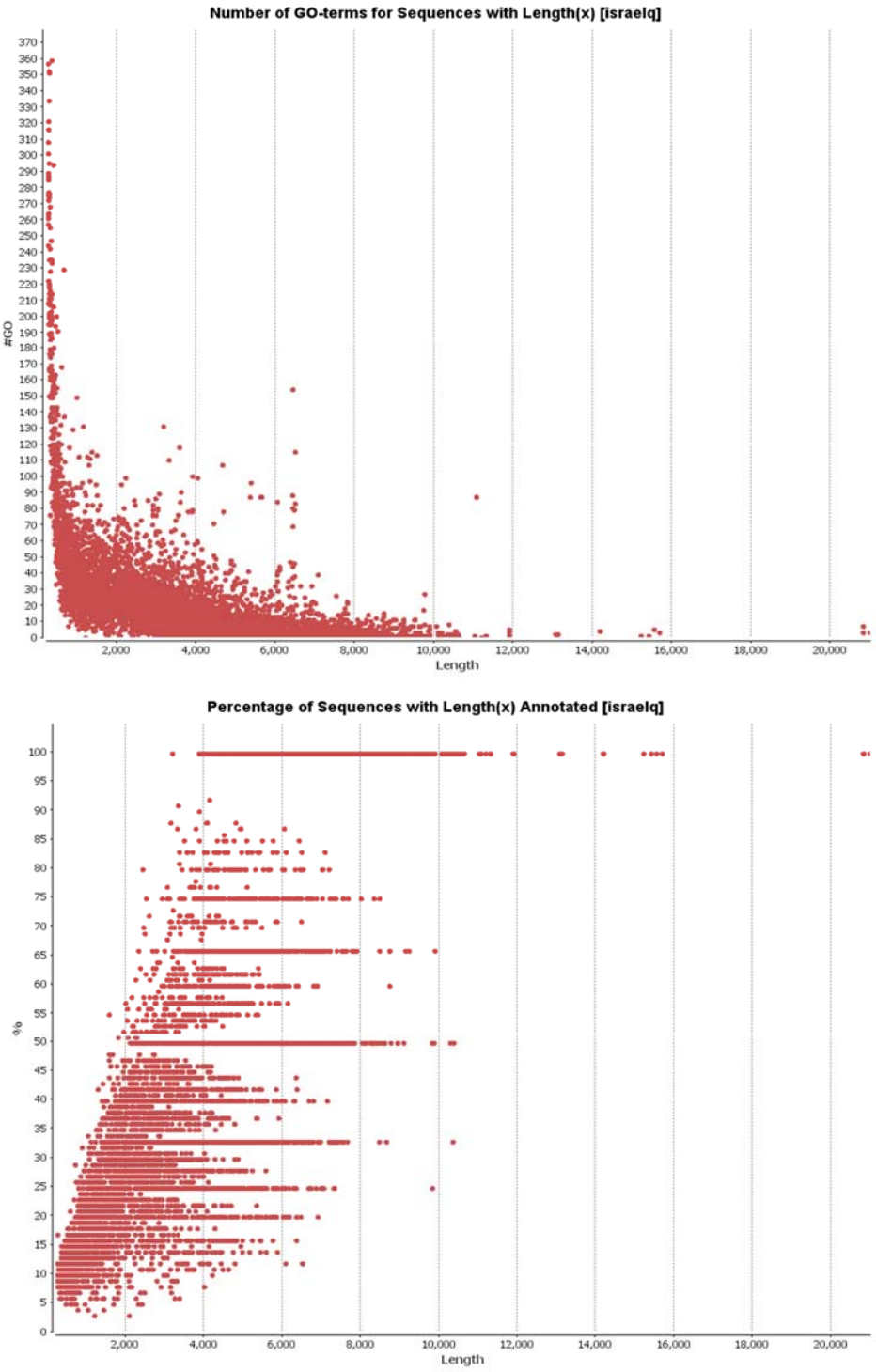


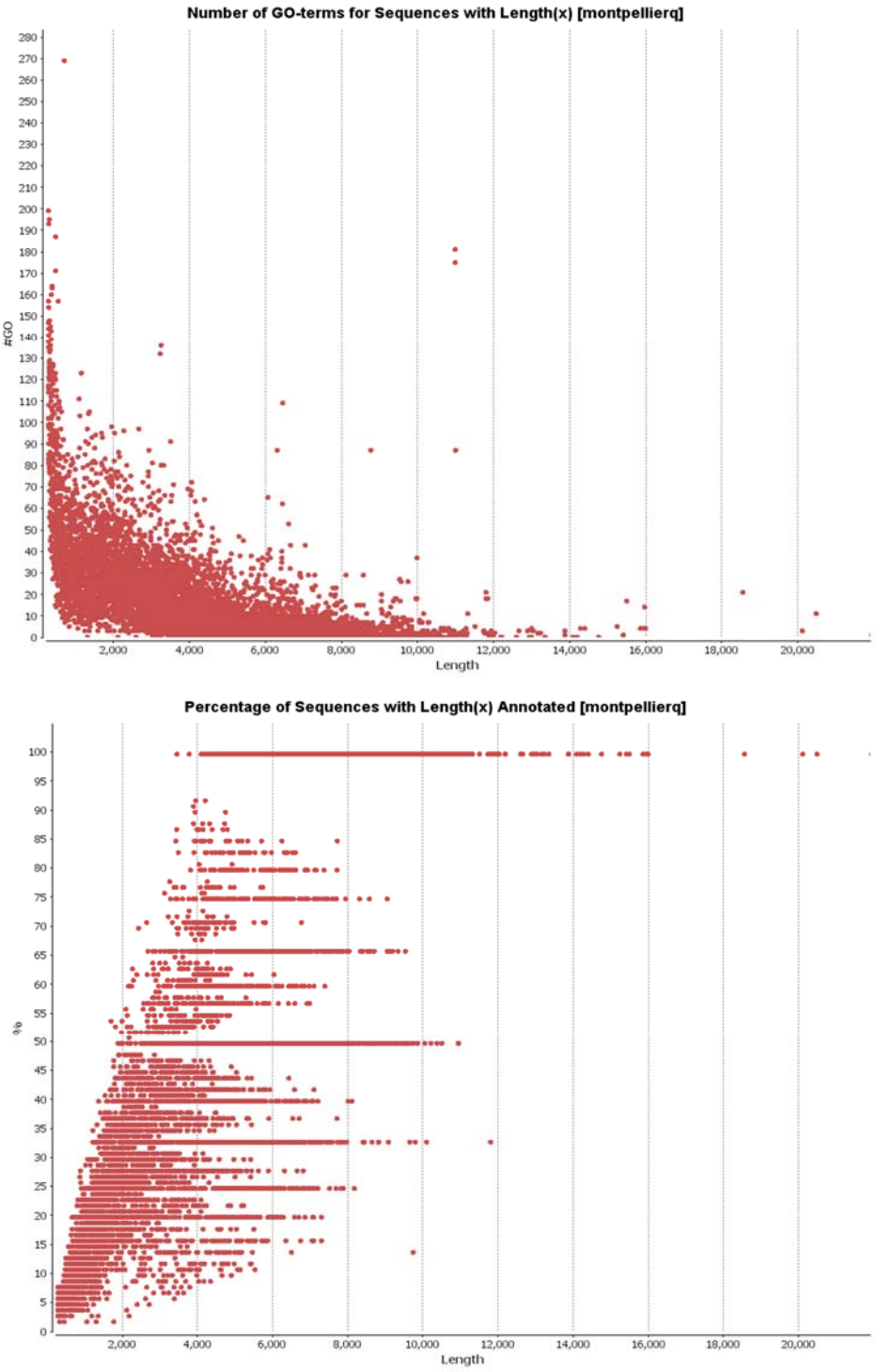
Figure 4.6: Histogram distribution of GO functional categories of samples IB, IQ, MQ, AfC, Ss and NigC. The results were categorized based on biological process (A), cellular component (B) and molecular function (C) and were further grouped into level-2 sub-categories on X-axis with number of contigs present in each sub-category on Y-axis.


Number of GO-terms for Sequences with Length(x) [israelb]
#GO
Length
Percentage of Sequences with Length(x) Annotated [israelb]
%
Length

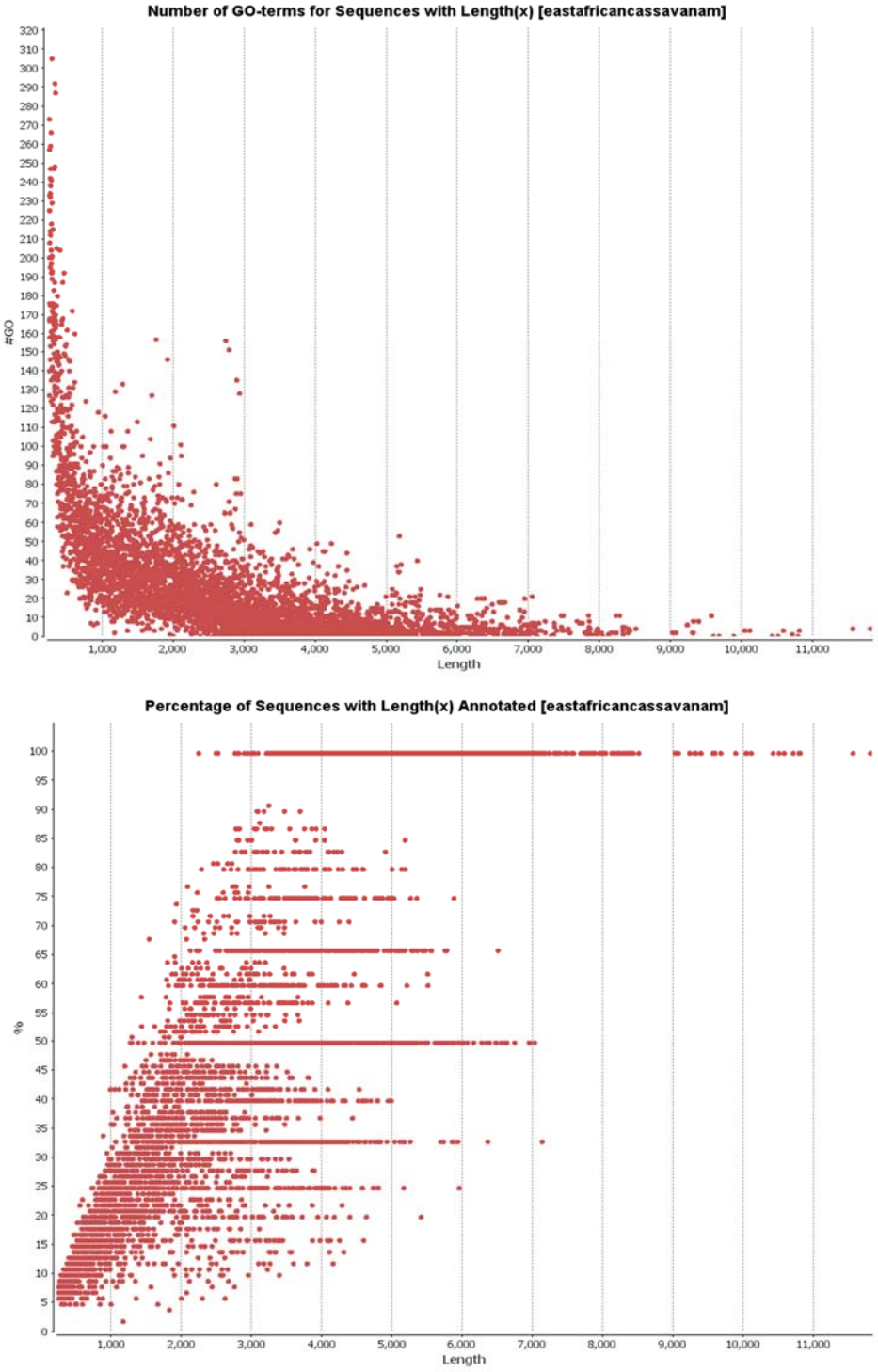
(A)


Number of GO-terms for Sequences with Length(x) [israelq]
#GO
Length
Percentage of Sequences with Length(x) Annotated [israelq]
%
Length

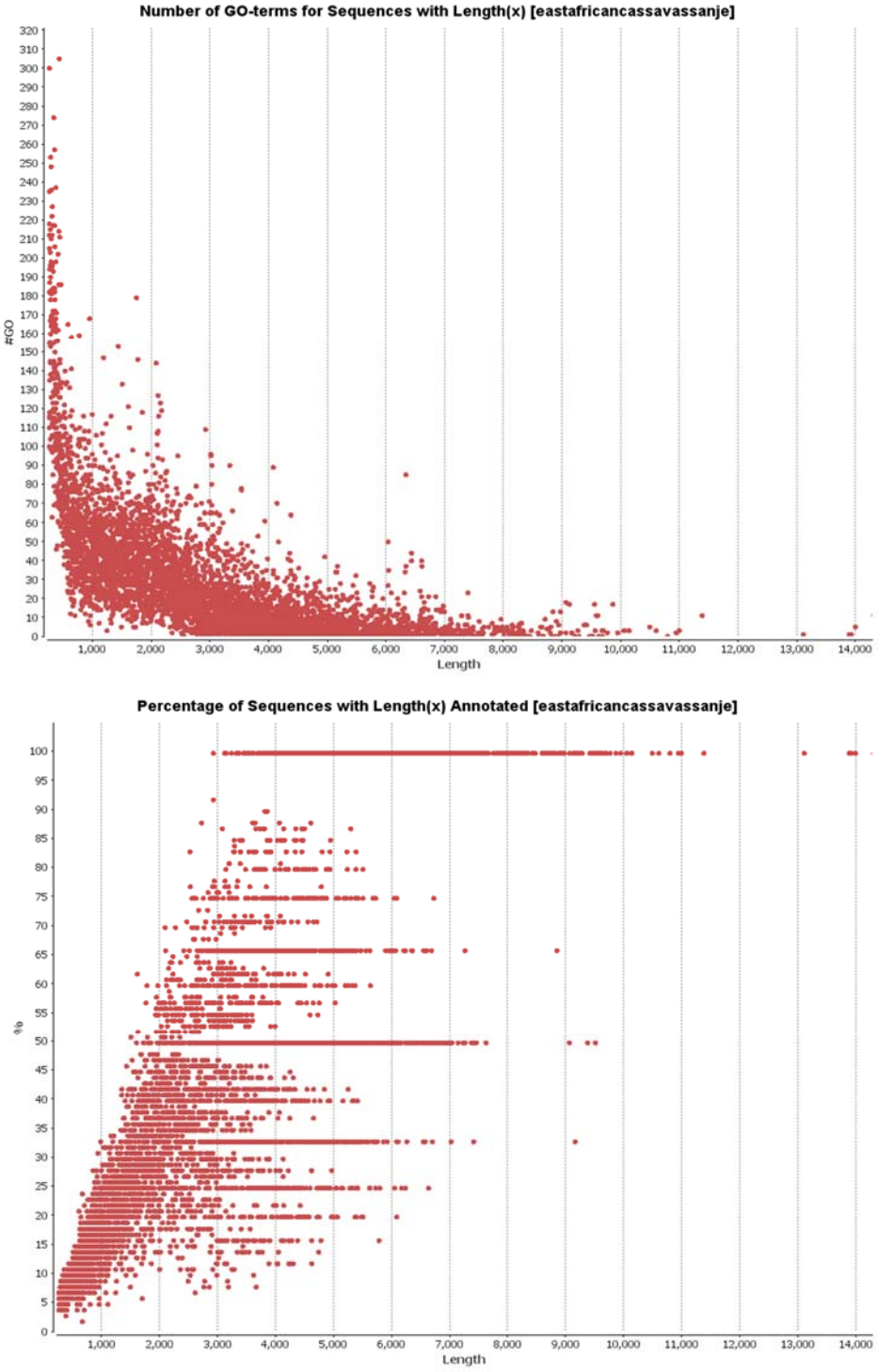
(B)


Number of GO-terms for Sequences with Length(x) [montpellierq]
#GO
Length
Percentage of Sequences with Length(x) Annotated [montpellierq]
%
Length

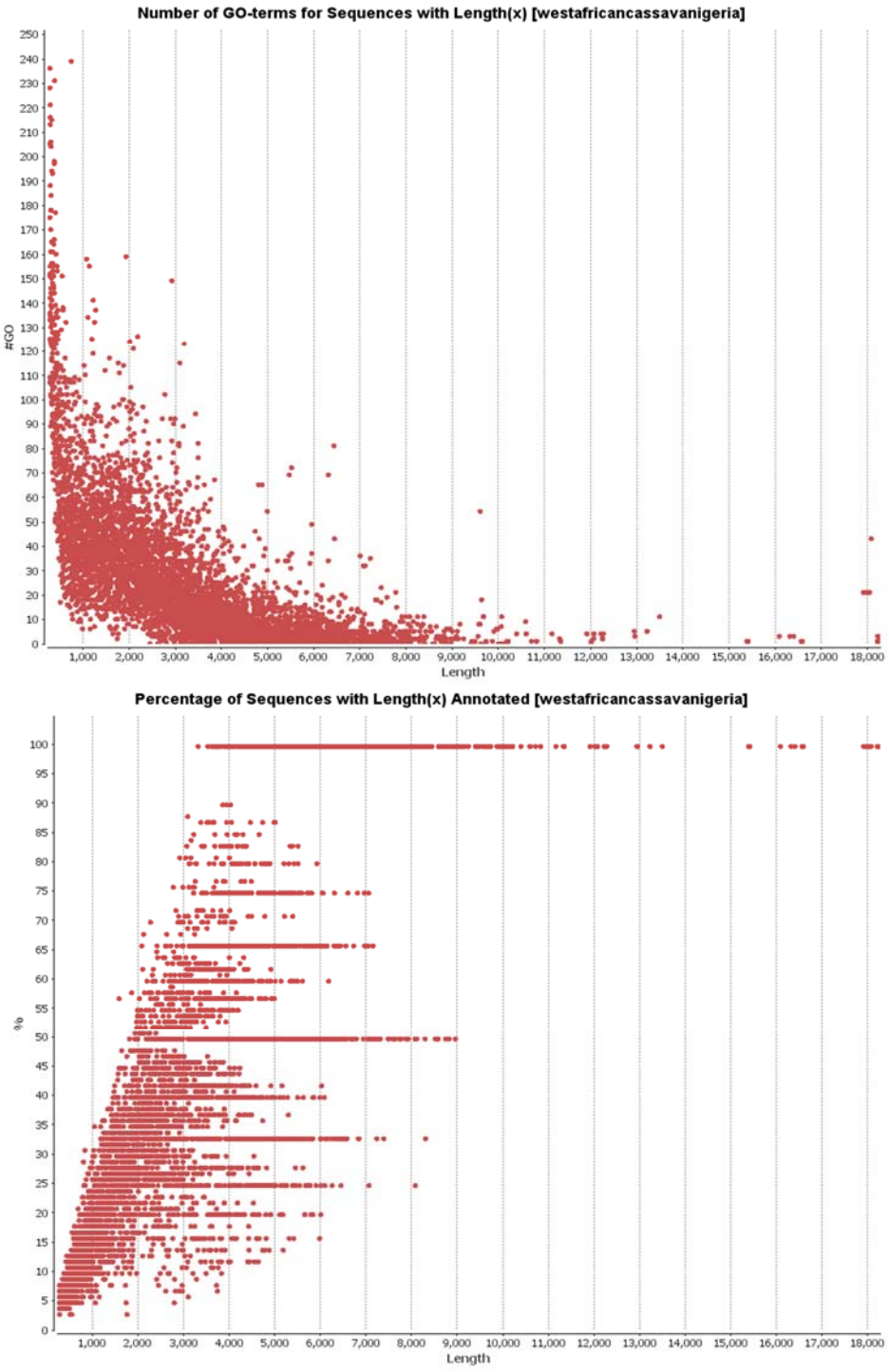
(C)


Number of GO-terms for Sequences with Length(x) [eastafricancassavanam]
#GO
Length
Percentage of Sequences with Length(x) Annotated [eastafricancassavanam]
%
Length

(D)


Number of GO-terms for Sequences with Length(x) [eastafricancassavassanje]
#GO
Length
Percentage of Sequences with Length(x) Annotated [eastafricancassavassanje]
%
Length

(E)



(F)

Figure 4.7: Length distribution of contigs annotated with GO terms. The graph also shows the percentage of contigs annotated with length X. Here, A, B, C, D, E and F represents *B. tabaci* populations IB, IQ, MQ, AfC, Ss and NigC respectively.

4.3.4 Biological pathway and enzyme classification of *B. tabaci*

To identify the potential contigs involved in biological pathways in the *B. tabaci* samples, annotated contigs were searched against Kyoto Encyclopedia of Genes and Genomes (KEGG) database. A total of 12,557 (IB), 18,894 (IQ), 15,333 (MQ), 17,432 (AfC), 20,304 (Ss) and 18,775 (NigC) contigs were assigned to five main categories in KEGG database, i.e., metabolism, genetic information processing, environmental information processing, organismal systems and human diseases. A total of 128 (IB), 139 (IQ), 128 (MQ), 132 (AfC), 131 (Ss) and 132 (NigC) pathways were predicted using Blast2GO annotation tool. Among the 5 main categories, metabolism represented 93.27% (IB), 94.09% (IQ), 93.61% (MQ), 95.09% (AfC), 95.20% (Ss) and 94.76% (NigC) as shown in Figure 4.8. Within the metabolism, nucleotide metabolism represented most number of contigs (2,377 (IB), 3,354 (IQ), 2,848 (MQ), 3,197 (AfC), 3,545 (Ss) and 3,397 (NigC)) followed by metabolism of cofactors (2,031 (IB), 3,034 (IQ), 2,565 (MQ), 2,839 (AfC), 3,215 (Ss) and 3,088 (NigC) and vitamins and carbohydrate metabolism. In contrast, immune system of organismal systems and drug resistance of human diseases only matched 458 (IB), 645 (IQ), 505 (MQ), 415 (AfC), 490 (Ss), 486 (NigC) and 2 (IB), 3 (IQ), 1 (MQ), 1 (AfC), 2 (Ss), 2 (NigC) respectively.

By searching against the enzyme database using the Blast2Go, a total of 8,927 (IB), 12,616 (IQ), 10,935 (MQ), 11,292 (AfC), 12,652 (Ss) and 12,308 (NigC) contigs were annotated with enzyme codes. The annotated contigs were classified into main six categories: oxidoreductases (10,302 contigs), transferases (18,174 contigs), hydrolases (31,578 contigs), lyases (2,763 contigs), isomerases (2,029 contigs) and ligases (3,884 contigs) (Figure 4.9). The most abundant enzyme types were acting on acid anhydrides of hydrolases with total number of contigs present in sample IB were 1,674, 2,443 for IQ, 2,028 for MQ, 2,333 for AfC, 2,570 for Ss and 2,436 for NigC. The second most abundant enzyme sub-class belonged to transferring phosphorous-containing groups in class transferases.

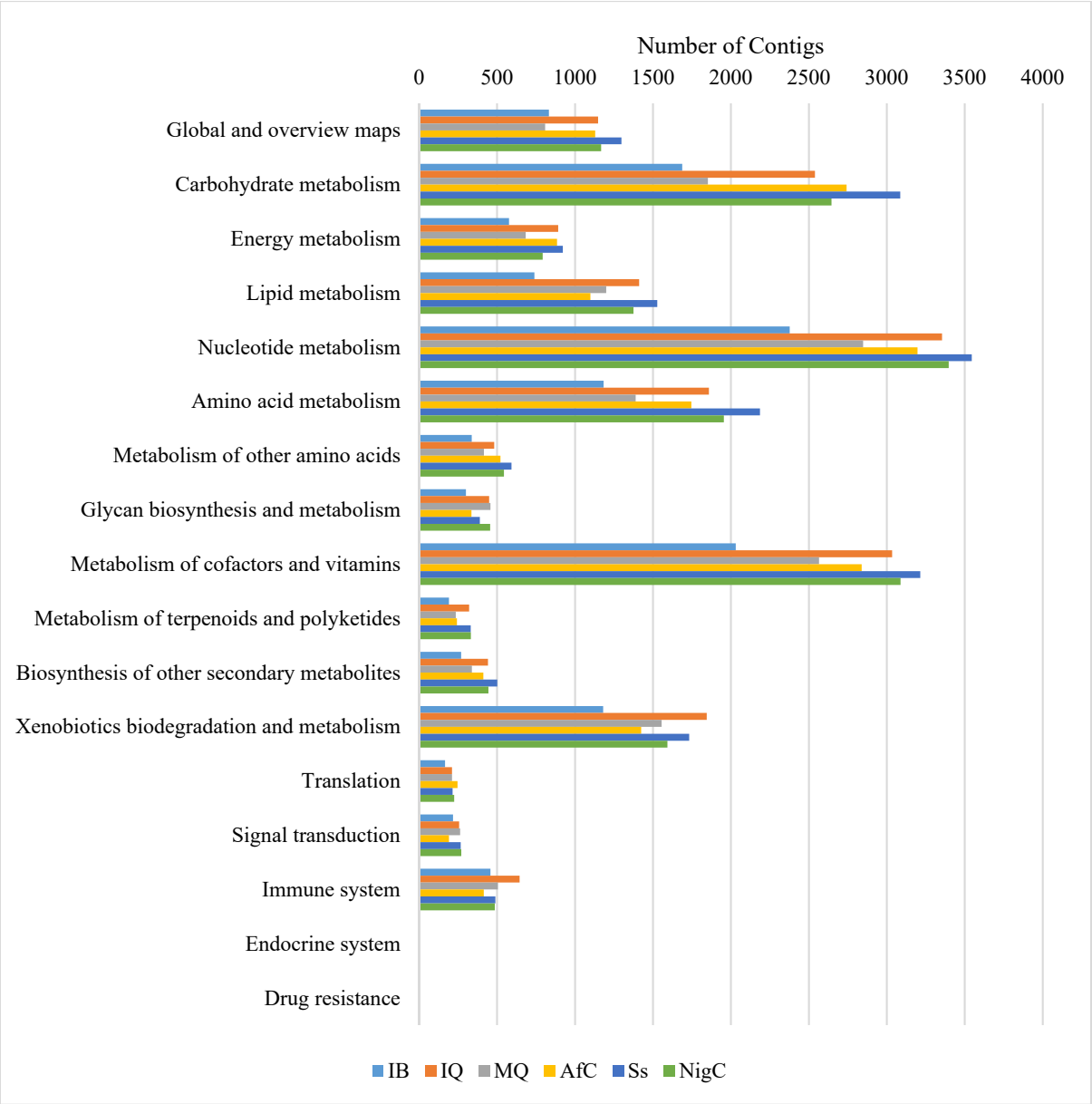
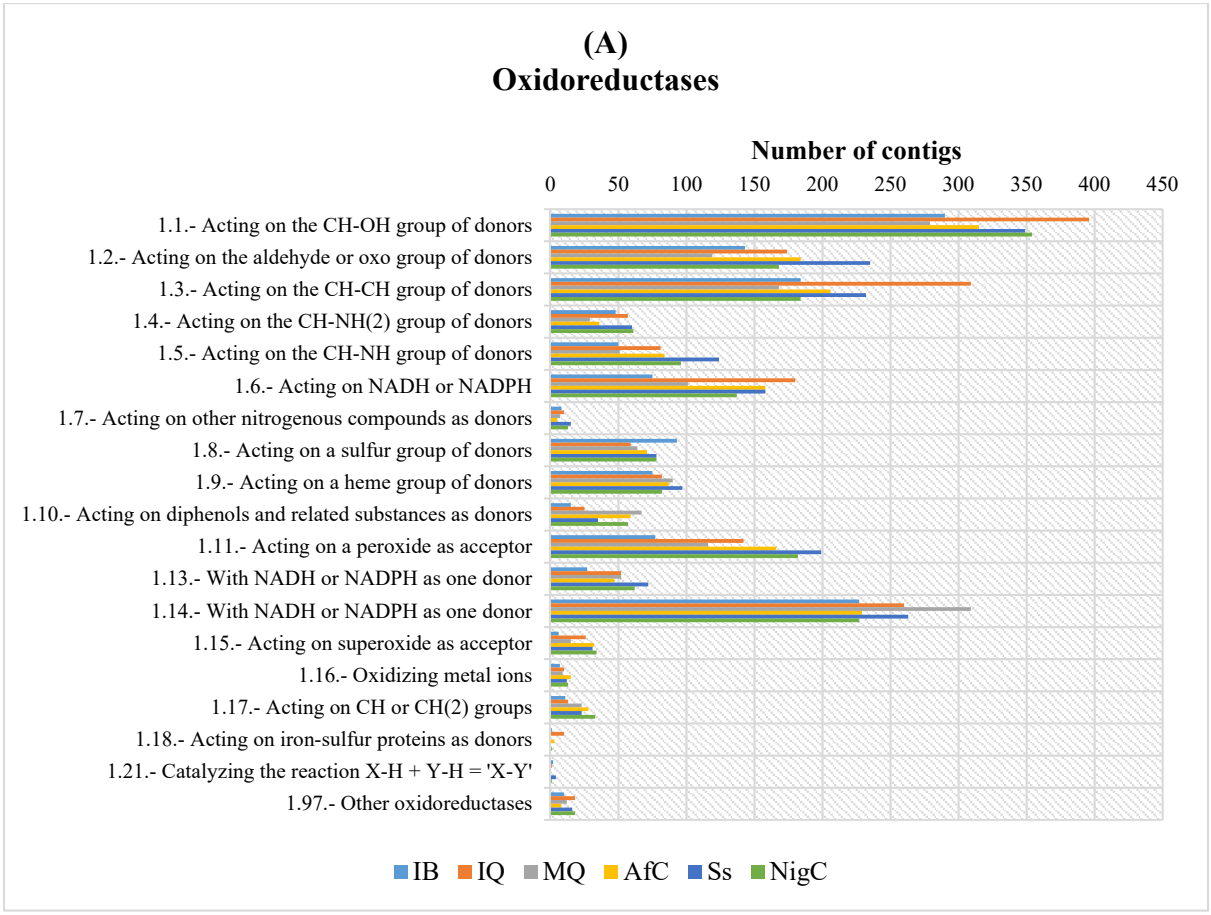
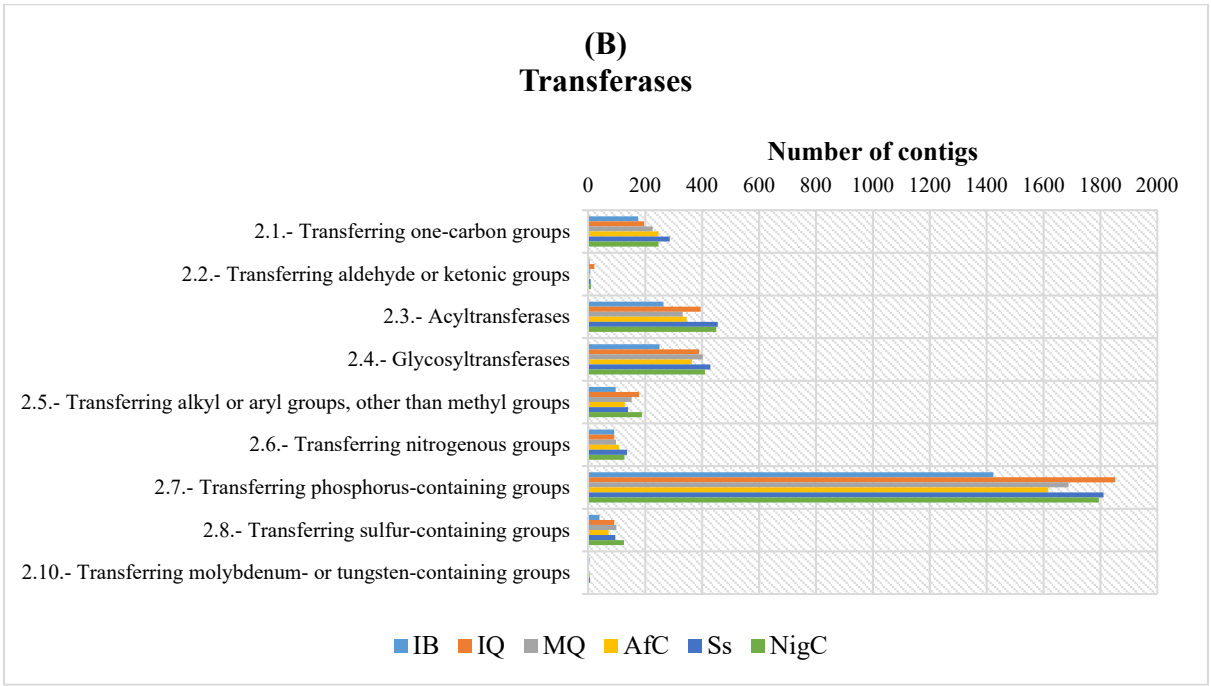
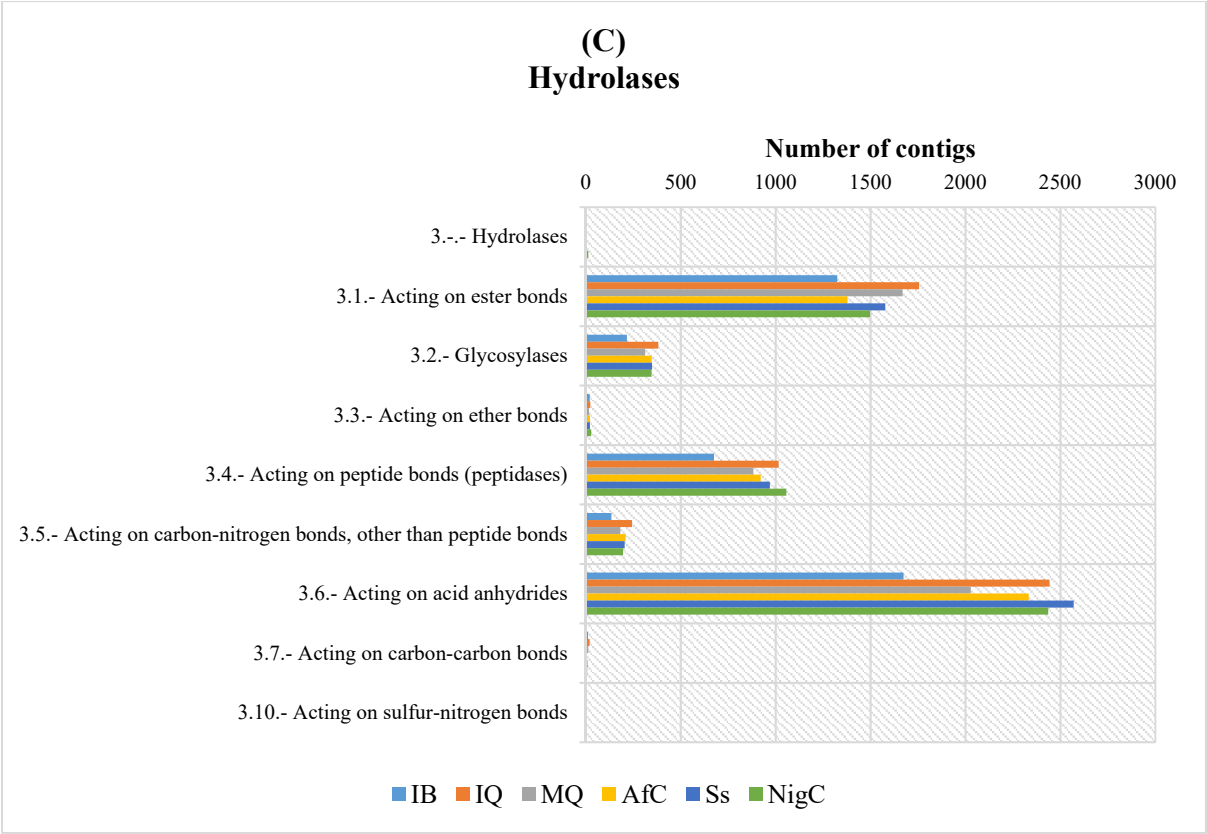
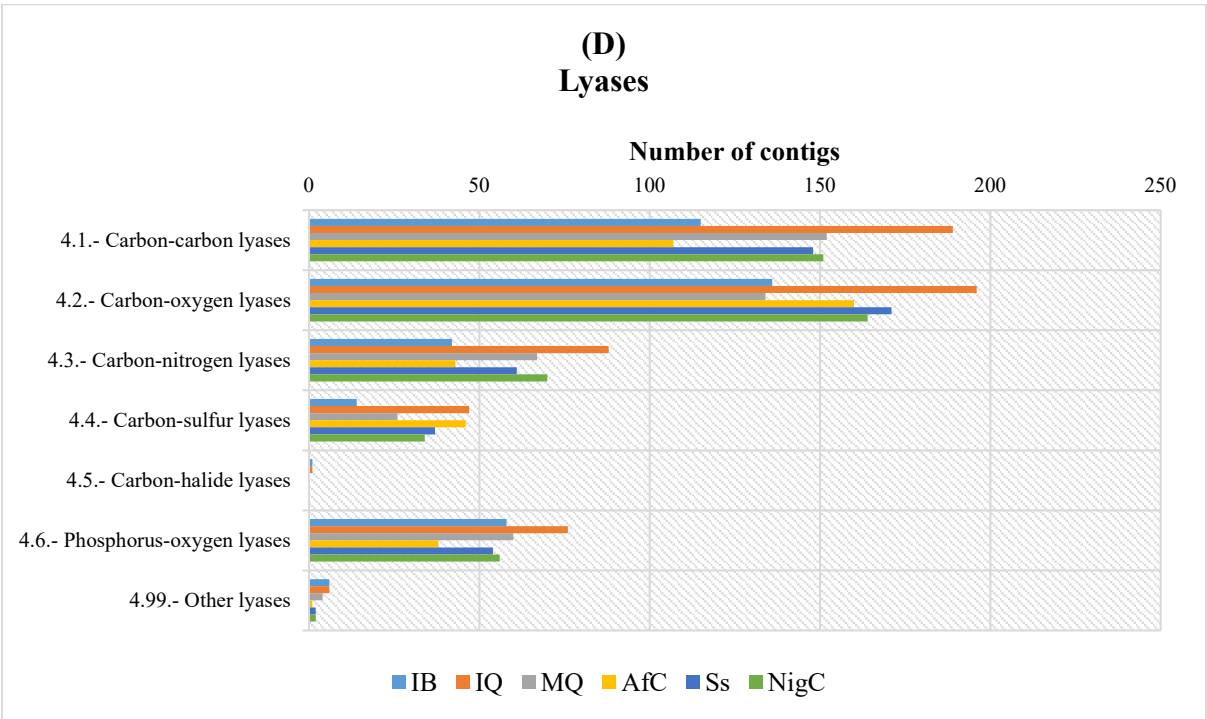


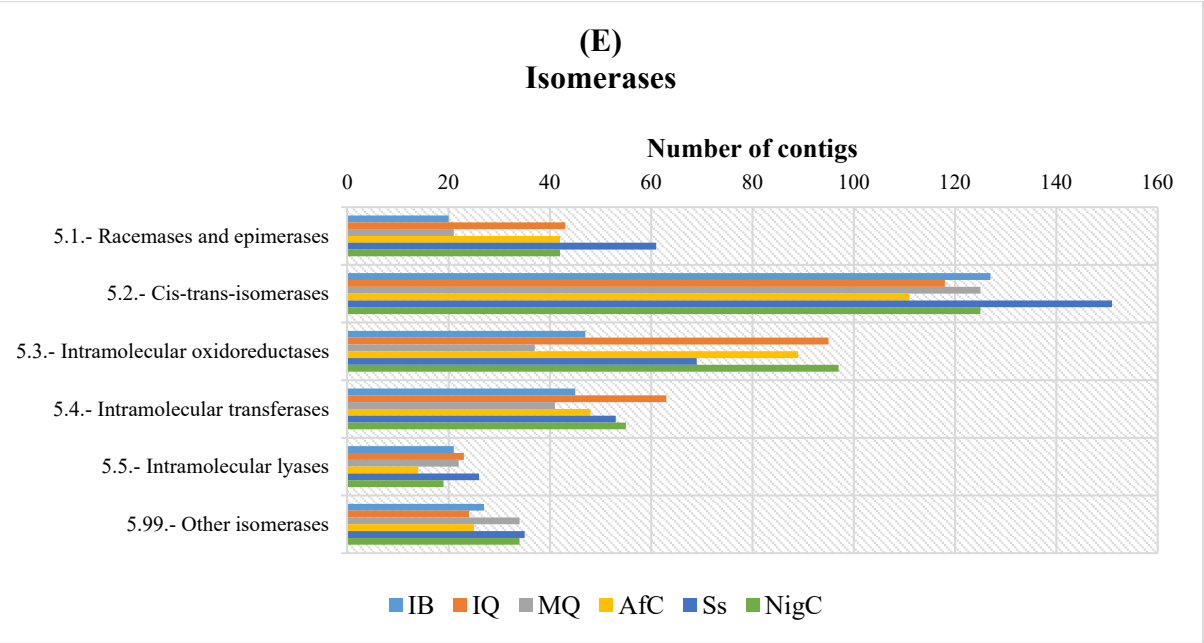
Figure 4.8: Analysis of KEGG pathway annotations of samples IB, IQ, MQ, AfC, Ss and NigC. Total number of pathways present in each sample were divided into main categories and were further divided into sub-categories with the number of contigs present in each sub-category.


(A)
Oxidoreductases
Number of contigs
0
50
100
150
200
250
300
350
400
450
1.1.- Acting on the CH-OH group of donors
1.2.- Acting on the aldehyde or oxo group of donors
1.3.- Acting on the CH-CH group of donors
1.4.- Acting on the CH-NH(2) group of donors
1.5.- Acting on the CH-NH group of donors
1.6.- Acting on NADH or NADPH
1.7.- Acting on other nitrogenous compounds as donors
1.8.- Acting on a sulfur group of donors
1.9.- Acting on a heme group of donors
1.10.- Acting on diphenols and related substances as donors
1.11.- Acting on a peroxide as acceptor
1.13.- With NADH or NADPH as one donor
1.14.- With NADH or NADPH as one donor
1.15.- Acting on superoxide as acceptor
1.16.- Oxidizing metal ions
1.17.- Acting on CH or CH(2) groups
1.18.- Acting on iron-sulfur proteins as donors
1.21.- Catalyzing the reaction X-H + Y-H = 'X-Y'
1.97.- Other oxidoreductases
IB
IQ
MQ
AfC
Ss
NigC


(B)
Transferases
Number of contigs
0
200
400
600
800
1000
1200
1400
1600
1800
2000
2.1.- Transferring one-carbon groups
2.2.- Transferring aldehyde or ketonic groups
2.3.- Acyltransferases
2.4.- Glycosyltransferases
2.5.- Transferring alkyl or aryl groups, other than methyl groups
2.6.- Transferring nitrogenous groups
2.7.- Transferring phosphorus-containing groups
2.8.- Transferring sulfur-containing groups
2.10.- Transferring molybdenum- or tungsten-containing groups
IB
IQ
MQ
AfC
Ss
NigC


(C)
Hydrolases
Number of contigs
0
500
1000
1500
2000
2500
3000
3.-.- Hydrolases
3.1.- Acting on ester bonds
3.2.- Glycosylases
3.3.- Acting on ether bonds
3.4.- Acting on peptide bonds (peptidases)
3.5.- Acting on carbon-nitrogen bonds, other than peptide bonds
3.6.- Acting on acid anhydrides
3.7.- Acting on carbon-carbon bonds
3.10.- Acting on sulfur-nitrogen bonds
IB
IQ
MQ
AfC
Ss
NigC


(D)
Lyases
Number of contigs
0
50
100
150
200
250
4.1.- Carbon-carbon lyases
4.2.- Carbon-oxygen lyases
4.3.- Carbon-nitrogen lyases
4.4.- Carbon-sulfur lyases
4.5.- Carbon-halide lyases
4.6.- Phosphorus-oxygen lyases
4.99.- Other lyases
IB
IQ
MQ
AfC
Ss
NigC



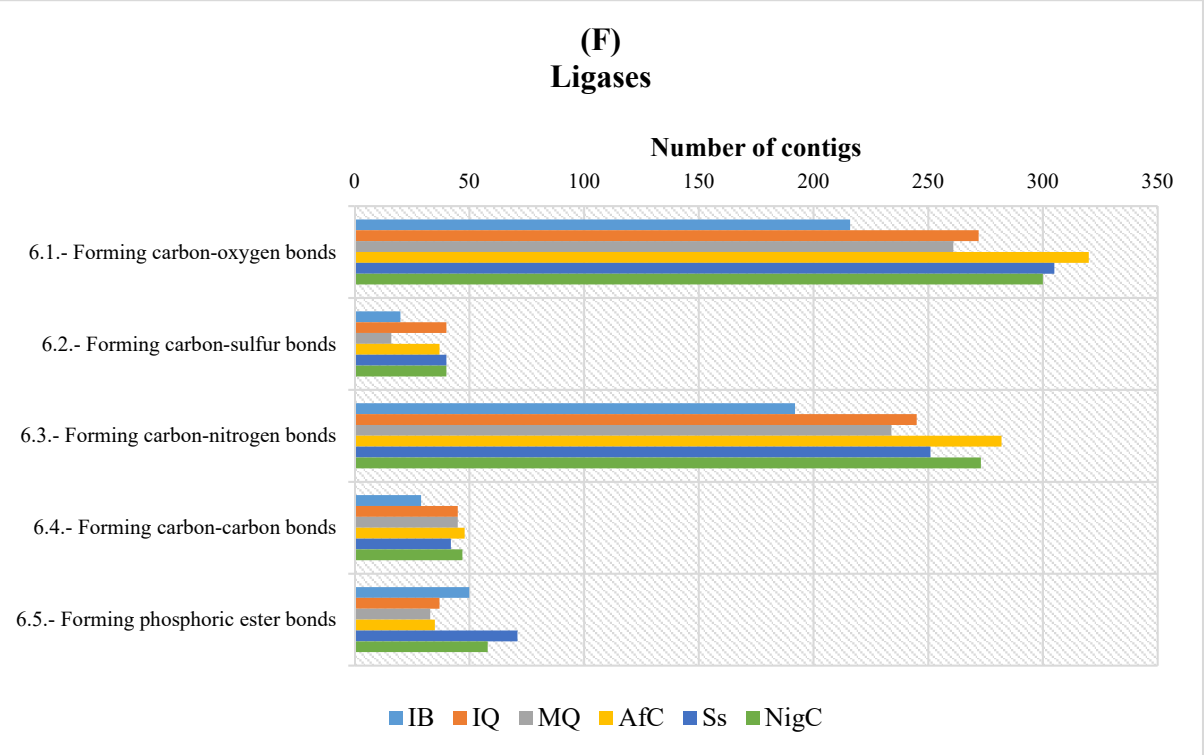


Figure 4.9: Classification of potential enzyme genes in six *B. tabaci* populations. Six main enzyme classes were further divided into sub-classes. A) Classification of enzyme genes into category oxidoreductases. B) Classification of enzyme genes into category transferases. C) Classification of enzyme genes into category hydrolases. D) Classification of enzyme genes into category lyases. E) Classification of enzyme genes into category isomerases. F) Classification of enzyme genes into category ligases.

4.3.5 Domain prediction

Using the hmmscan script for domain prediction, the total number of hits found were 5,190 in IB, 12,138 in IQ, 11,779 in MQ, 11,371 in AfC, 11,743 in Ss and 11,703 in NigC. There were 11 domains found in IB that were absent in IQ and MQ. Whereas, 1553 domains found in IQ that were absent in IB and MQ and there were 1197 domains found in MQ that were absent in IB and IQ. Figure 4.10 also shows that the common domains found between IQ and MQ were much higher in number compared to found common between IB and IQ and IB and MQ. While in the case of cassava colonizing group, 640 domains were found in AfC that were absent in Ss and NigC. Likewise, 718 domains were found in Ss that were absent in AfC and NigC and 702 domains were found in NigC that were absent in AfC and Ss.

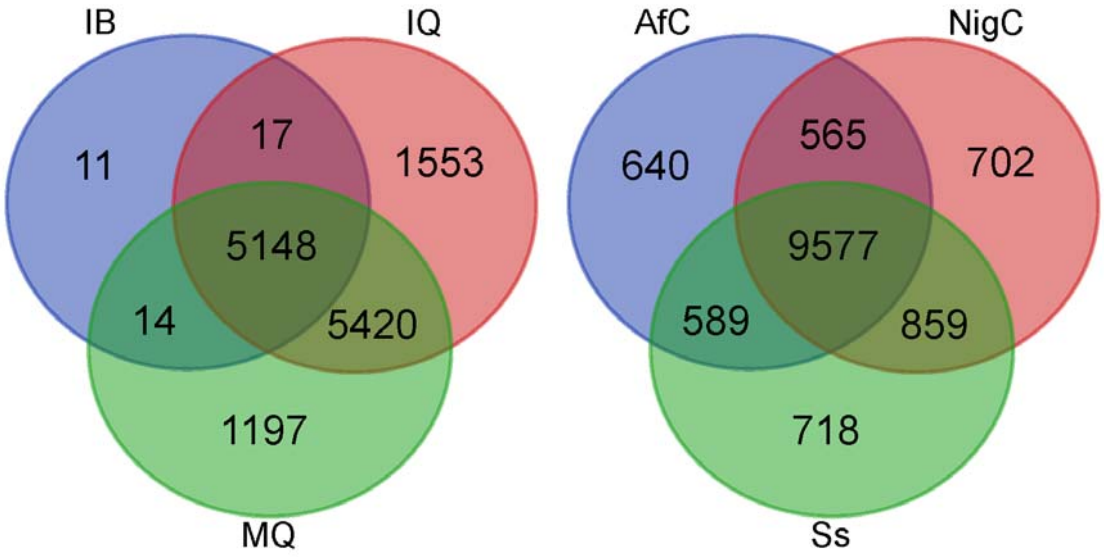


Figure 4.10: Venn diagram showing number of Pfam domains found in cassava and non-cassava colonizing *B. tabaci* populations.

4.3.6 Estimation of transcriptome completeness

To estimate the transcriptome completeness, each assembled contig and its best BLASTx hit were considered orthologs to calculate the "Ortholog hit ratio". For this study, hit region of the contig is considered as an estimator of "putative coding region". The ortholog hit ratio can be calculated by dividing the length of putative coding region of contig by the total length of ortholog matched against that contig (O'Neil *et al*., 2010). It is an important indicator of percentage of relative insertions present in both contigs and *B. tabaci* orthologs. The results as shown in Figure 4.11 indicate that 1613 (IB), 2000 (IQ), 2067 (MQ), 1837 (AfC), 2175 (Ss) and 2208 (NigC) transcripts were fully assembled based on "Ortholog hit ratio" of 1.0 (O'Neil *et al*., 2010), whereas a high number (IB (66,353), IQ (55,200), MQ (89,552), AfC (76,924), Ss (91,600), NigC (90,843)) contigs fall below 1.0. A ratio greater than 1.0 indicates large insertions in contig sequences. The total number of sequences found for ratios >1.0 were 2682 (IB), 3000 (IQ), 4276 (MQ), 2262 (AfC), 3382 (Ss) and 3518 (NigC).

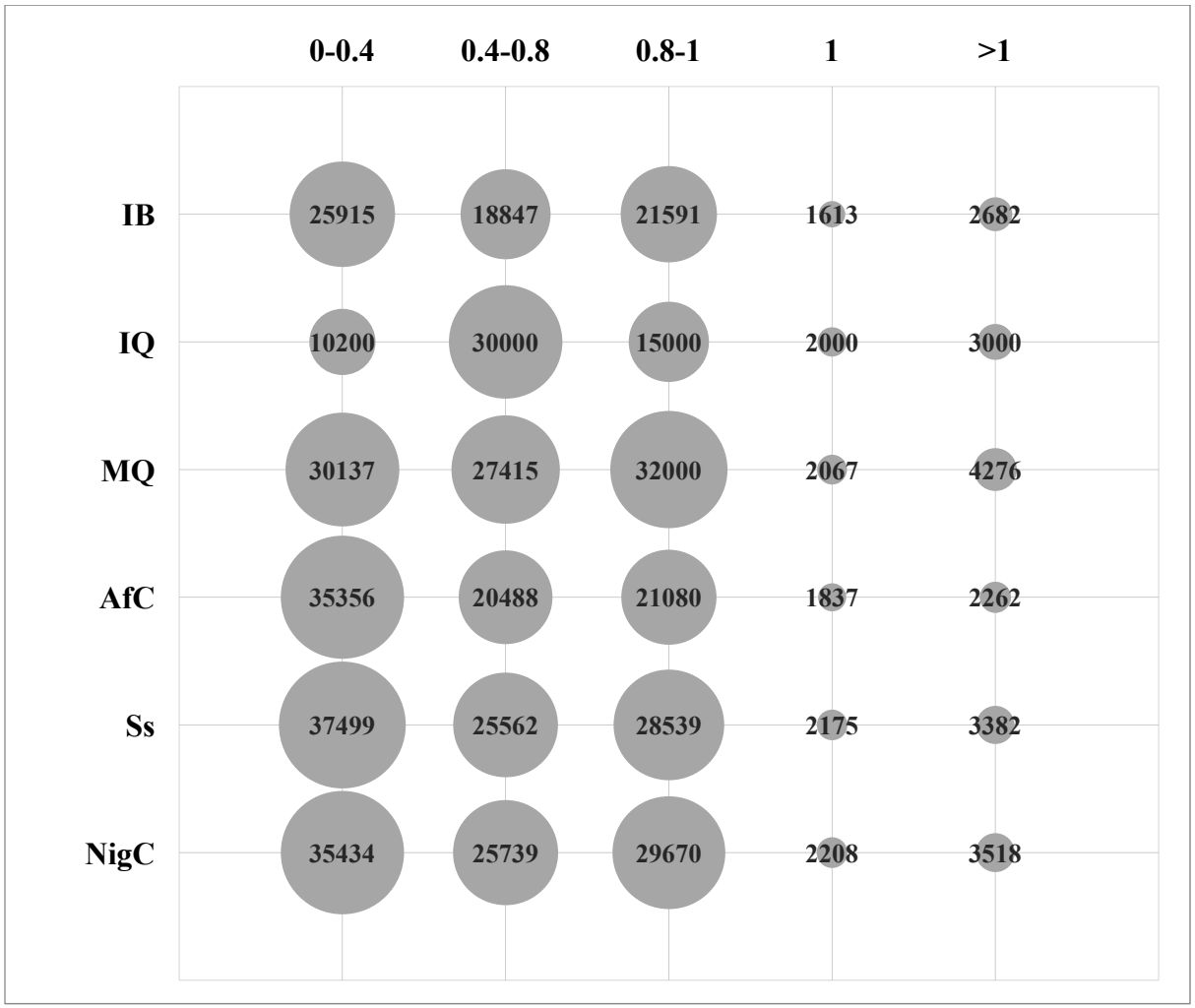


Figure 4.11: Relationship between ortholog hit ratio and ortholog length for different *B. tabaci* populations. Ortholog hit ratios were calculated based on BLASTx top hits, where the ratio of 1.0 indicates the contig is likely to be fully assembled. Ratios >1.0 indicates insertions in contigs and ratios <1.0 indicates partially assembled contigs.

4.3.7 Secretome of *B. tabaci*

Secreted pathogenic proteins also known as secretome are crucial for establishing infection on the host plant. These secretory proteins may affect the plant defence mechanism and cellular processes to support the needs of invading pathogens (Thakur *et al*., 2013). Here we used SignalP to predict the presence of signal peptides and TMHMM to predict the presence of transmembrane helices in all *B. tabaci* samples. Total number of contigs that contain the signal peptides were 7492 in sample IB, 10,322 in IQ, 10,422 in MQ, 7358 in AfC, 9876 in Ss and 10,361 in NigC (Figure 4.12). The number of contigs that contain transmembrane helices were 13,718 in IB, 18,907 in IQ, 20,740 in MQ, 13,685 in AfC, 18,219 in Ss and 19,097 in NigC.

Those proteins which contain signal peptides but lack transmembrane helices are considered as secreted proteins. Following these criteria, a total of 5724 (IB), 7663 (IQ), 8477 (MQ), 5507 (AfC), 7511 (Ss) and 7552 (NigC) were predicted to be secreted (Figure 4.13).

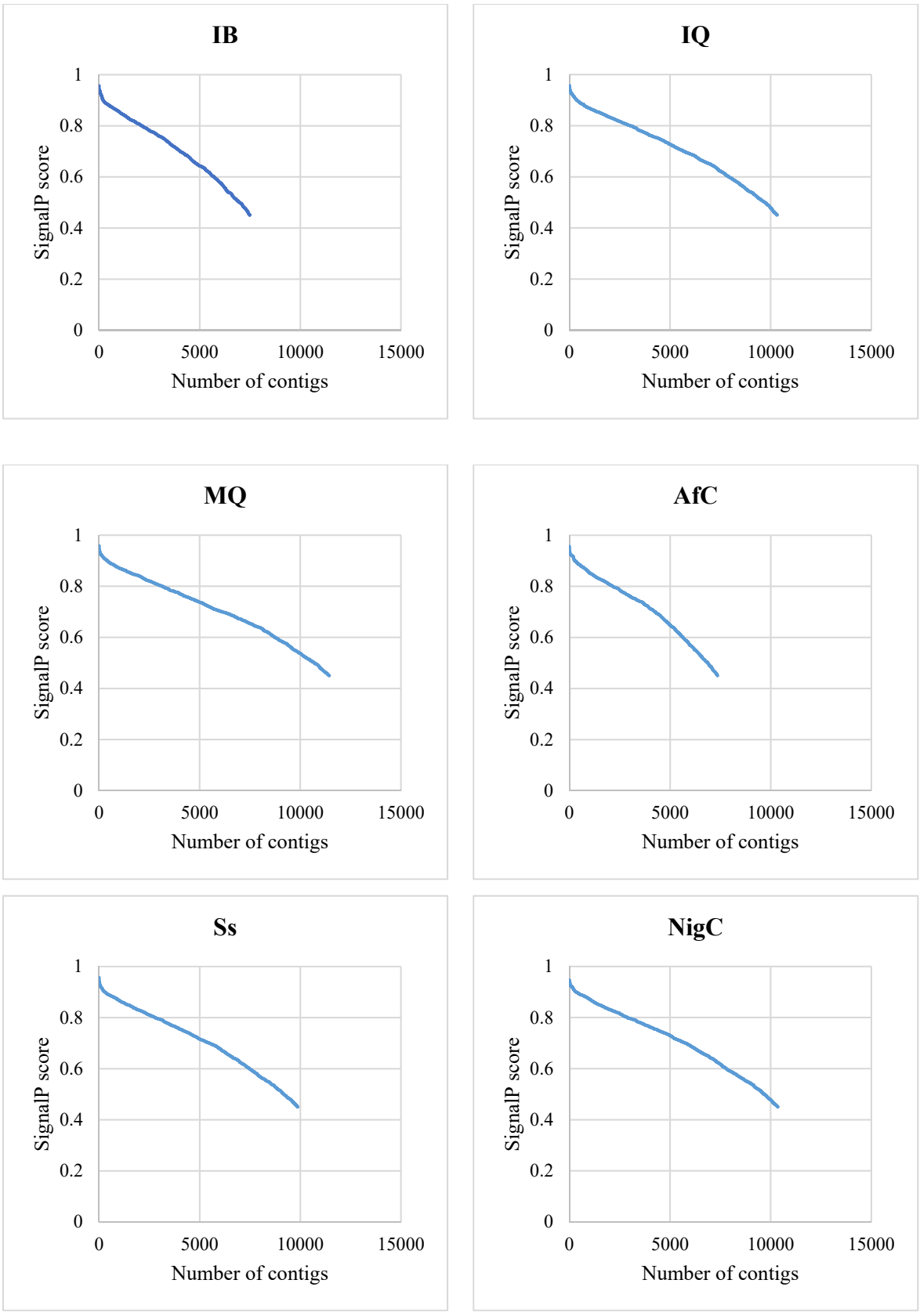


Figure 4.12: The distribution of predicted signal peptides based on probability score. Here, the probability score of 1 indicates the contig likely to contain signal peptide motif.

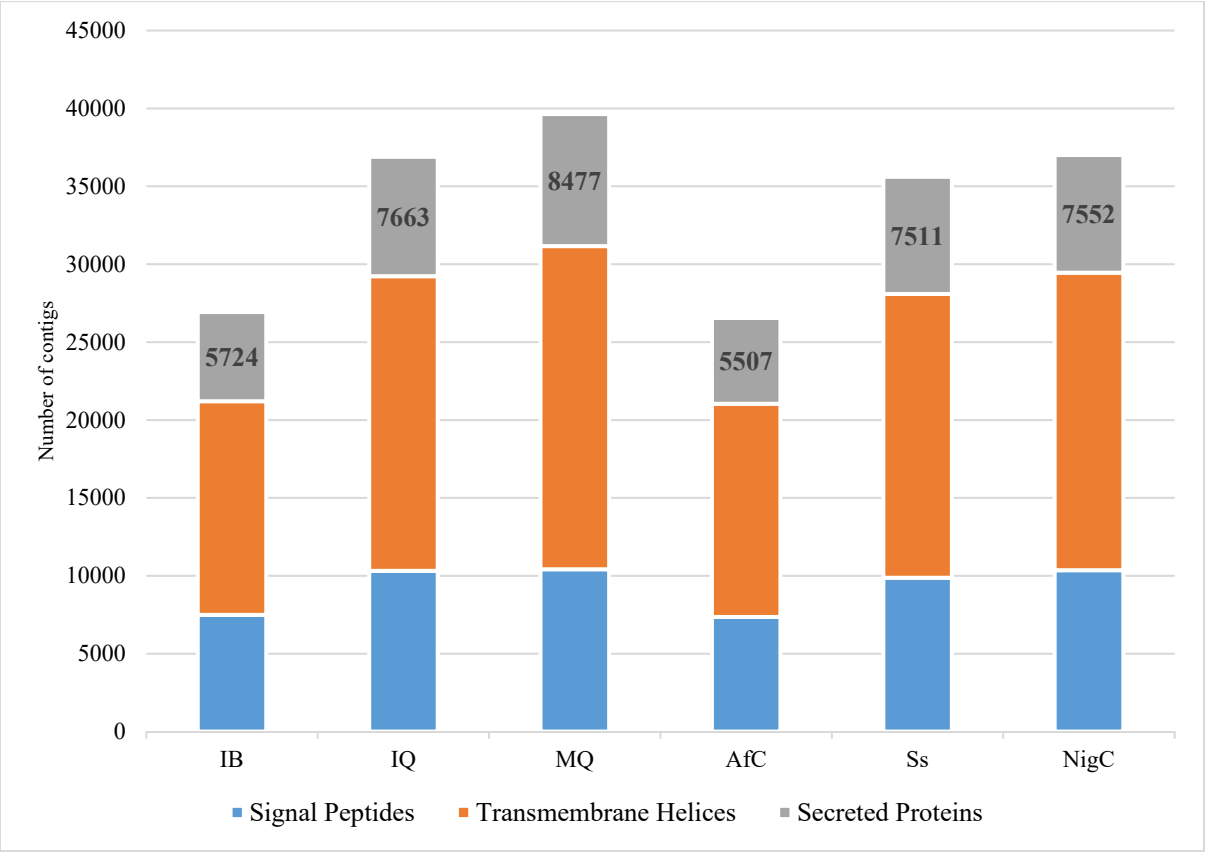


Figure 4.13: The distribution of predicted secretome proteins in *B. tabaci* populations.

4.3.8 SSR discovery

For identification of SSRs, contigs of *B. tabaci* samples were analysed using MISA (MIcroSAtellite identification tool) Perl script. SSRs or microsatellites are polymorphic repeat regions of 2 to 6 base pairs in length present in genomic DNA. A total of 65,410 (IB), 105,149 (IQ), 94,941 (MQ), 50,844 (AfC), 65,830 (Ss) and 74,353 (NigC) SSRs were identified in 185,895 (IB), 257,163 (IQ), 256,401 (MQ), 242,664 (AfC), 287,559 (Ss) and 280,616 (NigC) contigs, of which 14,239 (IB), 23,976 (IQ), 21,249 (MQ), 9,515 (AfC), 12,135 (Ss) and 14,426 (NigC) contigs contained more than 1 SSR (Table 4.2). Of these SSRs, the largest fraction was mononucleotides, followed by tri-nucleotides and di-nucleotides. The most abundant of mononucleotide was poly-T, accounting for 40.48% (IB), 41.73% (IQ), 39.48% (MQ), 40.05% (AfC), 39.83% (Ss) and 39.51% (NigC) followed by the poly-A motif. In the 10 types of trinucleotide repeats, AAC/GTT was the most common motif for samples IB, IQ, MQ and Ss, while AAG/CTT found to be most common motif in samples AfC and NigC. Within di-nucleotide repeats, AG/CT was the most common type of motif, followed by AC/GT and AT/AT (Figure 4.14).

The number distribution of predicted SSRs is summarized in Table 4.3. The results show that the repeat number of most SSRs was between 5 and 7, and there are very few that are greater than 10. The number of repeats for samples IB, IQ and MQ are higher than the sample AfC, Ss and NigC but most number of hexanucleotide repeats were present in sample NigC. The highest repeat number for dinucleotide was 36 for sample MQ.

	IB	IQ	MQ	AfC	Ss	NigC
Total number of sequences examined	185895	257163	256401	242664	287559	280616
Total number of identified SSRs	65410	105149	94941	50844	65830	74353
Number of sequences containing SSR	44405	68997	64179	37495	49605	54375
Number of sequences containing more than 1 SSR	14239	23976	21249	9515	12135	14426
Number of SSRs present in compound formation	9982	17025	13202	7111	7248	8753

Table 4.2: Summary of SSRs found in transcriptome assemblies of *B. tabaci* populations (IB, IQ, MQ, AfC, Ss and NigC).

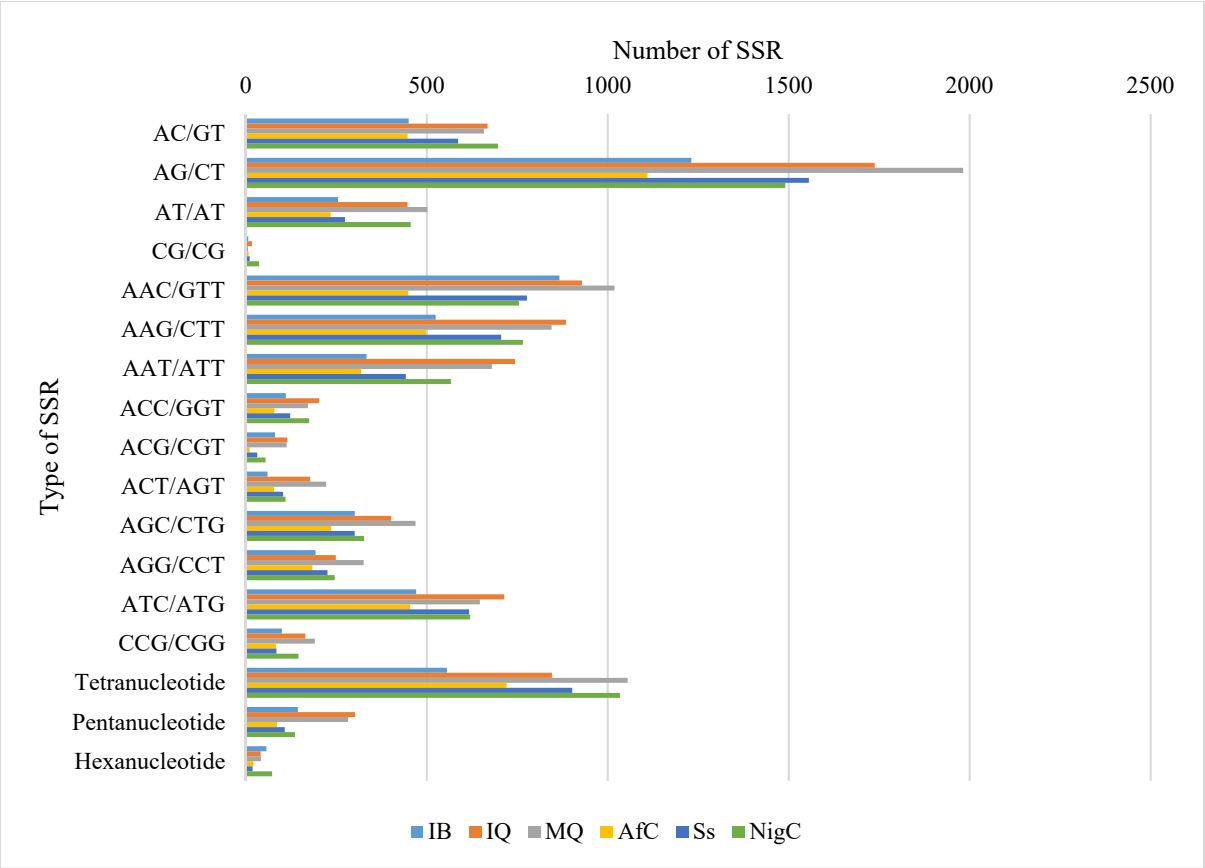


Figure 4.14: Overview of type and frequency of repeat motif found in six *B. tabaci* populations.

	SSR motifs	Number of repeats							
		5	**6**	**7**	**8**	**9**	**10**	**>10**	**Total**
Sample IB	Dinucleotide	0	1017	424	206	98	85	113	1943
	Trinucleotide	1709	598	231	229	100	44	132	3043
	Tetranucleotide	300	131	62	12	28	1	22	556
	Pentanucleotide	73	10	7	11	12	13	18	144
	Hexanucleotide	17	6	9	11	1	0	10	54
Sample IQ	Dinucleotide	0	1580	579	174	196	95	246	2870
	Trinucleotide	2695	831	382	329	119	42	186	4584
	Tetranucleotide	485	196	67	25	26	12	35	846
	Pentanucleotide	173	33	27	30	16	10	13	302
	Hexanucleotide	23	11	6	1	0	0	0	41
Sample MQ	Dinucleotide	0	1679	644	271	147	110	297	3148
	Trinucleotide	2900	949	345	202	101	37	150	4684
	Tetranucleotide	666	224	77	36	21	14	17	1055
	Pentanucleotide	144	52	28	18	8	9	24	283
	Hexanucleotide	20	10	7	0	0	0	1	38
Sample AfC	Dinucleotide	0	1046	398	119	109	32	96	1800
	Trinucleotide	1391	437	250	150	56	15	92	2391
	Tetranucleotide	361	186	59	27	28	29	31	721
	Pentanucleotide	58	6	6	10	1	1	4	86
	Hexanucleotide	12	4	1	1	3	0	0	21
Sample Ss	Dinucleotide	0	1417	535	159	145	46	126	2428
	Trinucleotide	2025	607	290	275	52	19	144	3412
	Tetranucleotide	434	211	114	31	45	22	45	902
	Pentanucleotide	64	10	11	15	3	1	3	107
	Hexanucleotide	8	8	1	0	0	0	1	18
Sample NigC	Dinucleotide	0	1496	591	255	140	54	145	2681
	Trinucleotide	2127	761	384	233	77	38	147	3767
	Tetranucleotide	490	260	90	54	37	47	56	1034
	Pentanucleotide	120	7	2	3	1	0	3	136
	Hexanucleotide	46	2	7	7	3	4	3	72

Table 4.3: Distribution of SSRs in *B. tabaci* populations.

4.4 Discussion

Transcriptomes of three cassava and three non-cassava *B. tabaci* populations were sequenced from one lane of Illumina HiSeq 2000 sequencer. The Illumina HiSeq 2000 sequencing platform used in this study is capable of generating two billion 100 bp paired-end reads with a yield up to 200 Gb per run. Previous transcriptome studies by Wang *et. al*. 2010; 2011; 2012 for analysing *B. tabaci* populations MED, MEAM1 and Asia II 3 produced 640 million 150 bp paired-end reads with a yield capacity of 95 Gb per run (Source: www.illumina.com). In the absence of a reference genome for *B. tabaci* at the time of this study, the approach used was to assemble *B. tabaci* transcriptome sequencing reads. Results from Chapter 2 indicated that the Trinity assembler performed best for assembling full length transcripts based on percentage of reads successfully mapped back to contigs, number of CRB BLAST hits obtained against *A. pisum* and *D. citri* and based on BUSCO analysis. Previous transcriptome studies on yeast, mouse and whitefly by Grabherr *et. al*. 2013 also reported that the Trinity outperformed all other *de novo* assemblers such as ABySS, Trans-Abyss and SOAPdenovo. The results were reported based on the number of transcripts generated for each sample, the average length of contig and also the number of full-length contigs. The results also suggested that the Trinity resolved ~99% of the sequencing errors, substantial polymorphism as well as transcripts derived from splicing and duplication events (Grabherr *et. al*., 2013).

The total number of contigs assembled for all samples (185,895 (IB), 257,163 (IQ), 256,401 (MQ), 242,664 (AfC), 287,559 (Ss) and 280,616 (NigC)) used in this study were higher than previous findings where the contigs assembled for MED (Wang et al., 2010), MEAM1 (Wang et al., 2011) and Asia II 3 (Wang et al., 2012) (170,115, 123,055 and 144,103, respectively). The difference in numbers could be the result of amount of reads, assembly program and the parameters used to assemble the contigs, especially the k-mer size. The transcriptome assemblies for MED, MEAM1 and Asia II 3 were generated using the SOAPdenovo with k-mer size 21. Whereas, the assemblies produced in this study were generated using the Trinity by combining three k-mer sizes 25, 27 and 29. The mean length of contigs assembled for MED, MEAM1 and Asia II 3 were 266 bp, 269 bp and 201 bp, respectively compared to 1155 bp (IB), 1221 bp (IQ), 1355 bp (MQ), 881 bp (AfC), 987 bp (Ss) and 1081 bp (NigC). These results suggested that the size of k-mer is an important factor for assembling raw sequencing reads.

While annotating assembled contigs using BLASTx, we discovered that 68.69% (IB), 71.76% (IQ), 75.46% (MQ), 68.79% (AfC), 72.95% (Ss) and 73.53% (NigC) contigs could not be annotated with known biological functions. This may have been due to fact that only limited (12,094 EST, 14,359 protein and 260,065 nucleotide sequences) were available for *B. tabaci* in NCBI (October 3, 2016). The unannotated contigs may play essential roles in the biology of *B. tabaci*, and hence further research is required to understand the role of these unknown genes. The E-value distribution results for MED (Wang *et al*., 2010) and MEAM1 (Wang *et. al*., 2011) showed that 28.03% and 34.04%, respectively of the contigs have high similarity with E-value smaller than $1.0E^{-40}$. Whereas, the results in this study found 39% (IB), 38% (IQ), 35% (MQ), 43% (AfC), 40% (Ss) and 38% (NigC) of the contigs ranged between $1.0E^{-4}$ to $1.0E^{-40}$. The sequence similarity distribution graph (Figure 4.3) showed that 12.66% (IB), 14.73% (IQ), 10.89% (MQ), 13.51% (AfC), 12.20% (Ss) and 11.70% (NigC) of the contigs have a similarity higher than 80% compared to 18% (MED) and 11.9% (MEAM1).

The taxonomic distribution results (Figure 4.4) were found startling with termite, *Z. nevadensis* having highest number of hits followed by *A. pisum*. Despite having distinct lineages, high percentage of contigs matched to *Z. nevadensis* was surprising as both *B. tabaci* and *Z. nevadensis* do not share common characteristics like diet and life cycle stages. Though, when we compared OHR statistics for the *Z. nevadensis and A. pisum,* we found that high percentage of contigs fall between the OHR score of 0.8 to 1.0 for *Z. nevadensis* than for *A. pisum* (Figure 4.6). Here, the OHR score of 1.0 indicates the contig is likely to be fully assembled. These results indicating high similarity between *Z. nevadensis* and *B. tabaci* may be false positives due to error in assembling or contamination in sampling and a further study is required to study sequence identity and alignment coverage between two to understand sequence similarity.

The species distribution results for MED (Wang *et al*., 2010) and MEAM1 (Wang *et al*., 2011; Xie *et al*., 2012) showed that the highest percentage of contigs were matched to pea aphid (*Acyrthosiphon pisum*). This is probably due to the genome sequence of *Z. nevadensis* were only available recently (March 2014) in NCBI databases and were not observed in previous whitefly transcriptome analysis results (Wang *et al*., 2010; Wang *et al*., 2011). As these results do not support the close relationship between Isoptera and Hemiptera group in the taxonomic status, further research is required. In the absence of full genome sequences

for *B. tabaci*, only 1055 (IB), 1082 (IQ), 1246 (MQ), 1295 (AfC), 1508 (Ss) and 1484 (NigC) contigs were matched to *B. tabaci.* Whereas, the results of MED (Wang *et al*., 2010), MEAM1 (Wang *et al*., 2011) and MEAM1 (Xie *et al*., 2012) showed 126, 97 and 215 contigs, respectively matched to *B. tabaci*. This is probably due to the number of sequences deposited in NCBI is growing exponentially than it was before for *B. tabaci* as the number of nucleotide sequences for MED and MEAM1 were only 1445 (Wang *et al*., 2010) compared to nucleotide sequences available for *B. tabaci* are 260,065 on October 3, 2016 (Source: http://www.ncbi.nlm.nih.gov).

The GO annotations revealed that "metabolic process", "catalytic activity" and "cell" were the most abundant sub-categories for all six *B. tabaci* populations within three main categories biological process, molecular function and cellular component, respectively. We also noticed a high percentage of contigs matched to "cellular process", "binding" and "cell part". These results indicates that cassava and non-cassava *B. tabaci* population shows similar functional distribution as observed by Wang *et al*., 2012 when they compared the GO annotation results of MED, MEAM1 and Asia II 3, with "metabolic process", "cell" and "binding" were found to be highly represented, indicating different *B. tabaci* populations shares common conserved genes regardless of differences in the amount of sequencing data used (Wang *et al*., 2011, 2012). Interestingly, when we compared the results with other GO sub-categories, we found that contigs associated with biological process: "anatomical structure formation", "death" and "pigmentation" were missing in all six samples used in this study and were present in MED, MEAM1 and Asia II 3. Whereas, the contigs associated with "detoxification" were not found in MED, MEAM1 and Asia II 3. Likewise, "envelope" in cellular component, "auxiliary transport protein" and "enzyme regulator" in molecular function were missing in IB, IQ, MQ, AfC, Ss and NigC. The difference in results may be due to the amount of sequencing reads obtained for samples (IB (46 million), IQ (81 million), MQ (82 million), AfC (30 million), Ss (38 million), NigC (43 million)) used in this study were higher than MED (43.7 million), MEAM1 (17 million) and Asia II 3 (16.8 million) (Wang *et al*., 2012). Another possibility could be that the samples used in MED, MEAM1 and Asia II 3 were prepared using egg & nymph, pupa, female and male adult, whereas the samples used in this study were obtained from male and female adults only and therefore some of the genes might have not expressed at different levels of life cycle stage.

Biological pathways and enzymes provide key information about processes and functions that are active during different life stages of *B. tabaci*. When we searched the assembled contigs to identify possible biological pathways present in five main categories (Figure 9), purine metabolism within the nucleotide metabolism represented 2111 (IB), 3042 (IQ), 2502 (MQ), 2854 (AfC), 3177 (Ss) and 3028 (NigC) contigs compared to 458 for MED (Wang *et al*., 2010). However, the starch and sucrose metabolism pathway in MED contained 553 contigs compared to 130 (IB), 253 (IQ), 206 (MQ), 294 (AfC), 325 (Ss) and 247 (NigC) contigs. Whereas, galactose metabolism matched 68 (IB), 127 (IQ), 87 (MQ), 150 (AfC), 183 (Ss) and 141 (NigC) contigs compared to 183 contigs for MED (Wang *et al*., 2010). Here, we also noticed that the number of contigs matched to galactose metabolism are much higher for cassava populations than for non-cassava populations. These functional annotations provide a key reference to understand and investigate functions and metabolic activities of essential genes in different *B. tabaci* populations.

To identify full length contigs, we used OHR statistics proposed by O'Neil *et al*., 2010. Based on that, we found that the large number of contigs with score <1.0 were either partially assembled or were not aligned fully due to insertions in ortholog sequences based on scoring matrices (Figure 4.12). This is probably due to the length of assembled contigs (~60%-70% were between 200 bp to 1000 bp) as shown in Figure 4.1. These results suggested that the assembled sequences were either failed to match the best BLASTx ortholog sequences probably due to the error in assembly or may be that the limited genomic resources are available for *B. tabaci* in public domains. The resulted contigs were also analysed to predict SSRs present in all six *B. tabaci* populations used in this study. SSR markers are useful for assessing genetic variation in organisms when full genome sequences are not available (Kalra *et al*., 2013). SSR markers developed using NGS data are much cheaper than the traditional isolation methods as the large amount of data can be produced using NGS technologies (Zhang *et al*., 2014). Microsatellites identified from *B. tabaci* transcriptome of MED, MEAM1 and Asia II 3 also showed that A/T motifs were most abundant in mononucleotide repeats and AG was the most frequent di-nucleotide class as found in this study. However, most prevalent tri-nucleotide repeats obtained for MED, MEAM1 and Asia II 3 were from AAG, AAG and ATG class, respectively (Wang *et al*., 2014) whereas, it is AAC for samples IB, IQ, MQ, Ss and AAG for AfC and NigC. These findings indicate that the potential SSR markers identified here may provide a useful resource for studying genetic diversity within *B. tabaci* populations. However, many of the identified SSRs could be the

result of isoforms present in transcriptome data and therefore further research is required to validate the real ones.

Chapter 5: Analysing transcriptome data for other potential mechanisms of evolution and diversity

5.1 Introduction

Bemisia tabaci is considered one of the world's top 100 invasive species capable of causing extensive damage not only to cassava but also a wide range of vegetable, grain and fiber crops through direct feeding or by transmitting plant pathogenic viruses (Boykin *et al*., 2013; Wang *et al*., 2013). The *B. tabaci* species complex contains at least 11 major clades which are morphologically indistinguishable and hence complete mitochondrial genome information would be useful to identify and understand conserved and divergence patterns within species apart from traditional delimitation method based on partial mtCOI gene (Tay *et al*., 2014). To identify genetic diversity between populations used in this study and published mitochondrial genome of Asia I (Tay *et al*., 2014) and Mediterranean (Wang *et al*., 2013), we compared the transcriptome data of three non-cassava Israel B (IB), Israel Q (IQ), Montpellier Q (MQ) and three cassava East African cassava Nam (AfC), East African cassava Ssanje (Ss) and West African cassava Nigeria (NigC) *B. tabaci* populations.

Symbiotic bacteria play an important role in eukaryotic evolution and diversity (Kikuchi, 2009). Almost all insects are associated with heritable endosymbiotic bacteria which are widespread in nature and can be divided into several groups (Moran and Baumann, 2000). The most general types are: primary endosymbionts (P-endosymbionts) and secondary endosymbionts (S-endosymbionts) (Bing *et al*., 2013). P-endosymbionts are present in all host types and provide important nutritional requirements to their insect host. S-endosymbionts are facultative and play negative role in host survival as well as provide nutrients, parasitiods resistance and improve thermotolerance (Bing *et al*., 2013).

As a sap-feeding pest, whiteflies have developed relationships with a wide range of primary and secondary endosymbionts (Sloan and Moran, 2012). All whiteflies harbour the primary endosymbiont *Candidatus Portiera aleyrodidarum* with an extremely reduced genome that provides only essential amino acids and carotenoids to host (Santos-Garcia *et al*., 2014). In addition to P-endosymbionts, genetically distinct whitefly *B. tabaci* hosts different combination of S-endosymbionts that may include *Wolbachia*, *Rickettsia*, *Hamiltonella*, *Cardinium*, *Arsenophonus* and *Fritschea* (Tajebe *et al*., 2014).

Recent studies on importance of endosymbionts within *B .tabaci* complex shows that these endosymbionts have diverse effects on their host and therefore it is essential to understand their exact functions and mechanisms within species (Bing *et al*., 2013). This study was carried out to identify the primary and secondary endosymbionts present in cassava and non-cassava *B. tabaci* populations.

5.2 Methods

5.2.1 Assembling mitochondrial genes

To obtain mitochondrial gene sequences of samples IB, IQ, MQ, Afc, Ss and NigC, complete mitogenome sequence of published Asia I species of *B. tabaci* was downloaded from GenBank (KJ778614) and was used as a reference to search against Illumina reads. All mapped sequences were then retrieved and stored based on highest number of reads obtained for particular gene sequence using CLC Genomics Workbench 7.

5.2.2 Comparative sequence analysis of 13 PCGs

After mapping, all matched sequences were aligned using ClustalW (Thompson *et al*., 1994) to identify conserved regions within different *B. tabaci* populations and to see the differences in sequence lengths. The aligned sequences were then manually edited using MEGA6 alignment explorer and were used to predict evolutionary trees based on concatenated alignment of 13 PCGs and also by individual sequence analysis of 13 PCGs using MEGA6 (Tamura *et al*., 2013). The best scoring models from maximum likelihood models were selected based on automated model selection method.

5.2.3 Identifying primary and secondary endosymbionts

All assembled files were stored in a FASTA format and were searched using BLASTX (Altschul *et al*., 1990) against the non-redundant (nr) NCBI nucleotide database using a cut-off E-value of 10^{-3}.The top-hits from BLASTX results were used to identify primary and secondary endosymbionts present among all six samples.

5.2.4 Phylogenetic analysis of *Portiera, Cardinium*, *Hamiltonella* and *Rickettsia*

The 16S rDNA sequences of *Portiera*, *Cardinium*, *Hamiltonella* and *Rickettsia* taken from NCBI were aligned against assembled contigs using BLASTN for all samples. Sequences with >90% identity were retrieved and stored for phylogenetic analysis. The resulted sequences of 16S rDNA were then aligned using ClustalW and were manually edited to get best possible alignment score using MEGA6 sequence alignment explorer. All trimmed sequences were then used to predict best substitution model for phylogenetic tree construction. Phylogenetic trees were constructed using maximum-likelihood method by generating 1000 bootstrap values in MEGA6. The pair wise distance values for each alignment were generated using p-distance method using MEGA6.

5.3 Results

5.3.1 Identifying mitochondrial genes

To obtain mitochondrial gene sequences, 46 million paired end reads for sample IB, 81 million paired end reads for sample IQ, 82 million paired end reads for sample MQ, 30 million paired end reads for sample AfC, 38 million paired end reads for sample Ss and 43 million paired end reads for sample NigC were used to retrieve the mitogenome sequences for each sample by mapping them against Asia I mitogenome sequences downloaded from GenBank. Number of reads mapped to each gene is shown in Table 5.1. Total of 31 contigs for IB, 31 contigs for IQ, 33 contigs for MQ, 14 contigs for AfC, 16 contigs for Ss and 16 contigs for NigC were identified against published Asia I mitochondrial genes. These numbers varied with the published mitochondrial genome of MED and Asia I which encodes 37 genes (Wang *et al*., 2013; Tay *et al*., 2014). The complete mitogenomes of MED and Asia I contain 13 PCGs (Protein Coding Genes) found in most metazoan genomes are present in all six samples used in this study apart from ND4L gene that was lacking in sample AfC. The 2 rRNAs genes were also present in all six samples with a variable length sizes compare to one found in MED and Asia I mitogenome. All tRNAs were absent in all cassava (AfC, Ss and NigC) samples apart from tRNA-Asp present in Ss and NigC compare to non-cassava (IB, IQ and MQ) samples as shown in Table 5.2.

Name	IB	IQ	MQ	AfC	Ss	NigC
COI	497423	1019930	1010439	342569	342260	418909
tRNA-Leu2	67	32	13	255	311	149
COII	103962	156439	137847	47895	49510	76100
tRNA-Lys	130	52	54	74	69	43
ATP8	21	5	2	182	184	189
ATP6	14182	7650	3520	432	378	1022
tRNA-Ser1	0	0	0	0	0	0
tRNA-Glu	1	0	0	5	6	1
tRNA-Phe	2	2	1	21	7	6
ND5	35904	66316	40977	26867	17549	29001
tRNA-His	8	11	24	44	53	45
ND4	50606	112338	73860	58722	41423	45948
ND4L	150	242	324	0	4	5
tRNA-Thr	1	3	1	7	2	0
tRNA-Pro	11	1	1	3	0	2
ND6	548	882	355	1166	712	664
CYTB	24897	39019	23176	71788	49163	86432
tRNA-Ser2	0	0	0	0	0	0
ND1	37300	77888	58636	15575	14457	20212
tRNA-Leu1	0	7	7	14	19	18
rRNA-L	3319122	8684092	4724191	4167062	2685676	3881589
tRNA-Val	4	16	20	17	6	7
tRNA-Asp	2	0	0	9	3	2
tRNA-Gln	2	2	0	39	42	14
rRNA-S	18530	20511	17075	10327	12746	14280
tRNA-Asn	5	5	4	4	2	0
tRNA-Arg	0	0	3	22	40	32
tRNA-Ala	16	21	5	131	153	110
ND3	14	19002	9420	4188	4934	7544
tRNA-Gly	12	23	4	41	26	12
COIII	46882	72760	42407	12185	9670	15591
tRNA-Ile	3	0	2	6	8	3
tRNA-Met	101	66	70	14	23	14
ND2	13922	11409	5146	13201	7771	10032
tRNA-Trp	6	7	1	2	1	1
tRNA-Tyr	4	7	1	8	8	10
tRNA-Cys	31	0	0	19	33	1

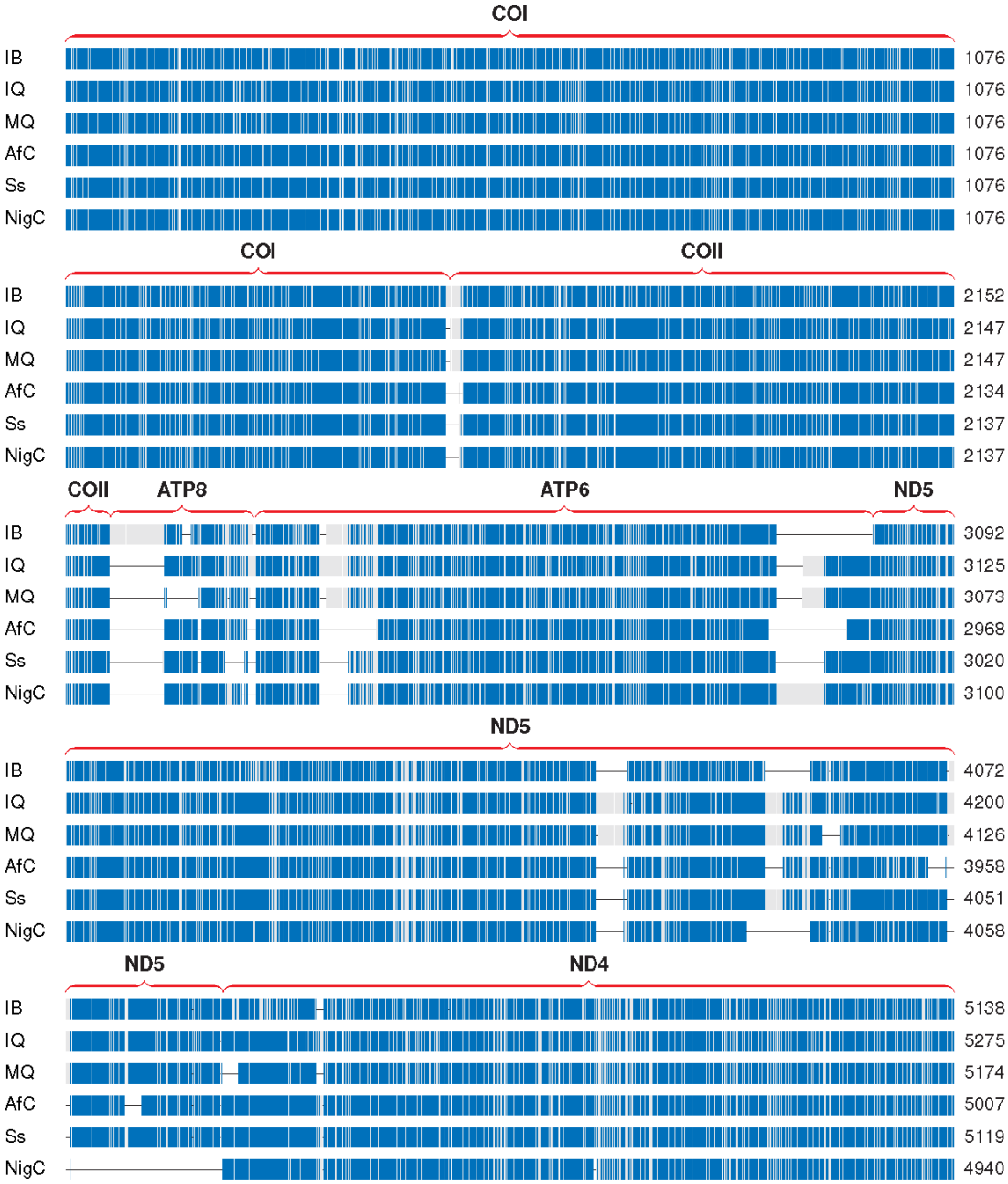
Table 5.1: Total number of mitochondrial genes against the total read number for each gene and sample. Here 0 indicates the absence of particular gene sequence in reads.

Name	IB	IQ	MQ	AfC	Ss	NigC	Asia I	MED
COI	1542	1537	1537	1541	1541	1541	1542	1537
tRNA-Leu2	65	65	65	-	-	-	65	65
COII	664	664	664	647	650	650	684	664
tRNA-Lys	68	67	67	-	-	-	68	68
ATP8	163	110	64	96	75	97	237	234
ATP6	625	631	625	554	599	597	651	651
tRNA-Ser1	-	-	-	-	-	-	60	57
tRNA-Glu	-	-	-	-	-	-	63	63
tRNA-Phe	69	67	70	-	-	-	69	67
ND5	1268	1448	1426	1285	1370	1175	1671	1654
tRNA-His	65	65	65	-	-	-	68	65
ND4	1285	1279	1256	1292	1222	1204	1293	1293
ND4L	213	279	276	-	156	195	285	285
tRNA-Thr	61	65	64	-	-	-	64	78
tRNA-Pro	-	61	62	-	-	-	62	62
ND6	416	432	401	335	412	297	447	547
CYTB	1127	1135	1093	1075	1134	1115	1134	1137
tRNA-Ser2	-	-	-	-	-	-	57	64
ND1	884	902	870	842	846	878	903	936
tRNA-Leu1	-	68	68	-	-	-	71	68
rRNA-L	608	609	609	607	607	607	1182	1211
tRNA-Val	68	68	68	-	-	-	67	68
tRNA-Asp	71	-	-	-	73	73	76	72
tRNA-Gln	64	64	61	-	-	-	64	64
rRNA-S	742	745	746	732	738	738	751	671
tRNA-Asn	63	63	63	-	-	-	64	64
tRNA-Arg	-	-	68	-	-	-	69	69
tRNA-Ala	63	65	63	-	-	-	65	65
ND3	291	298	274	267	258	308	354	354
tRNA-Gly	64	63	63	-	-	-	63	63
COIII	787	794	792	843	841	842	843	786
tRNA-Ile	64	-	63	-	-	-	66	65
tRNA-Met	69	69	68	-	-	-	70	68
ND2	959	959	960	824	795	833	960	954
tRNA-Trp	68	68	68	-	-	-	69	68
tRNA-Tyr	63	63	63	-	-	-	63	63
tRNA-Cys	57	54	54	-	-	-	62	74

Table 5.2: Total number of mitochondrial genes against the length of that gene and sample compared with published Asia I and MED. The – sign indicates the absence of particular gene sequence for that sample.

5.3.2 Sequence divergence between cassava and non-cassava *B. tabaci* populations

The sequence divergence of 13 mitochondrial PCGs for cassava and non-cassava *B. tabaci* populations is shown in Figure 5.1. The multiple alignment of concatenated 13 PCGs shows differences within *B. tabaci* populations. These may be due to the sequences not having assembled fully due to lack of sufficient coverage of genomic data, assembly errors or sequencing errors. Gaps within genes also represent the absence of particular portion of gene and the dark blue bar shows conserved regions of the genes.


COI
IB 1076
IQ 1076
MQ 1076
AfC 1076
Ss 1076
NigC 1076
COI
COII
IB 2152
IQ 2147
MQ 2147
AfC 2134
Ss 2137
NigC 2137
COII
ATP8
ATP6
ND5
IB 3092
IQ 3125
MQ 3073
AfC 2968
Ss 3020
NigC 3100
ND5
IB 4072
IQ 4200
MQ 4126
AfC 3958
Ss 4051
NigC 4058
ND5
ND4
IB 5138
IQ 5275
MQ 5174
AfC 5007
Ss 5119
NigC 4940

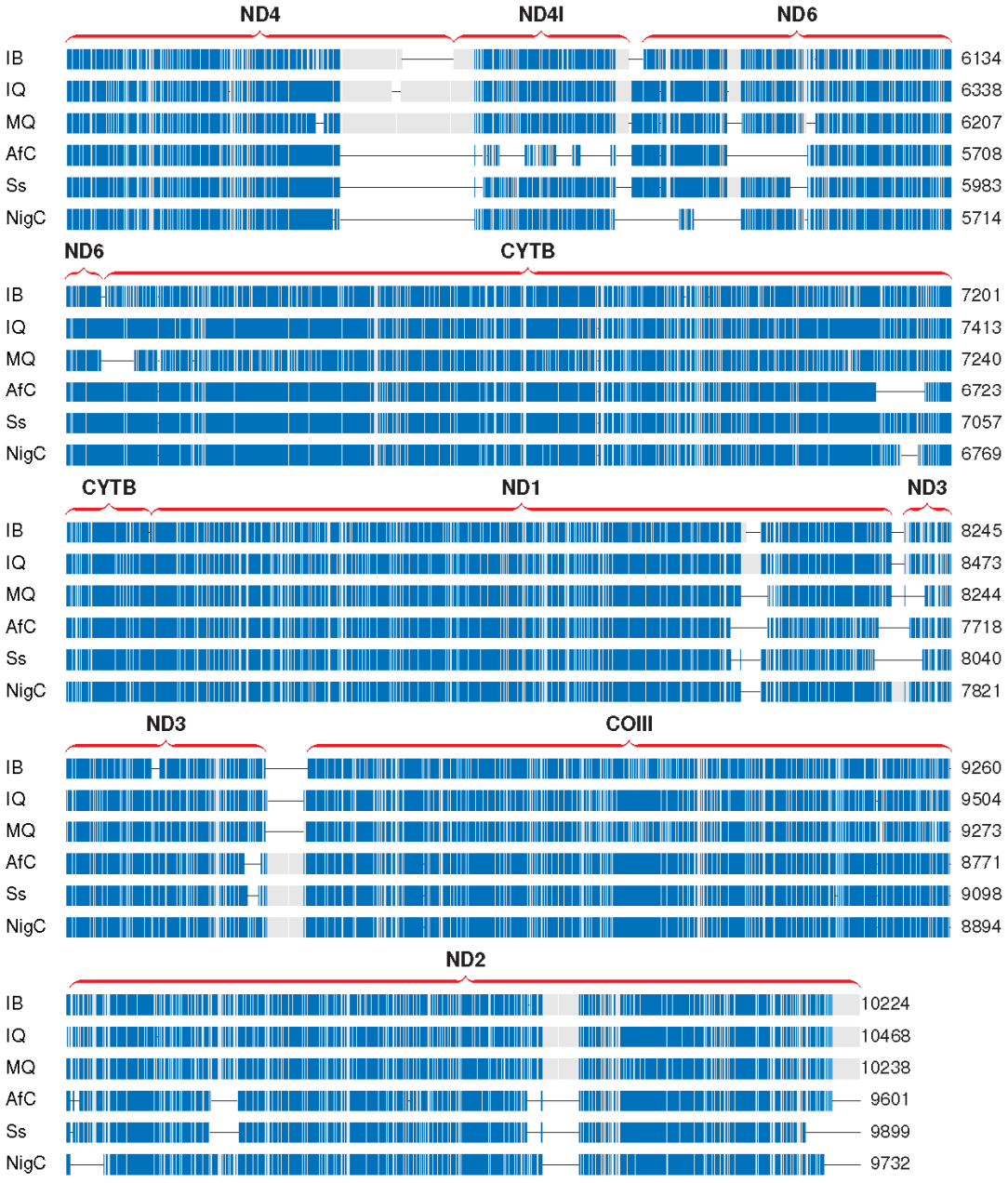


Figure 5.1: Multiple sequence alignment of concatenated genomic sequences of 13 mitochondrial PCGs. The alignment shows that the sequences are of different length and were not assembled fully. The conserved nucleotides are coloured according to % identity score from higher (blue) to lower (grey). Gaps are shown as solid horizontal line.

5.3.3 Evolutionary analysis of 13 mitochondrial PCGs

To analyse the evolutionary relationships between different *B. tabaci* mitochondrial genes used in this study, 13 PCGs were compared against MED and Asia I genes (Wang *et al*., 2013; Tay *et al*., 2014). The phylogenetic analysis on concatenated nucleotide sequences of 13 mitochondrial PCGs is shown in Figure 5.2(A). The result shows that the two East African cassava *B. tabaci* populations AfC and Ss share a common ancestry path and are more homologous than the West African population NigC. Similarly, IQ, MQ and MED share a common node compared to IB and Asia I. Similar patterns were observed in individual phylogenetic analysis of 12 PCGs (Figure 5.2) apart from ND4L, due to the absence of gene sequence in sample AfC. The comparative analysis of ATP8 gene sequences produced less significant alignment due to relatively short gene length and were not included in phylogenetic analysis.

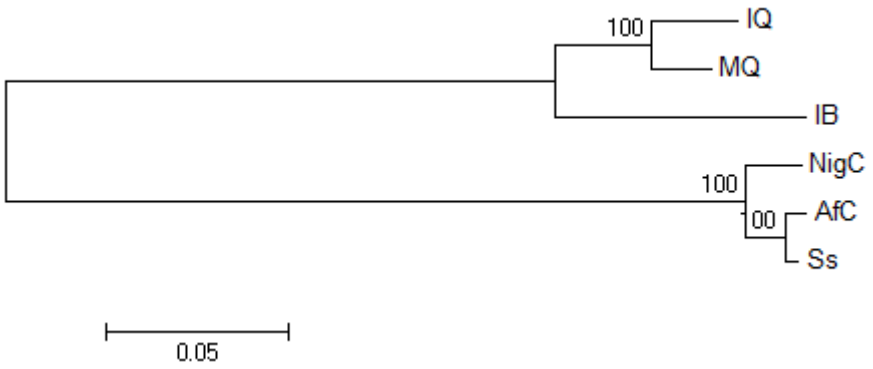


Figure 5.2(A): Maximum likelihood tree on concatenated nucleotide sequences of 13 mitochondrial PCGs for all samples was generated using HKY + G substitution model in MEGA6. The rate variation among sites was modelled with a gamma distribution (5 categories (+G, parameter = 0.3218)). The analysis involved six nucleotide sequences. All positions with less than 100% site coverage were eliminated. There were a total of 9056 positions in the final dataset.

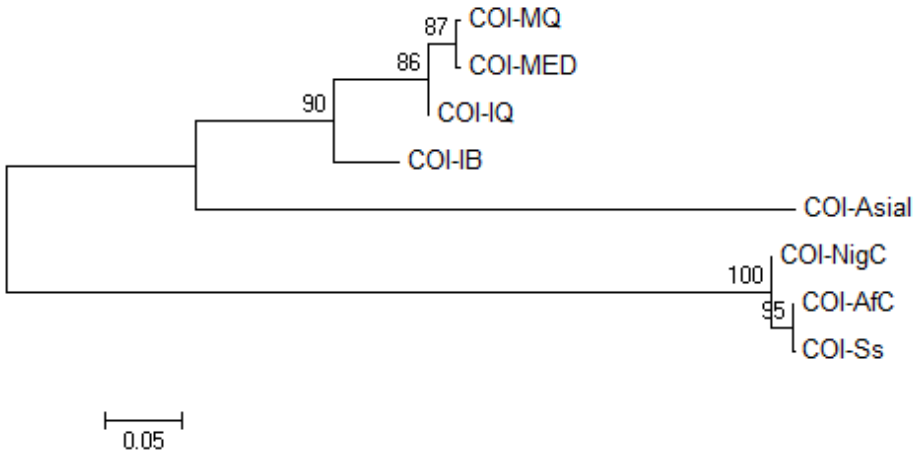


Figure 5.2(B): Maximum likelihood tree showing the relationship of COI nucleotide sequences obtained from different *B. tabaci* populations. Tree was constructed using the Hasegawa-Kishino-Yano model with a discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 0.3675)). The analysis involved eight nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 1537 positions in the final dataset.

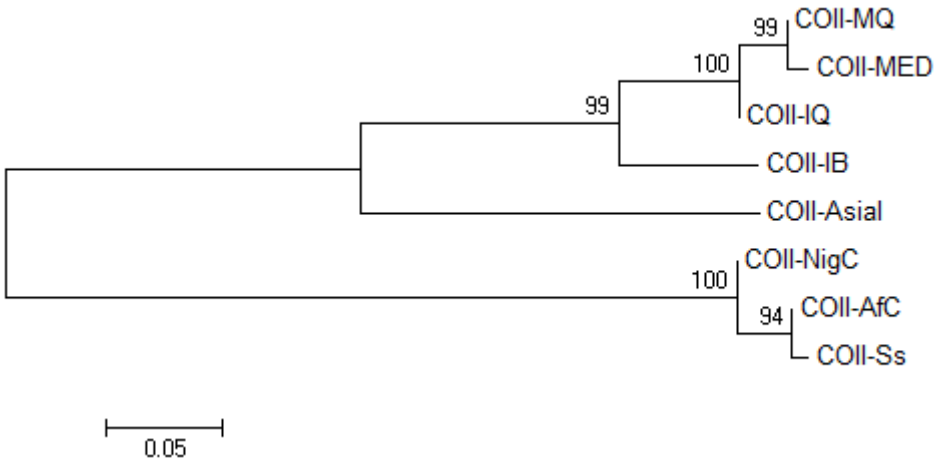


Figure 5.2(C): Maximum likelihood tree showing the relationship of COII nucleotide sequences. Tree was constructed using the Hasegawa-Kishino-Yano model with a rate variation model allowed for some sites to be evolutionarily invariable ([+I], 58.3654% sites). The analysis involved 8 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 647 positions in the final dataset.

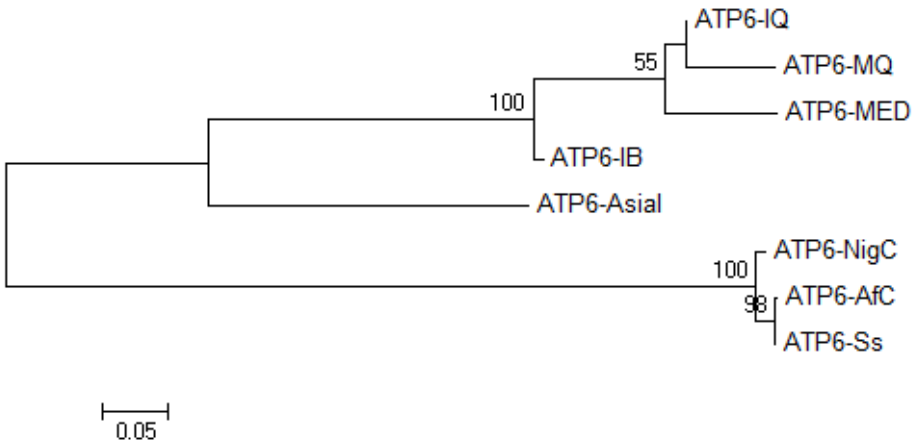


Figure 5.2(D): Maximum likelihood tree showing the relationship of ATP6 nucleotide sequences. Tree was constructed using the Hasegawa-Kishino-Yano model with a rate variation model allowed for some sites to be evolutionarily invariable ([+I], 53.9459% sites). The analysis involved 8 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 474 positions in the final dataset.

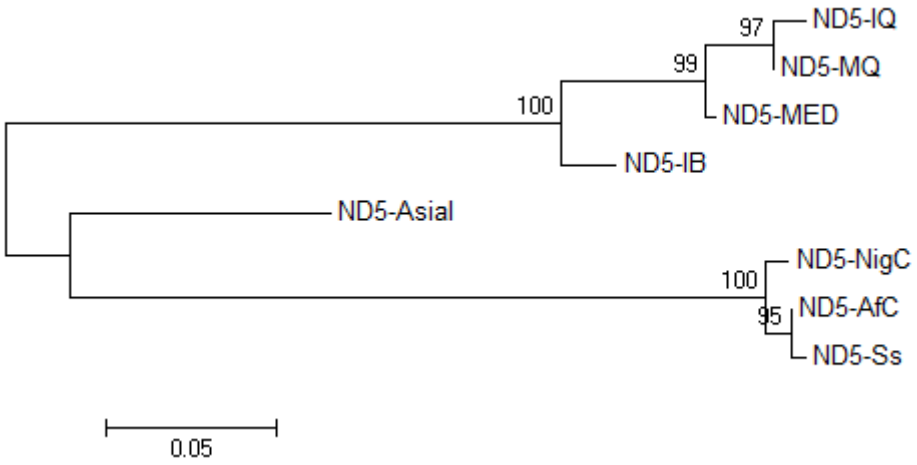


Figure 5.2(E): Maximum likelihood tree showing the relationship of ND5 nucleotide sequences. Tree was constructed using the Tamura 3-parameter model with a rate variation model allowed for some sites to be evolutionarily invariable ([+I], 0.0010% sites). The analysis involved 8 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 740 positions in the final dataset.

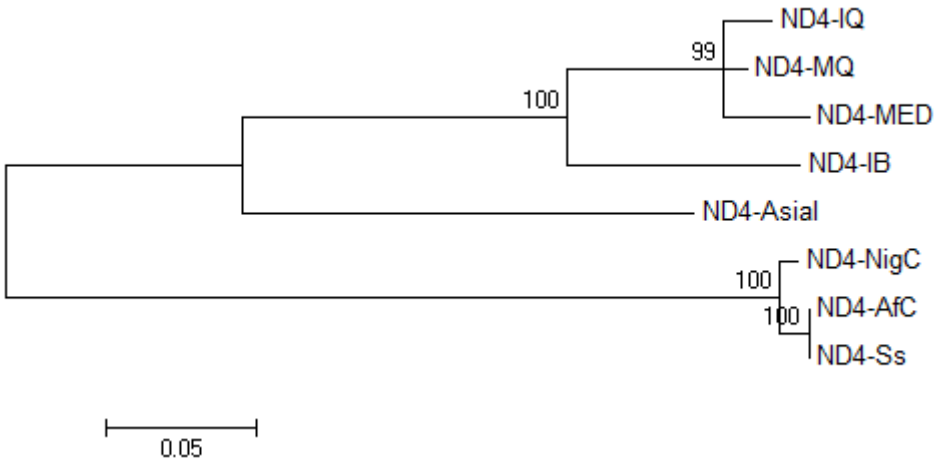


Figure 5.2(F): Maximum likelihood tree showing the relationship of ND4 nucleotide sequences. Tree was constructed using the Tamura 3-parameter model with a rate variation model allowed for some sites to be evolutionarily invariable ([+I], 24.4732% sites). The analysis involved 8 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 1156 positions in the final dataset.

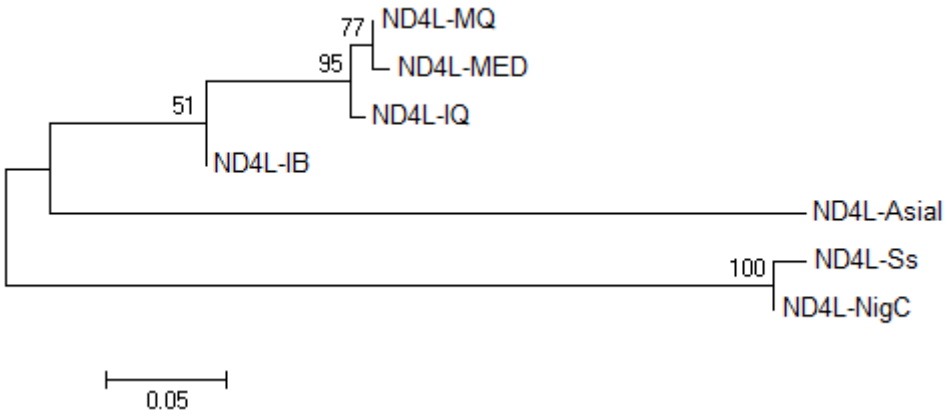


Figure 5.2(G): Maximum likelihood tree showing the relationship of NDL4 nucleotide sequences. Tree was constructed using the Tamura 3-parameter model with a discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 0.5268)). The analysis involved 7 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 156 positions in the final dataset.

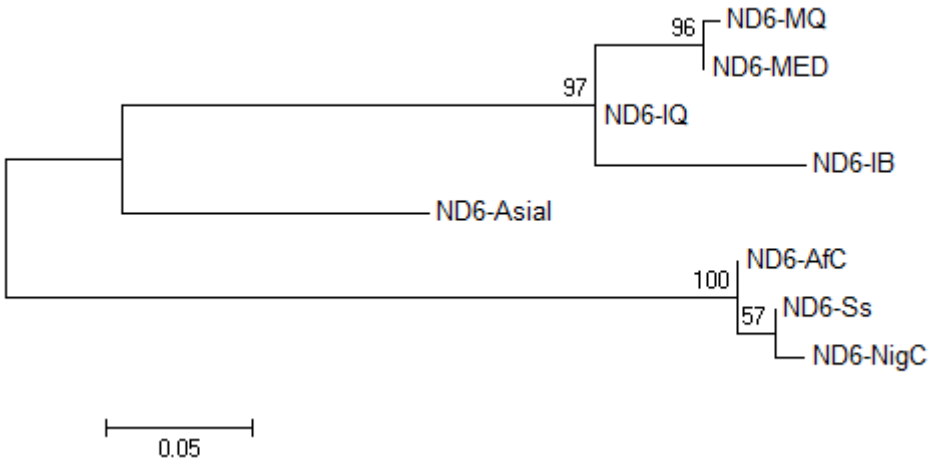


Figure 5.2(H): Maximum likelihood tree showing the relationship of ND6 nucleotide sequences. Tree was constructed using the Tamura 3-parameter model with a discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 0.3304)). The analysis involved 8 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 206 positions in the final dataset.

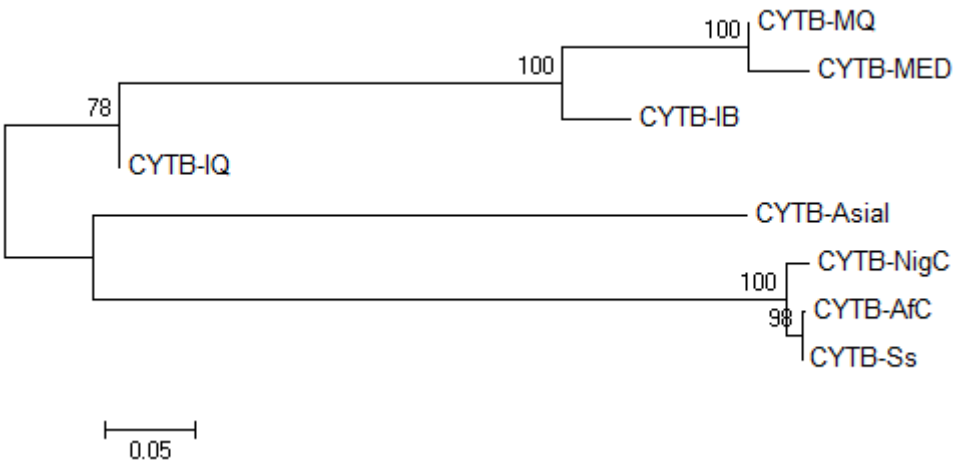


Figure 5.2(I): Maximum likelihood tree showing the relationship of CYTB nucleotide sequences. Tree was constructed using the Hasegawa-Kishino-Yano model with a rate variation model allowed for some sites to be evolutionarily invariable ([+I], 54.2063% sites). The analysis involved 8 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 897 positions in the final dataset.

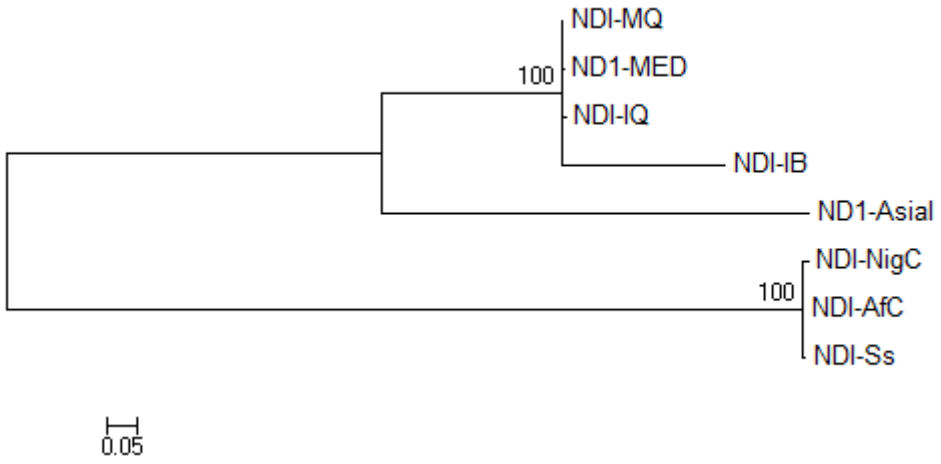


Figure 5.2(J): Maximum likelihood tree showing the relationship of ND1 nucleotide sequences. Tree was constructed using the Tamura 3-parameter model with a rate variation model allowed for some sites to be evolutionarily invariable ([+I], 59.5461% sites). The analysis involved 8 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 705 positions in the final dataset.

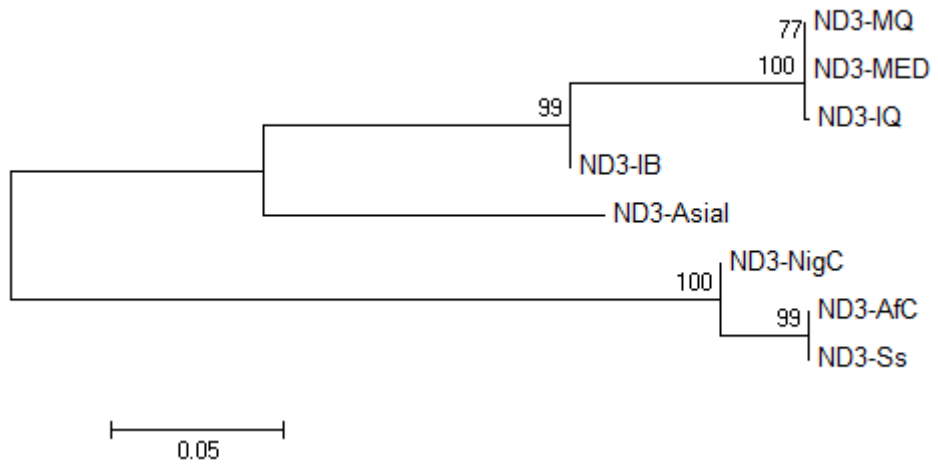


Figure 5.2(K): Maximum likelihood tree showing the relationship of ND3 nucleotide sequences. Tree was constructed using the Tamura 3-parameter model with a discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 0.4757)). The analysis involved 8 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 239 positions in the final dataset.

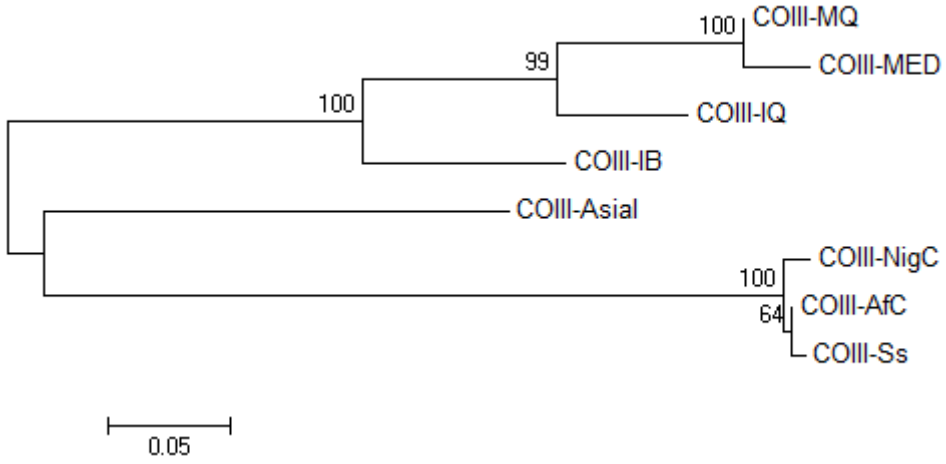


Figure 5.2(L): Maximum likelihood tree showing the relationship of COIII nucleotide sequences. Tree was constructed using the Tamura 3-parameter model with a discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 0.5393)). The analysis involved 8 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 780 positions in the final dataset.

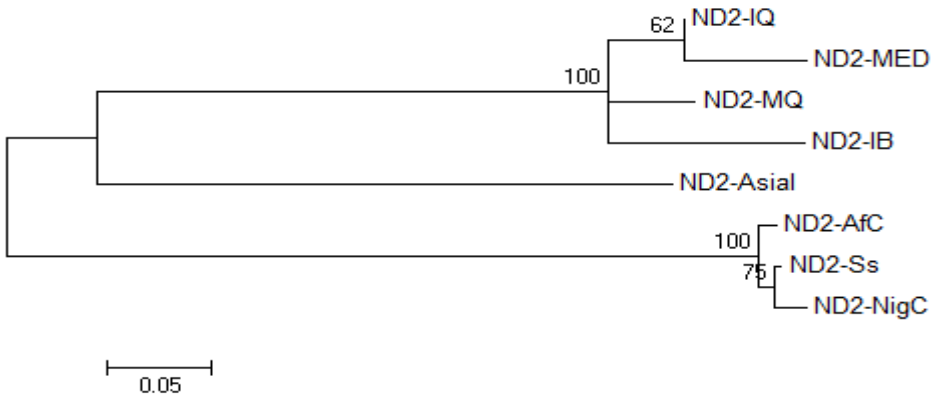


Figure 5.2(M): Maximum likelihood tree showing the relationship of ND2 nucleotide sequences. Tree was constructed using the Hasegawa-Kishino-Yano model with a discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 0.4689)). The analysis involved 8 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 756 positions in the final dataset.

Figure 5.2: The phylogenetic analysis of concatenated nucleotide sequences of 13 mitochondrial PCGs and individual PCGs using maximum likelihood tree.

5.3.4 Identification of primary and secondary endosymbionts among *B. tabaci* populations

Analysis of three cassava and three non-cassava *B. tabaci* population shows presence of P-endosymbiont *Portiera* as well as three S-endosymbionts *Wolbachia*, *Rickettsia* and *Cardinium* in all samples as shown in Table 5.3. These results confirm that the *B. tabaci* populations used in this study contain a combination of endosymbionts. *Portiera* showed the highest number of hits followed by *Wolbachia.* There was no sequence homology found with *Fritschea* and the presence of *Arsenophonus* was found only in non-cassava populations IB, IQ and MQ. The number of hits found for *Cardinium* in MQ (n=19) was much higher than in other populations with 2 hits for IB, 3 for IQ, Ss and NigC, and 4 for AfC.

	IB	IQ	MQ	AfC	Ss	NigC
Portiera	151	228	83	37	67	55
Wolbachia	28	33	42	42	47	47
Rickettsia	23	50	31	24	28	24
Hamiltonella	1	1	0	0	0	0
Cardinium	2	3	19	4	3	3
Arsenophonus	2	2	2	0	0	0
Fritschea	0	0	0	0	0	0

Table 5.3: Summary of primary and secondary endosymbionts present in *B. tabaci* populations.

5.3.5 Phylogenetic analysis of primary endosymbiont *Portiera* based on 16S rDNA sequence

The phylogenetic tree of 16S rDNA sequences of primary endosymbiont *Portiera* were constructed using sequences taken from NCBI database and sequences obtained from this study against *Trialeurodes vaporariorum* and *Buchnera aphidicola* as outgroup is shown in Figure 5.3. Analyses revealed that the phylogeny of three African *B. tabaci* primary endosymbionts of AfC, Ss and NigC were most closely related to *B. tabaci* collected from cassava KEN3 (AF400460) and SSA1-SG3 (KM386389) than UG25 and UG26. Interestingly, primary endosymbiont of MED is closely related to Asia I instead of MQ as observed in mitochondrial gene analysis (Figure 5.3). This may suggest that the phylogeny of *B. tabaci* host tree based on mtCOI was partially congruence with the phylogeny of *B. tabaci* primary endosymbiont. Similar results were found when calculated against pair wise sequence divergence between samples where IB, IQ and MQ shows 100% similarity with MEAM1 but not with MED as observed in phylogenetic analysis (Figure 5.4).

The pair wise sequence divergence result shows that the divergence between primary endosymbionts of *B. tabaci* were generally observed between 0 to 1.5% apart from UG26 and UG25, that diverged by 4.1 to 5.2% and 1.3 to 2.4% respectively.

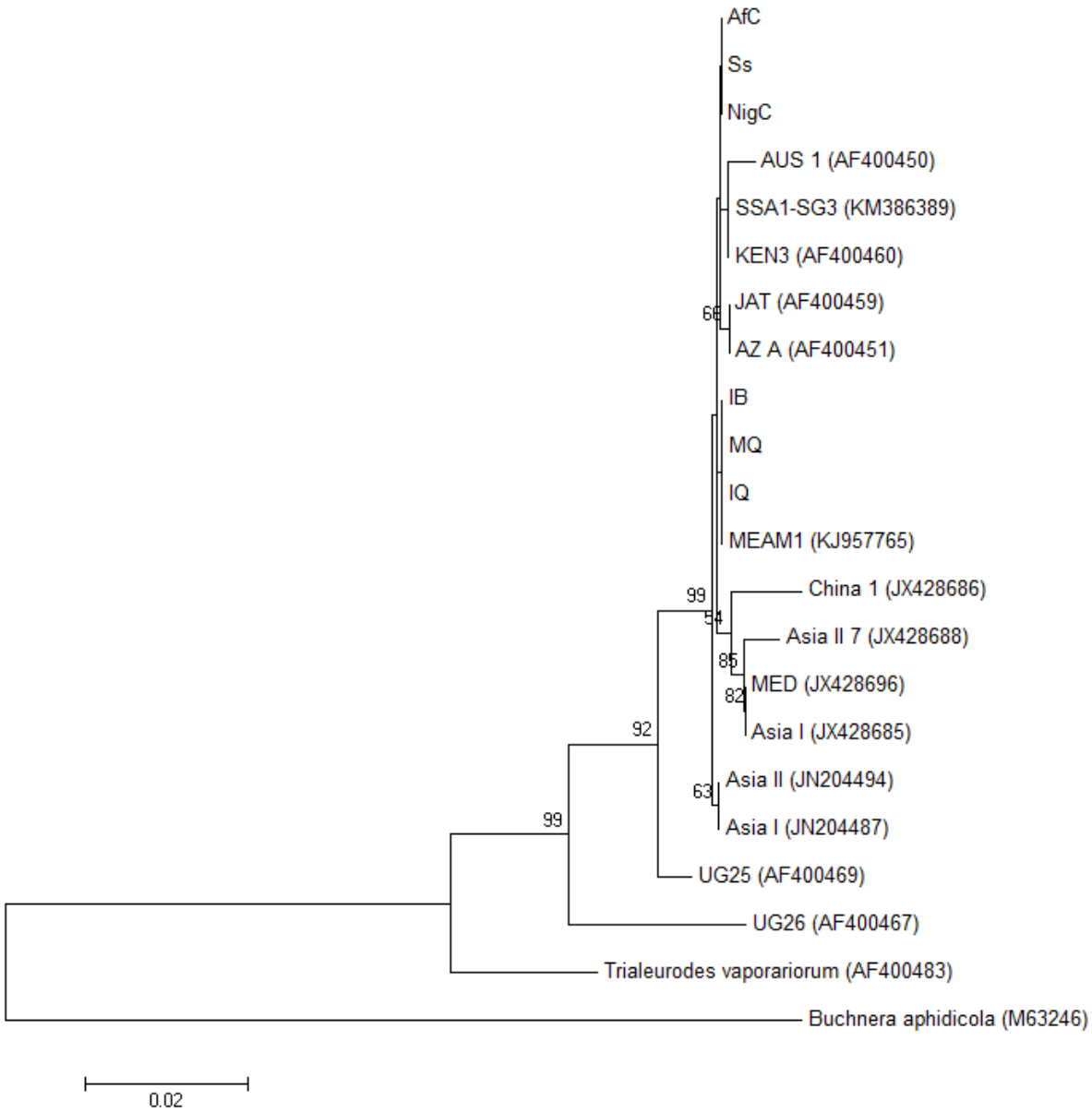


Figure 5.3: Phylogenetic tree showing evolutionary relationships for primary endosymbiont of *B. tabaci* based on 16S rDNA sequence predicted using maximum likelihood method based on Kimura 2-parameter model. The analysis involved 22 nucleotide sequences. All positions with less than 95% site coverage were eliminated. There were a total of 928 positions in the final dataset.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
IB																					
IQ	100																				
MQ	100	100																			
AfC	99.9	99.9	99.9																		
Ss	99.9	99.9	99.9	100																	
NIgC	99.9	99.9	99.9	100	100																
Asia II 7 (JX428688)	99.2	99.2	99.2	99.1	99.1	99.1															
MED (JX428696)	99.7	99.7	99.7	99.6	99.6	99.6	99.6														
China 1 (JX428686)	99	99	99	98.9	98.9	98.9	98.5	98.9													
Asia I (JX428685)	99.7	99.7	99.7	99.6	99.6	99.6	99.6	100	99												
MEAM1 (KJ957765)	100	100	100	99.9	99.9	99.9	99.2	99.7	99	99.7											
SSA1-SG3 (KM386389)	99.8	99.8	99.8	99.9	99.9	99.9	99	99.5	98.8	99.5	99.8										
KEN3 (AF400460)	99.8	99.8	99.8	99.9	99.9	99.9	99	99.5	98.8	99.5	99.8	100									
Asia II (JN204494)	99.8	99.8	99.8	99.9	99.9	99.9	99	99.5	98.8	99.5	99.8	99.8	99.8								
Asia I (JN204487)	99.8	99.8	99.8	99.9	99.9	99.9	99	99.5	98.8	99.5	99.8	99.8	99.8	100							
UG26 (AF400467)	95.8	95.8	95.8	95.9	95.9	95.9	95.3	95.7	94.8	95.7	95.8	95.8	95.8	95.9	95.9						
AUS 1 (AF400450)	99.5	99.5	99.5	99.6	99.6	99.6	98.7	99.1	98.5	99.1	99.5	99.7	99.7	99.5	99.5	95.5					
JAT (AF400459)	99.8	99.8	99.8	99.9	99.9	99.9	99	99.5	98.8	99.5	99.8	99.8	99.8	99.8	99.8	95.8	99.5				
AZ A (AF400451)	99.8	99.8	99.8	99.9	99.9	99.9	99	99.5	98.9	99.5	99.8	99.8	99.8	99.8	99.8	95.8	99.6	100			
UG25 (AF400469)	98.6	98.6	98.6	98.7	98.7	98.7	98.2	98.5	97.6	98.5	98.6	98.6	98.6	98.7	98.7	96.5	98.3	98.6	98.6		
Buchnera aphidicola (M63246)	83.4	83.4	83.4	83.4	83.4	83.4	83.1	83.2	82.4	83.2	83.4	83.4	83.4	83.4	83.4	83.4	83	83.2	83.2	83.7	
Trialeurodes vaporariorum (AF400483)	95.4	95.4	95.4	95.5	95.5	95.5	94.6	95	94.4	95	95.4	95.6	95.6	95.5	95.5	94.4	95.3	95.5	95.5	94.9	84.9

Figure 5.4: Percentage similarity between 22 nucleotide sequences of the primary endosymbiont *Portiera* identified in *B. tabaci*. Analyses were conducted using MEGA6 and all positions with less than 95% site coverage were eliminated. There were a total of 928 positions in the final dataset.

5.4 Discussion

As advances in next-generation sequencing are growing rapidly, many RNA-seq datasets of *B. tabaci* have been generated. However, only three MED (Wang *et al*., 2013), Asia I (Tay *et al*., 2014) and New World I (Thao *et al*., 2004) *B. tabaci* mitogenome were available when carrying out analyses. To obtain more information about mitochondrial genes and their roles in different *B. tabaci* populations, RNA-seq data of six different *B. tabaci* populations were used to analyse their mitochondrial genes. The published mitogenome of Asia I was used to identify mitochondrial gene sequences present in each sample using Illumina sequencing reads. Using this approach, most mitochondrial genes for each sample were identified.

The comparison of six samples used in this study against published MED and Asia I mitogenome sequences indicated the number of genes present varied in numbers and also in their lengths. Expression profiles of all mitochondrial genes are different in different developmental stages as discovered in MED mitochondrial gene expression analysis (Wang *et al*., 2013). Although the complete mitogenome of all six samples was not retrieved using this technique, the data has provided necessary information for designing PCR primer sequences for further analysis. These data then can be used to identify differences within population, gene expression analysis, coding and non-coding regions and gene arrangements.

Many phloem-feeding insects harbour array of vertically transmitted prokaryotes within a population and also in different populations as observed in aphids and psyllids (Zchori-Fein and Brown, 2002). Similar results were found in this study, when we compared different populations of *B. tabaci* to identify primary and secondary symbionts present within populations. The results show that the symbionts present in each *B. tabaci* population varied and appeared compositionally diverse in nature as found in other phloem-feeding insects (Zchori-Fein and Brown, 2002). The primary endosymbiont of *B. tabaci* were found in all six samples as expected. The secondary symbionts were only present in some samples. Similar endosymbiotic associations were observed in aphids and other insect species and thus it is supported that primary symbionts have long association with their host and are obligate and necessary for host survival (Zchori-Fein and Brown, 2002).

6. Conclusion

Advances in high-throughput sequencing has led to sequencing and annotating genetic profiles of model and non-model organisms effectively and efficiently (O'Neil *et al*., 2013). To understand the genetic profile of different *B. tabaci* populations, we sequenced three cassava and three non-cassava *B. tabaci* populations using Illumina paired-end sequencing. Due to the absence of *B. tabaci* reference genome, we had to use *de-novo* assembly programs to assemble and annotate all samples. Due to the number of assembly programs available, it was crucial to compare and validate the assembly results using different evaluation methods available for RNA-seq datasets. Other important factors while assembling raw RNA-seq data are quality, input parameters and importantly the size of k-mer. Here, we used a multi k-mer approach to assemble the data as the transcriptome assembly using higher k-mer length can recover longer and contiguous fragments, while low k-mer length recovers poorly expressed transcripts (Surget-Groba and Montoya-Burgos, 2010). To ensure the quality of data, we used Cutadapt and FastQC program to remove adaptor contamination and base calling errors. In this study, we used four de novo assemblers: Trinity, CLC, SOAPdenovo-Trans and Velvet followed by Oases to assemble, compare and evaluate the performance of each assembly program based on contig statistics, assembly statistics and completeness of assembly. To compare and evaluate the performance of assemblies, we used TransRate, DETONATE, CRB-BLAST, BUSCO and TransDecoder programs and found that the assemblies generated using the Trinity performed best compared to other assembly programs but there was not any single dominant k-mer value observed in all results. To overcome this, we used a clustered assembly strategy by combining Trinity assemblies with the k-mer 25, 27 and 29 into one for further analysis. Doing this, it resulted in many redundant contigs which can cause bias in downstream analysis and may be much harder to filter out at next level than now. To do so, we used CD-HIT-EST program which removed all redundant contigs using the percentage similarity which is 100% in this case.

As currently (October 3, 2016), there are only about 12,094 EST, 14,359 protein and 260,065 nucleotide sequences available for *B. tabaci* on NCBI (Source: http://www.ncbi.nlm.nih.gov), identifying and annotating essential *B. tabaci* genes are important to understand the roles of those genes in the biology of *B. tabaci*. These datasets will also provide useful information to understand and compare sequence divergence

between cassava and non-cassava colonizing *B. tabaci* populations. We used final non-redundant clustered assembly for annotation using BLASTx to identify contigs that are identical with the ones available in non-redundant (nr) NCBI nucleotide database using a cut-off E-value of 10^{-3}. We found that ~60% of contigs showed strong homology with database sequences with ~40% of contigs have a sequence similarity ranging from 40 to 60 percent. Overall, this results showed significant assembly quality, despite having only 14359 protein sequence deposits in NCBI database. While analysing all contigs against GO database, differentiation between cassava and non-cassava samples was not possible, and these results were very similar to previous study undertaken by Wang *et al*., 2012 when they compared the GO annotation results of MED, MEAM1 and Asia II 3 and found that all different populations of *B. tabaci* shares common function distribution. Whereas, when we analysed all samples to find possible biological pathways, we discovered some differences between published *B. tabaci* results and our findings. This may be due to methods used to do so or that the data sets were generated using different methods. Domain prediction results clearly differentiate the domain sharing between cassava and non-cassava populations as the results were highly varied in terms of common sharing domains between samples.

We have also identified mitochondrial genes from all samples by mapping against Asia I mitochondrial genome. It was not possible to identify all mitochondrial genes. The primary endosymbiont *Portiera* was identified in all samples as well as a combination of secondary endosymbionts.

Overall, the assembly, annotation and evaluation of six *B. tabaci* transcriptome dataset provides useful information about the genes and their role in the biology of *B. tabaci* as well provide useful resource to study and compare different *B. tabaci* populations and their specificity with different hosts. Our results and methods used in this study will also provide some guidelines for further transcriptome analysis studies and will serve as a useful repository.

7. References

Altschul, S., Gish, W., Miller, W., Myers, E. and Lipman, D. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403-10.

Ansorge, W. J. 2009. Next-generation DNA sequencing techniques. *New Biotechnology* **25**: 195-203.

Ariyo, O. A., Atiri, G. I., Dixon, A. G. O. and Winter, S. 2006. The use of biolistic inoculation of cassava mosaic begomoviruses in screening cassava for resistance to cassava mosaic disease. *Journal of Virological Methods* **137**: 43-50.

Aubry, S., Kelly, S., Kumpers, B. M. C., Smith-Unna, R. D. and Hibberd, J. M. 2014. Deep evolutionary comparison of gene expression identifies parallel recruitment of trans-factors in two independent origins of C_4 photosynthesis. *PloS Genetics* **10**: e1004365.

Basit. M., Saeed. S., Saleem. M. A. and Sayyed. A. H. 2013. Can resistance in *Bemisia tabaci* (Homoptera: Aleyrodidae) be overcome with mixtures of neonicotinoids and insect growth regulators?. *Crop Protection* **44**: 135-141.

Bedford, I. D., Briddon, R. W., Brown, J. K., Rosell, R. C. and Markham, P. G. 1994. Geminivirus transmission and biological characterisation of *Bemisia tabaci* (Gennadius) biotypes from different geographic regions. *Annals of Applied Biology* **125**: 311-325.

Bellotti, A. C. and Arias, B. 2001. Host plant resistance to whiteflies with emphasis on cassava as a case study. *Crop Protection* **20**: 813-823.

Berglund, E. C., Kiialainen, A. and Syvanen, A. 2011. Next-generation sequencing technologies and applications for human genetic history and forensics. *Investigative Genetics* **2**: 23.

Bing, X., Ruan, Y., Rao, Q., Wang, X. and Liu, S. 2013. Diversity of secondary endosymbionts among different putative species of the whitefly *Bemisia tabaci*. *Insect Science* **20**: 194-206.

Bird, J. 1957. A whitefly transmitted mosaic *Jatropha gossypifolia*. *Technical Paper, Agricultural Experiment Station, Puerto Rico* **22**: 1-35.

Bird, J. and Maramorosch, K. 1978. Viruses and virus diseases associated with whiteflies. *Advances in Virus Research* **71**: 233-260.

Blagbrough, I. S., Bayoumi, S. A. L., Rowan, M. G. and Beeching, J. R. 2010. Cassava: an appraisal of its phytochemistry and its biotechnological prospects. *Phytochemistry* **71**: 1940-1951.

Boykin, L. and De Barro, P. 2014. A practical guide to identifying members of the *Bemisia tabaci* species complex: and other morphologically identical species. *Frontiers in Ecology and Evolution* DOI=10.3389/fevo.2014.00045

Boykin, L. M., Bell, C. D., Evans, G., Small, I. and Barro, P. J. D. 2013. Is agriculture driving the diversification of the *Bemisia tabaci* species complex (Hemiptera: Sternorrhyncha: Aleyrodidae)?: Dating, diversification and biogeographic evidence revealed. *BMC Evolutionary Biology* **13**: 228.

Boykin, L. M., Shatters, R. G., Rosell, R. C., McKenzie, C. L., Bagnall, R. A., De Barro, P. J. and Frohlich, D. R. 2007. Global relationships of *Bemisia tabaci* (Hemiptera: Aleyrodidae) revealed using Bayesian analysis of mitochondrial COI DNA sequence. *Molecular Phylogenetics and Evolution* **44**: 1306-1319.

Brown, J. K. and Bird, J. 1992. Whitefly-transmitted geminiviruses and associated disorders in the Americas and the Caribbean Basin. *Plant Disease* **76**: 220-225.

Brunt, A. A. 1986. Transmission of Disease. In: *Bemisia tabaci - A Literature Survey*, pp 43-50 (ed., Cock, M. J. W.). CAB International Institute of Biological Control, Ascot.

Buermans, H. P. J. and Dunnen, J. T. D. 2014. Next generation sequencing technology: advances and applications. *Molecular Basis of Disease* **1842**: 1932-1941.

Byrne, D. N. and Bellows, T. S. Jr. 1991. Whitefly biology. *Annual Review of Entomology* **36**: 431-458.

Cardoso, A. P., Mirione, E., Ernesto, M., Massaza, F., Cliff, J., Haque, M. R. and Bradbury, J. H. 2005. Processing of cassava roots to remove cyanogens. *Journal of Food Composition and Analysis* **18**: 451-460.

Chavez, A. L., Sanchez, T., Jaramillo, G., Bedoya, J. M., Echeverry, J., Bolanos, E. A., Ceballos, H. and Iglesias, C. A. 2005. Variation of quality traits in cassava roots evaluated in landraces and improved clones. *Euphytica* **143**: 125-133.

Cheek, S. and Macdonald, O. 1993. Preventing the establishment of *Bemisia tabaci* in the United Kingdom, pp 377-380. 1993 BCPC Monograph no.54. Plant Health and the European Single Market.

Chowda-Reddy, R. V., Kirankumar, M., Seal, S. E., Muniyappa, V., Valand, G. B., Govindappa, M. R. and Colvin, J. 2012. *Bemisia tabaci* phylogenetic groups in India and the relative transmission efficacy of *Tomato leaf curl Bangalore virus* by an indigenous and an exotic population. *Journal of Integrative Agriculture* **11**: 235-248.

Cohen, S. and Harpaz, I. 1964. Periodic, rather than continual, acquisition of a new tomato virus by its rector, the tobacco whitefly (*Bemisia tabaci* Gennadius). *Entomologica Experimentalis Applicata* **7**: 155-166.

Cohen, S., Kern, J., Harpaz, I. and Ben Joseph, R. 1988. Epidemiological studies of the *tomato yellow leaf curl virus* (TYLCV) in the Jordan Valley, Israel. *Phytoparasitica* **16**: 259-270.

Costa, A. S. and Russell, L. M. 1975. Failure of *Bemisia tabaci* to breed on cassava plants in Brazil (Homoptera: Aleyroideae). *Ciencia E Cultura* **27**: 390-399.

Czosnek, H. and Brown, J. K. 2010. The whitefly genome - White Paper: A proposal to sequence multiple genomes of *Bemisia tabaci*, pp 503-532 (ed., Stanley, P. A., Naranjo, S. E.). Bemisia, Bio-Nomics and Management of a Global Pest. Springer Science, Dordrecht, The Netherlands.

Dinsdale, L., Cook, C., Riginos, Y., Buckley, M. and De Barro, P. J. 2010. Refined global analysis of *Bemisia tabaci* (Hemiptera: Sternorrhyncha: Aleyrodoidea: Aleyrodidae)

mitochondrial cytochrome oxidase I to identify species level genetic boundaries. *Annals of Entomological Society of America* **103**: 196-208.

Dutt, N., Briddon, R. W. and Dasgupta, I. 2005. Identification of a second begomovirus, Sri Lankan cassava mosaic virus, causing cassava mosaic disease in India. *Archives of Virology* **150**: 2101-2108.

Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using Phred. II. error probabilities. *Genome Research* **8**: 186-194.

Fauquet, C.M. and Fargette, D. 1990. African cassava mosaic virus: Aetiology, epidemiology and control. *Plant Disease* **74**: 404-011.

Finn, R. D., Clements, J. and Eddy, S. R. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research* **39**: W29-W37.

Food and Agriculture Organization of the United Nations [FAO]. 2013. Cassava production statistics. http://faostat.fao.org accessed on 17/04/2014.

Fransen, J. J. 1994. *Bemsia tabaci* in the Netherlands; here to stay?. *Pesticide Science* **42**: 129-134.

Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W. 2012. CD-HIT: accelerated for clustering the next generation sequencing data. *Bioinformatics* **28**: 3150-3152.

Gotz, S., Garcia-Gomez, J., Terol, J., Williams, T., Nagaraj, S., Nueda, M., Robles, M., Talon, M., Dopazo, J. and Conesa, A. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research* **36**: 3420-3435.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N. and Regev, A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**: 644-652.

Gruenheit, N., Deusch, O., Esser, C., Becker, M., Voelckel, C. and Lockhart, P. 2012. Cutoffs and k-mers: implications from a transcriptome study in allopolyploid plants. *BMC Genomics* **13**: 92.

Herrera Campo, B. V., Hyman, G. and Bellotti, A. 2011. Threats to cassava production: known and potential geographic distribution of four key biotic constraints. *Food Security* **3**: 329-345.

Horowitz, A. R. 1986. Population dynamics of *Bemisia tabaci* (Gennadius): With special emphasis on cotton fields. *Agriculture, Ecosystems and Environment* **17**: 37-47.

Hu, J., De Barro, P. J., Zhao, H., Nardi, F., Wang, J. and Liu, S. S. 2011. An extensive field survey combined with a phylogenetic analysis reveals rapid and widespread invasion of two alien whiteflies in China. *PLoS ONE* **6**: e16061.

Kalra, S., Puniya, B., Kulshreshtha, D., Kumar, S., Kaur, J., Ramachandran, S. and Singh, K. 2013. *De nono* transcriptome sequencing reveals important molecular networks and metabolic pathways of the plant, *Chlorophytum borivilianum*. *PLoS ONE* **8**: e83336.

Kikuchi, Y. 2009. Endosymbiotic bacteria in insects: their diversity and culturability. *Microbes and Environments* **24**: 195-204.

Kogan, M. and Turnipseed, S. G. 1987. Ecology and management of soybean arthropods. *Annual Review of Entomology* **32**: 507-538.

Kristensen, S. B. P., Birch-Thomsen, T., Rasmussen, K., Rasmussen, L. V. and Traore, O. 2014. Cassava as an energy crop: a case study of the potential for an expansion of cassava cultivation for bioethanol production in Southern Mali. *Renewable Energy* **66**: 381-390.

Krogh, A., Larsson, B., Heijne, G. V. and Sonnhammer, E. L. L. 2001. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of Molecular Biology* **305**: 567-580.

Legg, J. P. 1999. Emergence, spread and strategies for controlling the pandemic of cassava mosaic virus disease in east and central Africa. *Crop Protection* **18**: 627-637.

Legg, J. P., Jeremiah, S. C., Obiero, H. M., Maruthi, M. N., Ndyetabula, I., Okao-Okuja, G., Bouwmeester, H., Bigirimana, S., Tata-Hangy, W., Gashaka, G., Mkamilo, G., Alicai, T. and Lava Kumar P. 2011. Comparing the regional epidemiology of the

cassava mosaic and cassava brown streak virus pandemics in Africa. *Virus Research* **159**: 161-170.

Legg, J. P., Owor, B., Sseruwagi, P. and Ndunguru. J. 2006. Cassava mosaic virus disease in East and Central Africa: epidemiology and management of a regional pandemic. *Advances in Virus Research* **67**: 355-418.

Li, B., Fillmore, N., Bai, Y., Collins, M., Thomson, J. A., Stewart, R. and Dewey, C. N. 2014. Evaluation of *de novo* transcriptome assemblies from RNA-Seq data. *Genome Biology* **15**: 553.

Liu, B., Preisser, E. L., Chu, D., Pan, H., Xie, W., Wang, S., Wu, Q., Zhou, X. and Zhang, Y. 2013. Multiple forms of vector manipulation by a plant-infecting virus: *bemisia tabaci* and *tomato yellow leaf curl virus*. *Journal of Virology* **87**: 4929-4937.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. and Law, M. 2012. Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology* DOI:10.1155/2012/251364.

Luan, J., Li, J., Varela, N., Wang, Y., Li, F., Bao, Y., Zhang, C., Liu, S. and Wang, X. 2011. Global analysis of the transcriptional response of whitefly to tomato yellow leaf curl china virus reveals the relationship of coevolved adaptations. *Journal of Virology* **85**: 3330-3340.

Mardis, E. R. 2008. Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics* **9**: 387-402.

Markham, P. G., Bedford, I. D., Liu, S. and Pinner, M. S. 1994. The transmission of geminiviruses by *Bemisia tabaci*. *Pesticide Science* **42**: 123-128.

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet* **17**: 10-12.

Martin, N. A. 1999. Whitefly: biology, identification and life cycle. *Crop and food research* **91**: 1-8.

Mound, L. A. 1963. Host-correlated variation in *Bemisia tabaci* (Gennadius) (Homoptera: Aleyroideae). *Royal Entomological Society of London* **38**: 171-180.

Mound, L. A. and Halsey, S. H. 1978. Whitefly of the world. A systematic catalogue of the Aleyrodidae (Homoptera) with Host Plant and Natural Enemy Data. *John Wiley and Sons, Chichester*.

Mouton, L., Thierry, M. and Henri, H. 2012. Evidence of diversity and recombination in Arsenophonus symbionts of the *Bemisia tabaci* species complex. *BMC Microbiology* **12**: S10.

Mugerwa, H., Rey, M. E. C., Alicai, T., Ateka, E., Atuncha, H., Ndunguru, J. and Sseruwagi, P. 2012. Genetic diversity and geographic distribution of *Bemisia tabaci* (Gennadius) (Hemiptera: Aleyrodidae) genotypes associated with cassava in East Africa. *Ecology and Evolution* **2**: 2749-2762.

Mundry, M., Bornberg-Bauer, E., Sammeth, M. and Feulner, P. G. D. 2012. Evaluating characteristics for *de novo* assembly software on 454 transcriptome data: a simulation approach. *PLoS ONE* **7**: e31410.

Nagalakshmi, U., Waern, K. and Snyder, M. 2010. RNA-Seq: a method for comprehensive transcriptome analysis. *Current Protocols in Molecular Biology* 4.11.1-4.11.13.

Nassar, N. M. A. and Ortiz, R. 2007. Cassava improvement: challenges and impacts. *Journal of Agricultural Science* **145**: 163-171.

Nielsen, H. 2017. Predicting secretory proteins with SignalP. *Methods in Molecular Biology* **1611**: 59-73.

Night, G., Asiimwe, P., Gashaka, G., Nkezabahizi, D., Legg, J. P., Okao-Okuja, G., Obonyo, R., Nyirahorana, C., Mukakanyana, C., Mukase, F., Munyabarenzi, I. and Mutumwinka, M. 2011. Occurrence and distribution of cassava pests and diseases in Rwanda. *Agriculture, Ecosystems and Environment* **140**: 492-497.

O'Neil, S. and Emrich S. J. 2013. Assessing *de novo* transcriptome assembly metrics for consistency and utility. *BMC Genomics* **14**: 465.

Otim-Nape, G. W., Bua, A., Baguma, Y. and Thresh, J. M. 1997. Epidemic of severe cassava mosaic disease in Uganda and efforts to control it. *African journal of root and tuber crops* **2**: 42–43.

Perring, T. M., Cooper, A. D., Rodriguez, R. J., Farrar, C. A. and Bellows Jr., T. S. 1993. Identification of a whitefly species by economic and behavioural studies. *Science* **259**: 74-77.

Perring, T.M. 2001. The *Bemisia tabaci* species complex. *Crop Protection* **20**: 725e737.

Polston, J. E, Hiebert, E., McGovern, R. J., Stansly, P. A. and Schuster, D. J. 1993. Host range of tomato mottle virus, a new geminivirus infecting tomato in Florida. *Plant Disease* **77**: 1181-1184.

Poulston, J. E. and Anderson, P. K. 1997. The emergence of whitefly-transmitted geminiviruses in tomato in the Western hemisphere. *Plant Disease* **81**: 1358-1369.

Price, J.F., Schuster, D.J. and Short, D.E. 1987. Managing sweetpotato whitefly. Greenhouse Grower December: 55-57.

Rauch, N. and Nauen, R. 2004. Characterization and molecular cloning of a glutathione S-transferase from the whitefly *Bemisia tabaci* (Hemiptera: Aleyrodidae). *Insect Biochemistry and Molecular Biology* **34**: 321-329.

Samudra I.M. and A. Naito. 1991. Varietal resistance of soybean to whitefly *Bemisia tabaci* Genn. In: Proceeding of FinalSeminar on the Strengthening of Pioneering Research for Palawija Crop Production (ATA-378). Central Research Institute for Food Crops, Bogor, Indonesia, pp. 51-55.

Schulz, M. H., Zerbino, D. R., Vingron, M. and Birney, E. 2012. Oases: robust *de novo* RNA-Seq assembly across the dynamic range of expression levels. *Bioinformatics* **28**: 1086-1092.

Sharma, J., Chakraborty, S. and Uddin, A. 2014. Codon usage bias in two hemipteran insect species: *Bemisia tabaci* and *Homalodisca coagulate*. *Advances in Biology* doi:10.1155/2014/145465

Shendure, J. and Ji, H. 2008. Next-generation DNA sequencing. *Nature Biotechnology* **26**: 1135-1145.

Shokralla, S., Spall, J. L., Gibson, J. F. and Hajibabaei, M. 2012. Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology* **21**: 1794-1805.

Simao, F. A., Waterhouse, R. M., Loannidis, P., Kriventseva, E. V. and Zdobnov, E. M. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* doi: 10.1093/bioinformatics/btv351

Sloan, D. and Moran, N. 2012. Endosymbiotic bacteria as a source of carotenoids in whiteflies. *Biology Letters* **23**: 986-989.

Smith-Unna, R. D., Boursnell, C., Patro, R., Hibberd, J. M. and Kelly, S. 2015. TransRate: reference free quality assessment of *de novo* transcriptome assemblies. *bioRxiv* doi: http://dx.doi.org/10.1101/021626

Surget-Groba, Y. and Montoya-Burgos, J. I. 2010. Optimization of *de novo* transcriptome assembly from next-generation sequencing data. *Genome Research* **20**: 1432-1440.

Tajebe, L. S., Guastella, D., Cavalieri, V., Kelly, S. E., Hunter, M. S. and Lund, O. S. 2014. Diversity of symbiotic bacteria associated with Bemisia tabaci (Homoptera: Aleyrodidae) in cassava mosaic disease pandemic areas of Tanzania. *Annals of Applied Biology* **166**: 297-310.

Tay, W. T., Elfekih, S., Court, L., Gordon, K. H. and Barro, P. J. D. 2014. Complete mitochondrial DNA genome of *Bemisia tabaci* cryptic pest species complex Asia I (Hemiptera: Aleyrodidae). *Mitochondrial DNA* doi:10.3109/19401736.2014.926511

Taylor, N., Chavarriaga, P., Raemakers, K., Siritunga, D. and Zhang, P. 2004. Development and application of transgenic technologies in cassava. *Plant Molecular Biology* **56**: 671-688.

Thao, M. L., Baumann, L. and Baumann, P. 2004. Organization of the mitochondrial genome of whiteflies, aphids, and psyllids (Hemiptera, Sternorrhyncha). *BMC Evolutionary Biology* **4**: 25.

Thiel, T., Michalek, W., Varshney, R. K. and Graner, A. 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theoretical and Applied Genetics* **106**: 411-422.

Thompson, J. D., Higgins, D. G. and Gibson, T. J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**: 4673-4680.

Vinogradov, A. 2003. DNA helix: the importance of being GC-rich. *Nucleic Acids Research* **31**: 1838-1844.

Wang, H., Yang, J., Boykin, L., Zhao, Q., Li, Q., Wang, X. and Liu, S. 2013. The characteristics and expression profiles of the mitochondrial genome for the Mediterranean species of *Bemisia tabaci* complex. *BMC Genomics* **14**: 401.

Wang, H., Yang, J., Boykin, L., Zhao, Q., Wang, Y., Liu, S. and Wang, X. 2014. Developing conserved microsatellite markers and their implications in evolutionary analysis of the *Bemisia tabaci* complex. *Scientific Reports* **4**: 6351.

Wang, J., He, W., Su, Y., Bing, X. and Liu, S. 2014. Molecular characterization of soluble and membrane-bound trehalases of the whitefly, *Bemisia tabaci*. *Archives of Insect Biochemistry and Physiology* **85**: 216-233.

Wang, X., Luan, J., Li, J., Bao, Y., Zhang, C. and Liu, S. 2010. *De novo* characterization of a whitefly transcriptome and analysis of its gene expression during development. *BMC Genomics* **11**: 400.

Wang, X., Luan, J., Li, J., Su, Y., Xia, J. and Liu, S. 2011. Transcriptome analysis and comparison reveal divergence between two invasive whitefly cryptic species. *BMC Genomics* **12**: 458.

Wang, X., Zhao, Q., Luan, J., Wang, Y., Yan, G. and Liu, S. 2012. Analysis of a native whitefly transcriptome and its sequence divergence with two invasive whitefly species. *BMC Genomics* **13**: 529.

Wei, L., Liu, Y., Dubchak, I., Shon, J. and Park, J. 2002. Comparative genomics approaches to study organism similarities and differences. *Journal of Biomedical Informatics* **35**: 142-150.

Wydra, K. and Verdier, V. 2002. Occurrence of cassava diseases in relation to environmental, agronomic and plant characteristics. *Agriculture, Ecosystems and Environment* **93**: 211-226.

Xie, W., Meng, Q., Wu, Q., Wang, S., Yang, X., Yang, N., Li, R., Jiao, X., Pan, H., Liu, B., Su, Q., Xu, B., Hu, S., Zhou, X. and Zhang, Y. 2012. Pyrosequencing the *Bemisia tabaci* transcriptome reveals a highly diverse bacterial community and a robust system for insecticide resistance. *PLoS ONE* **7**: e35181.

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Haung, W., He, G., Gu, S., Li, S., Zhou, X., Lam, T., Li, Y., Xu, X., Wong, G. K. and Wang, J. 2014. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **30**: 1660-1666.

Yang, N., Xie, W., Jones, C. M., Bass, C., Jiao, X., Yang, X., Liu, B., Li, R. and Zhang, Y. 2013. Trancriptome profiling of the whitefly Bemisia *tabaci* reveals stage-specific gene expression signatures for thiamethoxam resistance. *Insect Molecular Biology* **22**: 485-496.

Zchori-Fein, E. and Brown, J. K. 2002. Diversity of prokaryotes associated with Bemisia tabaci (Gennadius) (Hemiptera: Aleyrodidae). *Annals of the Entomological Society of America* **95**: 711-8.

Zerbino, D. R. and Birney, E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* **18**: 821-829.

Zhang, Y., Hao, Y., Si, F., Ren, S., Hu, G., Shen, L. and Chen, B. 2014. The de novo transcriptome and its analysis in the worldwide vegetable pest, *Delia antiqua* (Diptera: Anthomyiidae). *G3: Genes|Genomes|Genetics* **4(5)**: 851-859.

Zhao, Q., Wang, Y., Kong, Y., Luo, D., Li, X. and Hao, P. 2011. Optimizing *de novo* transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics* **12**(S 14): S2.

Zhou, X., Ren, L., Meng, Q., Li, Y., Yu, Y. and Yu, J. 2010. The next-generation sequencing technology and application. *Protein Cell* **1**: 520-536.