A study of the Southern African begomovirus pathosystem:

Determining the diversity of whitefly transmitted geminiviruses (WTG) infecting indigenous plants in South Africa.

Onesimo Hendrieta Matole
(201495902)

Dissertation
Submitted in fulfilment of the requirements for the degree
Magister Scientiae (M.Sc.)
in
Biochemistry
in the
Faculty of Science
at the
University of Johannesburg

Supervisor: Dr. Lindy L. Esterhuizen
Co-supervisors: Prof. Michelle van der Bank

December 2018

ABSTRACT

With an ever increasing world population, there is growing pressure on agricultural production to keep up with the demand for food, fuel and fiber products. Agricultural intensification has resulted in the conversion of approximately one-half of the Earth's terrestrial surface to agricultural crops and range-land. This has significantly changed the natural landscape around agricultural borders, with the development of agro-ecological interface' between wild ecosystems and crop or pasture lands. These areas are populated with crop plants, weeds of crop systems and non-crop plants that consist of both native and introduced plants. One very important aspect of this change in landscape, is the influence of microroganisms, particularly plant viruses, that move across the interface between crop and wild plants (virus reservoirs) and significantly impact agricultural production. Due to the potential role of these essentially unknown or poorly studied plant viruses in crop diseases and yield loss, they should be a focus of study. The aim of this project, was to characterize the circular viral DNA viral diversity in infecting indigenous or non-cultivated plant species in various cropping systems in South Africa, using a metagenomics approach.

A survey was carried out in 2014 and 2015 in three Provinces (Limpopo, Mphumalanga and KwaZulu-Natal) of South Africa. Samples were collected from a total of 230 plants growing alongside cultivated crops and screened for viral infection. The viral diversity in these samples were explored using viral amplification by polymerase chain reaction (PCR), rolling circle amplification (RCA), RCA/restriction fragment length polymorphisms) (RCA/RFLP) and the results extended by next generation sequencing (NGS). Initial screening of the plants by PCR, using universal/degenerate begomovirus primers, found that 24% of the non-cultivated plants tested positive for possible begomovirus infection. To further reveal the full circular, viral diversity of the collected samples, they were screened by RCA/RFLP, followed by NGS for virus identification. The use of the RCA/RFLP was found to be a rapid, reproducible method to screen a large number of plant samples for viral biodiversity.

Because of the high cost of NGS, a combined strategy was used, where initially RCA positive samples were pooled, and sequenced in one NGS run (NGS 2014, Pool-1 and 2), and after further screening by PCR, the plant samples suspected to contain circular viral molecules, were sequencing in individual NGS runs (NGS 2015). The pooled NGS, strategy was found to be an effective method for viral discovery, as it did allow the detection and characterization of both previously known viruses and previously unknown, novel virus genomes (Table 4.2). The drawback of the pooling-strategy was that for most viral contigs, the full viral genome could not be assembled and, post-NGS, it required a lengthy follow up procedure to trace the original plant samples, where the viral molecules originated from.

The data provided by the individual NGS runs (NGS 2015) allowed for the identification of both previously known viruses (four strains or variants of Tomato Curly Stunt Virus), and previously unknown, novel virus species (one monopartite and two bipartite begomoviruses, two betasatellite molecules and one genomovirus) in four different, uncultivated host species (Table 4.3). Previous studies have shown the widespread presence of ToCSV throughout South Africa and indeed, four ToCSV genomes were recovered in this study, both from tomato plants and weed species. The two ToCSV genomes recovered from the weeds species, *Malvastrum coromandelianum* and *Acalypha indica*, represent new natural host report for this virus and should be targeted for removal as they are acting as begomovirus reservoir in these cropping areas.

The novel begomovirus species identified in this study, include one monopartite begomovirus, named *Malvastrum curly stunt virus* (MalCSV), that was identified in *M. coromandelianum*, along with an isolate of *Tomato leaf curl Togo* betasatellite (ToLCTGB) and one bipartite begomovirus, named *Sida corlifolia golden mosaic virus* (SiCGMV) that was identified in *S. corlifolia,* along with a novel betasatellite that was termed *Sida corlifolia yellow mosaic* betasatellite (SiCYMB).The MalCSV was identified from a symptomless host, but the SiCGMV and SiCYMB viral molecules were detected in a weed plant displaying symptoms of viral infection, i.e. bright yellow and green mosaic coloration of the apical leaves. The SiCGMV isolate displayed unique 'NW-like' genetic features and phylogenetic analysis grouped SiCGMV with other Corchoroviruses, a subgroup of whitefly transmitted viruses, genetically distinct, and basal to all other begomoviruses. These Old World (OW) viruses, with unique New Word (NW) genetic features, are likely due to distinct evolutionary histories or genetic isolation in their host species. It also lends support to the idea that NW begomoviruses may have originated in the OW and were subsequently disseminated to the NW, and that Corchoviruses may be a remnant of such NW begomoviruses that once populated the OW. Furthermore, this study provided the first reports of beta-satellite molecules in South Africa.

A new bipartite begomovirus was identified in *Phaseolus vulgaris* (bean), for which the name *Phaseolus vulgaris begomovirus* was proposed. Phylogenetic analysis grouped this viral isolate with ScCBV, a member of the'legumoviruses' group. Begomoviruses infecting legumes often cluster phylogenetically as a group between the OW and NW begomoviruses and have collectively been referred to as 'legumoviruses'. No symptoms were observed on the bean plant at the time of collection, thus further investigation into the biological characteristics and possible economic impact of this newly identified virus is required.

The last new circular, DNA-viral molecule identified in this study belongs to the newly formed virus family called *Genomoviridae*. The *Cynoglossum offinale* associated circular DNA virus (COasCV) was identified in the weed species *C. offinale* and phylogenetic analysis clustered COasCV with other gemycircularvirus-like sequences, including *Bemisia*-associated genomovirus

AdDF and Dragonfly associated gemyduguivirus 1, but clearly in a separate branch, suggesting that it may be classified in a distinct genus in this family. Members of this virus family have recently been discovered from a variety of different organisms and environmental samples, but all the viral members of this family are known solely from sequence data, except one species (*Sclerotinia gemycircularvirus-1*; SsHADV-1), for which some biological information has been established. There is a great need for biological characterization of COasCV, to determine the true host, vectors and possible economic impact of this newly identified viral species.

This study demonstrated the presence of numerous circular, DNA-viral molecules from four uncultivated host species, including *Abutilon indicum, Acalypha indica, Cynoglossium officinale, Malvastrum coromendalianum and Sida corlifolia*. At least one or more reports can be found in literature, that indicate each of these weed species to be begomovirus hosts. They are found widely distributed in South Africa, where they not only serve as an important source of primary inoculum, but are virus reservoirs that support mixed infection, recombination, pseudo-recombination and evolution of new viral species. The additional knowledge about alternative hosts of (begomo)-viruses in South Africa provided by this study, can be used to develop effective management practices, such as the implementation of host-free periods, and the removal or eradication of these viral-source plants in agricultural areas. These measures will assist with the successful control of begomovirus infection in crop plants.

As most of the collected plants were volunteer plants in the agricultural environment and at an early growth stage, flowering or fruit structures that are important features for morphological identification were absent. Plant identification was therefore accomplished by DNA barcoding, a molecular method that makes use of short, standardize region of a genome, i.e. the *rbcLa* and *matK* genes for the barcoding of land plants, as a unique identifier for species and provides identification of a specimen to species level. In this study, it allowed for rapid and easy identification of the non-cultivated plants, without any specialized botanical training being required.

In conclusion, the results of this study expanded our understanding of the diversity of small, circular, DNA viruses infecting both crop plants and non-cultivated hosts in three Provinces in South Africa. It highlighted the need for more in depth studies across the agricultural and botanical landscape. Although the crop-infecting viruses are of possible economic concern, the viral diversity discovered in indigenous plants and non-cultivated weeds are expected to be quite dynamic, and will give a truer picture of the evolution and the shaping of viral diversity in South Africa, and Africa, as a whole. The results also emphasize the potential for the emergence of novel begomoviruses by genetic recombination amongst these viral reservoirs. Lasty, the study further lends support to the use of a metagenomic (NGS)-approaches to explore viral diversity, as it once again proved to be a very powerful, although costly, method for viral discovery.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank the Lord God almighty for giving me the strength to do this project with success, without you I could not have achieved my goals. I would also like to express my sincere gratitude to the following people who supported me through this journey:

- To my parents, Mr and Mrs Matole and my family, I thank you so much for believing in me and your support and motivation kept me going when I felt like I cannot go anymore.
- To my wonderful supervisor Dr L.L Esterhuizen, I thank you so much for your guidance and motivation. You showed me kindness and took me through this journey with so much kindness and for that I thank you so much and I will forever be grateful for your support.
- To my co supervisor Prof M van der Bank thank you so much for your support on the project.
- To my former collegues at the African Centre for DNA Barcoding (ACDB), thank you for your motivation.
- I thank the South African National Biodiversity Institute (SANBI) and NRF for financial support.
- To Gershwin and Simon, thank you so much for driving me to field trips, I appreciate your support.

CONTENTS

LIST OF SYMBOLS AND ABBREVIATIONS

A

α	alpha

B

β	beta
BLAST	basic local alignment search tool
bp	base pair
BOLD	barcoding of life database

C

C	complementary sense strand
°C	degree celsius
cm	centimetre
CP	coat protein
CR	common region
cv	cultivar

D

DNA	deoxyribonucleic acid
dNTP	deoxyribonucleotide
ds	double strand

E

EC	Eastern Cape
EDTA	ethylene diamine tetra-acetic acid
EtBr	ethidium bromide

G

g	gram
xg	times gravitational force

H

h	hour
ha	hectare
Hz	hertz

I

ICTV	International Committee on Taxonomy of Viruses
Id	identity
IDT	integrated DNA technologies
IPM	integrated pest management
IR	intergenic region
ITS	internal transcribed spacer

K

kb	kilobase pair
kDa	kilodalton
kg	kilogram
km	kilometre
KZN	Kwazulu-Natal

L

lin	linear

M

M	molar
MB	megabytes
mg	milligram
min	minutes
ml	milliliter
mm	millimetre

mM	millimolar
μg	microgram
μl	microliter
μm	micrometre
μM	micromolar
MP	movement protein
mRNA	messenger ribonucleic acid
mtCOI	mitochondrial cytochrome oxidase I

N

NC	Northern Cape
nc	not computable
NDA	national department of agriculture
NJ	neighbour-joining
nm	nanometre
NSP	nuclear shuttle protein
nt	nucleotide
NW	new world

O

oc	open-circular
ORF	open reading frame
ori	origin of replication
OW	old world

P

PCR	polymerase chain reaction
%	percentage

R

RCA	rolling circle amplification
RCR	rolling circle replication
RDP	recombination detection program
REn	replication enhancer protein

Rep	replication-associated protein
RF	replicative form
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid

S

sc	supercoiled
SCR	sequence-conserved region
SD	standard deviation
sec	seconds
spp	species
ss	single strand
SSAF	sub-Saharan Africa
SWIO	south west indian ocean

T

TrAP	transcription activation protein
Tris	Tris(2-amino-2-hydroxymethyl)-1,3-propandiol

U

U	unit
UV	ultraviolet light

V

V	Virion-sense strand
V2	pre-coat protein
v/v	volume to volume
v/w	volume to weight

W

WC	Western Cape

x

X times

CHAPTER 1

Literature review

1.1 Introduction

With an ever increasing world population, agriculture has to find a way to meet the growing demand for food, fuel and fiber products. Agricultural intensification to meet this growing need has therefore resulted in the conversion of approximately one-half of the Earth's terrestrial surface to agricultural crops and range-land (Alexander et al. 2014; Rattray, 2012; Tilman et al. 2011). This intensified human agricultural activity has also notably changed the Earth's natural landscape on the borders of agricultural areas. It has led to the creation of what has been termed an 'agro-ecological interface' (Burdon and Thrall, 2008). This interface between wild ecosystems and crop or pasture lands (agricultural lands), are populated with crop plants, weeds of crop systems and non-crop plants that consist of both native and introduced plants. Interactions across this interface are the topic of scientific studies, as they have a direct influence on crop production and agricultural management (Alexander et al. 2014). One such topic is the movement of microorganisms across or between these different landscapes. The exchange of particularly plant viruses, between crop fields and wild plants at this interface, is a topic of importance, due to their ubiquity, their dispersal by insect vectors that increases their mobility across landscapes, their ability to colonize both crop and wild plants and lastly, their great potential for causing crop diseases and yield loss (Alexander et al. 2014; Malmstrom and Alexander, 2016; Roossinck et al. 2011).

Traditionally, studies have mainly focused on plant pathogens that cause diseases in crop plants. The vast majority of the approximately 1,000 known plant viruses recently reported by the International Committee on Taxonomy of Viruses (ICTV), have been reported from crop plants (King et al. 2012). This is not only due to the fact that crop pathogens are of great economical concern, but also because most virus infections of wild plants do not result in obvious viral symptoms and have therefore gone unnoticed. With the development of new technologies in molecular biology, particularly high-throughput, deep sequencing techniques (metagenomics approaches) over the last decade, more and more studies are focusing on virus biodiversity in natural environments. These metagenomics studies are not only unearthing a plethora of novel plant viruses from wild plants, from the tropics to Antarctica (Hopkins et al. 2014; Roossinck, 2012b), but are also providing a better understanding of what exist in the natural environment around agricultural areas. In addition, this information will help us to address important agricultural concerns, such as the sources of emerging viral infections, the role of viruses in plant ecology and the role of host biodiversity in disease emergence (Roossinck, 2011).

Numerous weeds of crop systems and non-crop plants growing in and around the agro-ecological interface have been reported to harbor *begomoviruses* (family *Geminiviridae*), many of which are

also known to infect crop plants (Alabi et al. 2016; Ambrozevicius et al. 2002; Mubin et al. 2010). This supports the notion that these plants act as natural hosts or begomovirus reservoirs, keeping them in the agro-system during non-cropping season and supporting mixed infections which lead to recombination, diversity and subsequent evolution of these viruses (Ambrozevicius et al. 2002; Azhar et al. 2011; García-Andrés, et al. 2006; Jones, 2009). Due to the importance of these alternative hosts or reservoirs in the possible evolution of crop-infecting begomoviruses, the objective of this study was to investigate the diversity of circular DNA viral genomic components associated with these hosts in South Africa. A metagenomics approach was used to characterise these begomoviral complexes at the molecular level.

1.2 Section A: Virus

1.2.1 *Geminiviridae:* classification and genome organization

Plant-infecting viruses within the *Geminiviridae* family (Rojas et al. 2005) cause severe economic losses to agricultural production around the world (Harisson, 1985). They infect a wide range of crop species and are considered one of the largest and most economically important plant infecting virus families (Brown et al. 2015). In sub-Saharan Africa, geminiviruses particularly endanger food security, due to the fact that they infect crops such as cassava and maize which are used as staple foods (Bernado et al. 2013; Varsani et al. 2008). Their genome is composed of single stranded circular DNA component(s), which are packed within twinned icosahedral capsules, and they are transmitted by various insect vectors (whiteflies, leafhoppers, treehoppers and aphids) (Bernado et al. 2013; Mansoor et al. 2003). The genome of germinivirus can be monopartite, i.e. it contains one single stranded DNA component, also called DNA-A, or it can be bipartite, i.e. it contains two single stranded DNA components of similar size, called DNA-A and DNA-B (Mubin et al. 2010; Rojas et al. 2005).

The family *Geminiviridae* consists of nine genera, namely *Begomovirus, Mastrevirus, Curtovirus*, *Topocuvirus* and five newly created genera *Becurtovirus, Turncurtovirus* and *Eragrovirus*, *Capulavirus* and *Grablovirus*. They are classified according to the host range, genome pair-wise sequence identities, insect vector and genome organization (Brown et al. 2015; Singht et al. 2012; Varsani et al. 2014; Varsani et al. 2017; Zerbini et al. 2017 Table 1.1).

Viruses in the genus *Topocuvirus* have a monopartite genome and infects dicotyledonous plants and are transmitted by treehoppers. This genus has only one member, which is *Tomato pseudo-curly top virus.* The two open reading frames on the sense strand encode for the coat protein (V1) and movement protein (V2). The replicase A (C1) and B (C2), replication enhancer protein (C3) and symptom determinant protein (C4) are encoded on the complementary strand. (Rojas et al. 2005).

Table 2.1. The taxonomy of the family *Geminiviridae* including the genome organization, host plant , insect vectors and type species of each of the nine genera (Brown et al. 2015; Fauquet et al. 2003; ICTV, 2017; Varsani et al. 2014, 2017; Zerbini et al. 2017**)**

Genus	Insect Vector	Host range	Genome organisation (Figure 1.1)	Type species
Becurtovirus	Leafhopper	dicotyledonous	Monopartite	*Beet curly top Iran virus*
Begomovirus	Whitefly	dicotyledonous	Monopartite OR Bipartite	*Tomato yellow leaf curl virus Bean golden mosaic virus*
Capulavirus	Aphid	dicotyledonous	Monopartite	*Euphorbia caput-medusae latent virus*
Curtovirus	Leafhopper	dicotyledonous	Monopartite	*Beet Curly Top Virus*
Eragrovirus	~	monocotyledonous	Monopartite	*Eragrostis curvula streak virus*
Grablovirus	Leafhopper	dicotyledonous	Monopartite	*Grapevine red blotch virus*
Mastrevirus	Leafhopper	monocotyledonous	Monopartite	*Maize streak virus*
Topocuvirus	Leafhopper	dicotyledonous	Monopartite	*Tomato pseudo-curly top virus*
Turncurtovirus	Leafhopper	dicotyledonous	Monopartite	*Turnip curly top virus*

~ characteristic unknown

Viruses in the genus *Mastrevirus* have a monopartite genome and are transmitted by leafhopper and infect either monocotyledonous plants or dicotyledonous plants. Their monopartite single stranded genome has four open reading frames (ORF) which are C2, V1, C1 and V2. Mastrevirus have two Rep proteins, the full length Rep protein and RepA. The full length Rep is translated from the spliced transcript of C1 and C2. RepA is translated from C1. The V1 and V2 open reading frames encodes movement protein and coat protein. Members of this genus infect crops such as sugar cane, maize, grass and weeds (Bigarre et al. 1999; Rojas et al. 2005). Mastreviruses are responsible for huge losses of maize production in Africa. Annual yield losses in maize production due to Maize streak disease have been estimated to be US $120-480 million per year (Bernardo et al. 2013; Martin and Shepherd, 2009).

The genus *Curtovirus* has a monopartite genome. Viruses in this genus are transmitted by the leaf hopper *Circulifer tenellus* in a persistent manner, meaning once the leaf hopper picks up the virus it will transmit it for the rest of its life. They infect a wide variety of plants including crops and indigenous plants or non-cultivated plants (Briddon, 2002). The genome of curtoviruses has seven

genes which are encoded by both the sense and anti-sense strand. The sense strand codes for the coat protein (V1), movement protein (V3) and regulatory protein (V2). The complementary strand encode for replication associated proteins (C1), replication enhancer (C3), C2 gene expressing protein with silencing suppressor (C3) and symptom determinant gene (C4). Symptoms displayed by a plant infected with *Curtoviruses* include servere stunting and yellowing and yield reduction (Creamer et al. 2005; Varsani et al. 2014).

Figure 2.1. Genome organisation of isolates of the nine geminivirus genera (LIR: Long intergenic region; SIR: Short intergenic region; CR: common region; The ORFs (V1, V2, V3, C1 etc are colour coded according to the function of their protein products: rep: replication associated protein; ren: replication enhancer; trap: transactivator protein; ss: silencing suppressor; tgs: transcriptional gene silencing; sd: symptom determinant; cp: capsid protein; mp: movement protein; reg: regulatory gene; nsp: nuclear shuttle protein). The hairpin which includes the origin of replication is indicated in the LIR. (Modified from Varsani et al. 2014; Zerbini et al. 2017).

The *Becurtovirus* genus include two well-known species, which are *Spinach Arizona Virus* and *Beet curly top Iran virus*. They infect dicotyledonous plants and are transmitted by leafhoppers

(*Cicrulifer tenellus*) (Varsani et al. 2014). The genome of Becurtoviruses is chimerical and it is believed that it originated from recombination between a Curtovirus and Mastrevirus (Hernandez-Zepend et al. 2013). *Beet curly top Iran virus (*BCTV) is the most important sugar beet pathogen in Iran. It consists of three sense open reading frames which encode for the regulatory protein (V2) and movement protein (V3). The complementary sense strand encodes for Rep and RepA. Symptoms displayed by a plant infected by BCTV include leaf yellowing, swelling of leaf veins and leaf curling (Kardani *et al*. 2013).

The genus *Turncurtovirus* consist of one species, *Turnip curly top virus* (TCTV). TCTV has a monopartite and is transmitted by leafhoppers (Varsani *et al*., 2014). The genome of turncurtoviruses consists of two ORF in the sense strand which encode for coat protein (V1) and movement protein (V2). The complementary strand encode for replication initiator protein (C1), transcription activator protein (C2), replication enhancer protein (C3) and symptom determinant protein (C4) (Briddon et al. 2010; Varsani et al. 2014). Symptoms of a plant infected by TCTV shows leaf thickening, curling and vein swelling (Briddon et al. 2010).

The genus *Eragrovirus* is represented by one species, *Eragrostis curvula streak virus* which has a monopartite genome is transmitted by leafhopper (Varsani et al. 2014). This virus was isolated from a weed (monocotyledonous) in South Africa. The genome organisation of eragroviruses is analogous to that of mastreviruses. It consist of two ORF in the sense strand and two on the complementary strand. The sense strand encodes for coat protein (V1), and movement protein (V2). The complementary strand encode for the replication initiation protein (C2) and transcription activator protein (Bernado et al. 2013; Briddon et al. 2010).

Recently, two new Geminivirus taxonomic groups have been described, viz. the genera *Capulavirus* and *Grablovirus*. The genus *Capulavirus* contains four, dicotyledonous infecting species, including the type member *Euphorbia caput-medusae* and *Alfalfa leaf curl virus*, *French bean severe leaf curl virus* and *Plantago lanceolata latent virus*. Isolates of the species *Alfalfa leaf curl virus* have been shown to be transmitted by an aphid (*Aphis craccivora*) (Varsani et al. 2017). The *Grablovirus* genus currently only contains one species, namely *Grapevine red blotch virus* that also infects dicotyledonous plants (Varsani et al. 2017). The natural vector is likely the three-cornered alfalfa treehopper (*Spissistilus festinus* Say).

The genus *Begomovirus* is the largest and the most diverse genus in the family Geminirividae (Singht et al. 2012) and has viruses that are significant pathogens of commercially important crops, such as tomato, cassava, pepper and cotton (Brown et al. 2001). Begomoviruses are transmitted by the whitefly *Bemisia tabaci* in a persistent and circulative manner. They infect a wide range of eudicot hosts and there are currently more than 288 species of begomoviruses known, that infect various crops and non-cultivated plants worldwide (http://www.ictvonline.org/virusTaxonomy.asp; Yadave et al. 2010). In general, begomoviruses

cause similar symptoms in plants, including upward stunting and leaf distortion and some combination of golden-light green-yellow mosaic/mottle, crumpling, curling, distortion, veinal or interveinal yellowing, and yellow spots in leaves, which results in flower abortion and no yield (Idris et al. 2005; ICTV, 2015). Begomoviruses have traditionally been subdivided into NW (Western Hemisphere) and Old World (OW) (Eastern Hemisphere). The NW members typically have a bipartite genome, with the DNA-A and DNA-B component and both DNA components are needed for infectivity (Hamilton et al. 1983). The OW members are monopartite, with only one DNA component, called a DNA-A. Although monopartite begomoviruses identified in the OW far outnumber those with bipartite genomes (Brown et al. 2015, Leke et al. 2015), the identification of indigenous NW monopartite tomato-infecting begomoviruses (*Tomato leaf deformation virus* (ToLDV) in Peru and *Tomato mottle leaf curl virus* (ToMoLCV) in Brazil has obscured the boundaries between these two groups.

The monopartite begomovirus genome has six genes which are encoded in both the sense and anti-sense strands. The sense strand codes for coat protein (V1) and pre-coat protein (V2). The complementary strand encodes for replication associated protein (C1), transcription activator protein (C2), replication enhancer protein (C3) and symptom determinant protein (C4). Genomic DNA component of monopartite begomovirus is larger (~2, 9 kb) than DNA-A component of New World bipartite begomovirus (Leke et al. 2015; Melgarejo et al. 2013; Rosario et al. 2016). Furthermore, these monopartite begomoviruses are usually found associated with DNA β/Beta satellites and/or satellite-like components called alpha satellite. Both of these satellites have a recognized origin of replication (*ori*) known as nanonucleotide motif (NANTATTAC) in the stem loop structure however, alpha satellites have a nanonucleotide motif similar to that of nanoviruses (family *Nanoviridae*), and therefore are capable of autonomous replication in a host cell (Idris et al. 2011; Lozano et al. 2016; Rosario et al. 2016). Alpha satellites have a highly conserved genome organization encompassing an adenine-rich region of nearly 200 nucleotides and they have been shown to attenuate disease symptoms. Biologically active DNA-*β* molecules are about half the size of the helper DNA-A component, with an average size of ~1350 nucleotides (Briddon et al. 2003). DNA-*β* molecules contain three conserved regions: an A-rich region, a sequence-conserved region (SCR) and an ORF termed *β* C1 (Briddon et al. 2003; Zhou et al. 2003; Bull et al. 2004). They require helper DNA for replication, insect transmission, encapsidation, movement in plants and are involved in symptom development. Betasatellites increase the accumulation of their helper begomovirus and enhance symptoms induced in some host plants due to suppression of RNA interference activity of *β* C1. (Idris et al. 2011; Mubin et al. 2010).

Begomoviruses from the NW are mostly bipartite, composed of two DNA components of about equal in size; DNA-A and DNA-B (Briddon et al. 2003; Bull et al. 2004; Zhou et al. 2003), but only the DNA-A is used in the taxonomic classification of bipartite begomoviruses (King et al. 2012). These components both have a common region (CR) which is a conserved region and protein

coding sequencing on the viral sense strand. In bipartite begomoviruses, DNA-A component contains one gene (AV1) on the sense strand which encodes the coat protein (CP) and three genes on the complementary strand which encode the replication activator protein (C1), transcription activator protein (C2), replication enhancer protein (C3) and symptom determinant protein (C4). The DNA-B component consists of two genes which encode two proteins, the nuclear shuttle protein (BV1) and the movement protein (BC1). In monopartite begomoviruses, the DNA-A component is larger than the DNA-A component of bipartite begomovirus and it consist of six open reading frames (Idris et al. 2011; Leke et al. 2015; Mubin et al. 2010).

1.2.2 Factors influencing Begomovirus diversity

An ever-increasing number of geminivirus species are being described every year. With 288 accepted begomovirus species (ICTV, 2015, http://www.ictvonline.org/virusTaxonomy.asp), it is currently the largest genus represented in virus taxonomy (Brown et al. 2015). Though begomoviruses are a very successful group of emerging viruses for various reasons, as discussed below (Melgarejo et al. 2013), the size of the genus is not necessarily due to them being more abundant than any other virus group. It is more likely due to the number of researchers working on them and the emergence of new technology, able to characterize them enmass, such as next generation sequencing (NGS), in combination with rolling circle amplification (RCA) (discussed in section D below). In 2012, there were about 1348 full-length begomovirus sequences deposited in Genbank. By January 2014, more than 3500 full-length begomovirus sequences had been deposited in public databases, emphasizing the volume of work being done on this virus group (Brown et al. 2015).

Geminiviruses in general, but especially those belonging to the genus *Begomovirus*, are considered to be emerging and important plant viruses. This is due to the increasing incidence (geographical and / or host range) of new strains of existing viruses or of completely new virus species being described (Anderson et al. 2004; Brown et al. 2015). It is the adaptive potential of begomoviruses, due to their genome plasticity that makes these viruses capable of adapting to so many new niches. The mechanisms that viruses in general use to generate this diversity, as well as the driving forces involved in their evolution and emergence, includes a number of factors, all of which appear to be closely related to human activities and can be attributed to major agricultural changes, worldwide invasion by the insect vector species and evolutionary changes that take place at the (genome) molecular level.

Agricultural changes reported to play a pivotal role in promoting virus adaptation include factors such as agricultural intensification and diversification, loss of genetic diversity due to introduction of new crops and vulnerable cultivars, excessive use of chemical control measures, all-year-round cropping, increased use of protected cropping practices (Morales & Anderson, 2001; Anderson

et al. 2004; Morales & Jones, 2004; Fargette et al. 2006; Morales, 2006; Seal et al.,2006a,b) and, very importantly, the international trade in vegetative plant material. Perhaps the most important driver for begomovirus emergence is the worldwide increase of two invasive members of their vector species, *Bemisia tabaci*, namely the Middle East-Asia minor 1 type (also referred to as biotype B) and Mediterranean (MED) (also referred to as biotype Q) (Dinsdale *et al.,* 2010). Both of these vector species have a host range that includes more than 600 plant species, which has led to the expansion of the host range of geminiviruses (Brown et al. 2015; Fauquet and Stanley, 2003).

Geminivirus diversity, at the molecular level, is thought to be generated by mutation, pseudorecombination, interspecies recombination, re-assortment and de novo gene and satellite acquisition (Briddon and Stanley, 2006; Seal et al. 2006b; García-Andrés et al. 2007; Navas-Castillo et al. 2011; Lefeuvre and Moriones, 2015). Recombination is a major driving force in geminivirus evolution within species, genera and family levels (Lefeuvre; et al. 2007; Padidam et al. 1999). There are three factors which contribute significantly to recombination: mixed infections, high levels of viral replication and the broad host range of the whitefly vector. Mixed infections are common in geminivirus diseases (García-Andrés et al. 2007; Padidam et al. 1999; Seal et al. 2006a). Under these circumstances, it is easier for the viruses to exchange fragments of their genomes; the fragments can be small, consisting of a few nucleotides, to very large fragments of 2000 nucleotides (Fauquet and Stanley, 2003). Geminiviruses also have high levels of viral replication, as they replicate via a double stranded replicative form and thus they achieve very high copy numbers (Hanley-Bowdoin et al. 1999; Padidam et al. 1999). Recombination between new and old world begomoviruses was previously limited because of geographical separation, but due to modern agricultural practices and long distance transport of plant materials, such recombination events have been recorded to occur much more frequently (Rojas and Gilbertson, 2008; Padidam et al. 1999; Polston et al. 1999). Begomoviruses are also said to be highly 'plastic' because they are able to adapt very well to a large number of different cultivated plant species, which makes them a very large threat (Morales, 2001; Padidam et al. 1999).

1.3 Section B: Begomovirus Vector

1.3.1 Whitefly*: Bemisia tabaci* (Hemiptera: Aleyrodidae)

The whitefly, *Bemisia tabaci* (Gennadius) (Hemptera), is the only known insect vector of the *Begomovirus* genus. It is classified under the family *Aleyrodidae*, which is characterized by small whiteflies that feed on undersides of plant leaves. *Bemisia tabaci* is considered to constitute a cryptic species complex (Brown, 2005), whose members display genetic, biological and behavioral variation, but are morphologically indistinguishable (Abdullahi et al. 2003; Ashfaq et al. 2014; Esterhuizen et al. 2013). Some members of this species complex have invaded well

beyond their respective home ranges and cause devastating damage to crops. Damage is caused directly through feeding and contamination of plant products with honeydew, as well as indirectly, by acting as a vector for pathogenic plant viruses (Boykin et al. 2007; De Barro et al. 2011; Perring, 2001). *Bemisia tabaci* transmits over 150 plant viruses to different plant families including *Geminiviridae* (*Begomovirus*) and *Torradoviruses* (*Secoviradea*) (Jones, 2003) The plant viruses transmitted by these insects cause diseases to economically important crops and huge economic losses to agricultural production around the world (Ambrozevicius et al. 2002; Esterhuizen et al. 2012). Cassava, cotton, cowpea, cucurbits, crucifers, tobacco, tomato, potato, soybean, sweet potato, okra, lettuce, pea, bean, pepper, poinsettia, and chrysanthemum are some of the crops which are vulnerable to infestation (Jones, 2003). This, among other damages, endangers food security in developing countries (Bernado et al. 2013).

The importance of *B. tabaci* as a pest and as a vector of viruses of food, fiber and ornamental plants has increased over the last three decades. This is because of the emergence of two members of this cryptic species complex, referred to by Dinsdale et al. (2010) as Middle East - Asia minor 1 (MEAM1, also called *B. argentifolii* and biotype B) and Mediterranean (MED, also called biotype Q), and because of the rapid expansion in geographic distribution and host range of each. The nearly global invasion of the MEAM1 and MED species can largely be attributed to the trade in ornamental plant species, their high fitness and adaptability, broad host ranges and propensity to develop insecticide resistance (Bethke et al. 2009; Chu et al. 2006; Dennehy et al. 2005, 2010; Delatte et al. 2009; Drayton et al. 2009; Frohlich et al. 1999; Hadjistylli et al. 2015; Horowitz et al. 2005; Mckenzie et al. 2009; Prabhaker et al. 2005). In many regions of the world, epidemics of plant diseases caused by begomoviruses transmitted by *B*. *tabaci* occurred soon after the invasion of the MEAM1 and MED types (Hogenhout et al. 2008; ; Seal et al. 2006a; Varma and Malathi, 2003).

Some *B. tabaci* species have a great capacity to transmit begomoviruses *(Hadjistylli et al. 2015; Hogenhout et al. 2008; Jones, 2003; Navas-Castillo et al.* 2011). The whitefly instar, nymphs, as well as the adults feed from plants by inserting their proboscises into the leaf, penetrating the phloem and withdrawing sap. It is during this feeding process that they acquire viruses (Jones, 2003). Begomoviruses are transmitted from infected plants to healthy plants in a persistent circular manner. Once the vector acquires the virus, transmission can occur within hours and can last for the entire lifespan of the whitefly (Lapidot and Friedmann, 2002). After the virus is taken up by the whitefly it crosses the midgut epithelium, is transported through the haemocoel and then enters the salivary glands. The adult whiteflies are then able to disperse the virus to other plants by feeding. Furthermore, since whiteflies are phloem feeders, the begomovirus is deposited directly into phloem cells. For efficient transmission of the virus, the capsid must protect the viral DNA in the insect gut and haemocoel, which involves the virus binding to receptors in the midgut and salivary epithelium (Jeske, 2009; Morales, 2001).

In addition to the invasive-like members of the Mediterranean and Middle East-Asia minor clades, the sub-Saharan African region harbours indigenous and possibly less invasive *B. tabaci* types that are known to vector many different begomovirus species (De La Rua et al. 2006; Legg et al. 2002). These indigenous populations cluster in the major sub-Saharan Africa non-silverleafing clade (SSAF), into five subclades (SSAF1-5) (Boykin et al. 2007, Dinsdale et al. 2010; Esterhuizen et al. 2013). *Bemisia tabaci* members within the SSAF major clade have been documented colonizing vegetable crops of which the majority associate with cassava (*Manihot esculenta*) (Abdullahi et al. 2003; Berry et al. 2004; Brown, 2010; Brown and Idris 2005; Carabali et al. 2010; De La Rua et al. 2006; Legg et al. 2002; Maruthi 2004; Sseruwagi et al. 2005; Sseruwagi et al. 2006). The cassava associated types transmit at least 9 species of begomoviruses to cassava that is a major staple food in sub-Saharan Africa (Abdullahi *et al.,* 2003; Carabali et al. 2010; Legg, 1996; Legg et al. 2002).

The presence of the *B. tabaci* in South Africa has been recorded since the 1960s, when it was associated with Tobacco leaf curl disease in tobacco-producing areas in the Gauteng and North West provinces (Hill, 1967; Thatcher, 1976). However, the specific *B. tabaci* type or species were only recently identified in several studies investigating the host range and distribution of these insects in South Africa. These studies revealed the presence of members from two endemic sub-Saharan Africa (SSAF) subclades co-existing with two introduced putative species: the SSAF-1 subclade, that includes cassava host-adapted *B. tabaci* populations, the SSAF-5 subclade that was found on both cassava and non-cassava hosts and lastly, the Mediterranean and Middle East-Asia minor 1 clades that include the B and Q types, respectively (Bedford et al. 1994; Berry et al. 2004; Esterhuizen et al. 2013).

1.4 Section C – Viral hosts

1.4.1 Cultivated Crop hosts

Geminiviruses are responsible for various economically significant crop diseases throughout the tropical and subtropical regions of the world, but are also a particularly serious problem in Africa, where they pose a serious risk to food security (Bernado et al. 2013; Rey et al. 2012). It threatens production of the continent's two main food crops namely, maize and cassava, as well as several vegetable crops, including cucurbits, okra, watermelon, beans, cowpea, pepper and tomatoes to name a few (Fauquet et al. 2008).

To date, several dicot-infecting geminivirus species have been reported in SA: two mastreviruses infecting bean crops (Liu et al. 1997) and eight begomoviruses infecting cassava, tomato and sweet potato crops (Table 2; Berrie et al. 1997; Esterhuizen, 2012; Esterhuizen et al. 2012b; Pietersen et al. 2000, 2008). Although the bean infecting mastreviruses (BeYDV) has been

repeatedly detected over the last ten years (Esterhuizen, 2012; Liu et al. 1997), little is known about the etiology of this disease in the country. Similarly, the two sweet potato infecting geminiviruses have been found to be widespread in South Africa, but information on its distribution and economic impact still remains to be determined, especially in light of the fact that sweet potato is an important food security crop in the SA, particularly in rural areas (Esterhuizen et al. 2012b). In 2013, an isolate of the recently formed Capulavirus genus, *Euphorbia caput-medusae latent virus*, was found infecting wild spurge, *Euphorbia caput-medusa* in South Africa. Although the virus does not induce symptoms in its natural host, it was shown to cause severe symptoms in both tomato and *N. benthamiana* plants, suggesting a potential threat to agricultural crops in future (Bernardo et al. 2013; Versani et al. 2017).

Table 1.2. List of the geminivirus species reported up to date in South Africa, indicating the region, host plant where the virus was detected and Genbank accession numbers (Bernado et al. 2016; Fauquet et al. 2008; Varsani et al. 2017).

Genus	Species	Region	Host	Accession no.
Begomovirus	*Tomato curly stunt virus* (ToCSV)	Onderburg, Mpumalanga	*Solanum lycopersicum*	AF261885
	Tomato curly stunt Mooketsi virus	Mooketsi, Limpopo	*Solanum lycopersicum*	
	Tomato curly stunt Lanseria virus	Lanseria, North West	*Solanum lycopersicum; Malva parviflora; Datura stramonium*	
	Tomato curly stunt Noordoewer virus	Noordoewer, Northern Cape	*Solanum lycopersicum*	
	South African cassava mosaic virus (SACMV)		*Manihot esculenta*	AF15506 AF15507
Mastrevirus	*Bean yellow dwarf virus-mild* (BeYDVm)	Pretoria, Gauteng	*Phaseolus valgaris*	DQ45879
	Bean yellow dwarf virus (BeYDV)	Mpumalanga	*Phasealus valgaris*	Y11023
	Maize streak virus A (MSV-A)	Komatipoort, Mpumalanga Vaalhart Makatini	*Zea mays*	AF003952 AF329884 Y00514
	Maize streak virus B (MSV-B)	Vaalhart	*Triticum spp*	AF239962
	Maize streak virus C (MSV-C)	Seteria		AF007881
	Maize streak virus D (MSV-D)	Rowsonville		AF329889
	Maize streak virus E (MSV-E)			AF329888
	Sugarcane streak virus (SSV)	Kwazulu-Natal	*Saccharum spp*	M82918 S64567
Eragrovirus	*Eragrostis curvular*		*Eragrostis*	FJ665633

	streak virus			
Capulavirus	*Euphorbia caput-medusae latent virus* (EcmLV)		*Euphorbia caput-medusae*	HF921459

Cassava, (*Manihot esculenta*), is a tuberous crop that is widely grown in tropical Africa, Asia and South America. This crop serves as a food source to almost one billion people worldwide (Nassar et al. 2002). This crop ranks as the second most important staple food crop after maize in sub-Saharan Africa (Nweke, 2001), but in SA it is mostly grown as a secondary food choice and/or for sale in the local markets (Berrie *et al*., 2001)**.** Cassava Mosaic Disease (CMD), caused by nine different cassava mosaic virus species, is one of the main biotic and economically important constraints of cassava cultivation in southern Africa. SACMV-[Za:99] in particular is endemic to South Africa. CMD affects the foliar parts of the plant and reductions in yield have been shown to range between 25 - 95% (Bisimwa et al. 2015).

Tomato (*Solanum lycopersicum*) is an economically important vegetable crop commonly grown by subsistence and resource poor farmers in South Africa. It is one of the main vegetables used for hawking by small-scale entrepreneurs in the informal sector. The crop is also grown commercially and provides a large number of employment opportunities in this country. The total annual production of tomatoes is about 566 180 tons (Abstract of Agricultural Statistics, 2013 NDA, 2009; Tshiala and Olwoch, 2010). Tomato-infecting begomoviruses have, however, become one of the major constraints to open field tomato production in SA. The presence of ToCSV was first reported in 1997 (Pietersen et al. 2000, 2008). In recent years, a sharp increase in the incidence of this begomovirus disease has been observed in most of the tomato production regions throughout the country, with severe yield and quality losses. Subsequent surveys have revealed that whereas ToCSV is the predominant tomato infecting begomovirus species in SA, there also exist at least three additional distinct monopartite begomovirus species (Table 1.2 Esterhuizen et al. 2010; Esterhuizen, 2012). While these other species appear to have a more limited geographical range than ToCSV, they nevertheless cause symptoms similar to ToCSV in tomato plants (Esterhuizen et al. 2012; Rossouw, 2016). Infected tomato plants show characteristic upward cupping, distortion and interveinal yellowing of leaves, severe stunting, degeneration, and flower abortion. These symptoms often lead to substantial yield losses, especially if plants are infected at an early growth stage. Although no formal crop loss assessment studies have been conducted since the first identification of the virus, yield losses of up to 95% have been recorded in many production seasons (Pietersen et al. 2008).

Whitefly transmitted geminiviruses cause devastating losses to agricultural production around the world and they are also a problem here in South Africa. Due to the wide host range of the vector,

B. tabaci, and the genetic plasticity of begomoviruses, these diseases pose a threat to many other crops as well. Many weed species, or non-cultivated indigenous plants that are growing alongside cultivated areas have been reported to harbor begomovirus infection. Since these plant species are there in the field throughout the year, these viruses is repeatedly transmitted from the weed species or indigenous plants to the crops, across the agro-ecological interface'. (Ambrozevicius et al. 2002; Azhar et al. 2011; Burdon and Thrall, 2008; Mubin et al. 2010).

1.4.2 Weed hosts

Recent advances that have been made in plant viral metagenomics studies (Discussed below in section D) have indicated that the number of viruses circulating in nature has vastly been underestimated. In addition, our knowledge of viral diversity in our environment have largely been limited towards viruses associated with diseased plants and very little is known about viral species that are living in mutualistic relationships with plants (Stobbe and Roossinck, 2014). Numerous metagenomic surveys exploring plant virus biodiversity have unearthed a plethora of new viral species and even new viral families (Alavandi and Poornima, 2012) in wild plants, insects, and other environments (Ng et al. 2011; Roossinck et al. 2010; Roossinck, 2012; Wren et al. 2006). This may be due to the fact that most viral infections in wild plants are not linked to disease or obvious symptoms and have therefore been overlooked. For example, despite only 2.3% of plant specimens collected in the Tallgrass Prairie Preserve in northeastern Oklahoma having any outward signs of virus infection, viruses were discovered in 25% of these (Muthukumar et al. 2009). While it is true that most viruses discovered in wild plants have not been linked to any pathology in their natural host, many of these viruses can, or have been shown to be pathogenic when infecting domesticated plants (Ambrozevicius et al. 2002; Castillo-Urquiza et al. 2008). For plant pathologists, the identification of these potential pathogens is very important for disease management, should they jump species into domesticated plants.

In addition to the crop species that begomoviruses infect, weed species (wild/non-cultivated plants) growing alongside cultivated crops have drawn considerable attention due to their widespread and perennial nature and the diversity of viruses they harbor. Numerous reports in literature of begomovirus infection in weeds from different botanical families, support the notion that these weed species act as natural hosts or begomovirus reservoirs, keeping the virus in the agroecosystem during the non-crop seasons and play an important role in the spread of these viruses into cultivated crops (Alabi et al. 2016; Fiallo-Olivé et al. 2013; García-Andrés et al.,2006; Graham et al. 2010; Jovel et al. 2004; Khashina et al. 2002; Prajapat et al. 2014; Tahir et al. 2010; Wyant et al. 2011). Additionally, the wide distribution and host range of the whitefly vector, *B. tabaci*, within agricultural areas further play an essential role in the distribution of these viruses (Hadjastylli et al. 2015). For this reason, the removal of volunteer plants and weeds from the border of a particular crop field or the use of bait plants within a crop field are two well-known

strategies used in the management of viral diseases. This will reduce the vector infestation and subsequent viral transmission. As seen in a number of countries around the world, the remarkable emergence of begomoviruses has been driven by the introduction of the more polyphagous biotypes of *B. tabaci* (primarily B, but also the Q type) that has spread throughout the world. Together, these are two important factors influencing the epidemiology of begomoviruses in an area. The wide host range of begomoviruses and their whitefly vector, together with the distribution of susceptible host plant species, including wild plants, improves the chances for begomoviruses to be present and in many cases, leads to mixed infections that can lead to recombination, diversity, and subsequent evolution of these viruses. (Ambrozevicius et al. 2002; Azhar et al. 2011; Bedford et al. 1998; García-Andrés, et al. 2006; Varsani et al. 2008). Infection by these essentially unknown or poorly studied plant viruses, can further complicated disease management.

The viral diversity and host range of tomato-infecting begomoviruses from Brazil has been well studied and can serve as a good example where viral diversity, found within wild/non-cultivated hosts, and the introduction of a polyphagous virus vector, culminated in the sudden emergence of tomato crop diseases (Rocha et al. 2013; Inoue-Nagata et al. 2016). The first reported tomato begomovirus disease in Brazil was caused by *Tomato golden mosaic virus* (TGMV) in the 1960's and this virus was thought to be transmitted by the indigenous *B. tabaci* biotype A. TGMV was never a serious economic problem, presumably because tomato was not a preferred host of the A biotype present in Brazil at that time. Today, a total of 14 different indigenous begomovirus species are causing tomato diseases in Brazil (Albuquerque et al. 2012b; Barbosa et al. 2011; Castillo-Urquiza et al. 2008; Fernandes et al. 2006; Ribeiro et al. 2003; Rocha et al. 2013). The situation is thought to have changed in the early 1990's, due to the introduction of the highly polyphagous biotype B that feeds on tomato and many other crops and non-cultivated plants. As in many other parts of the world, the invasive *B. tabaci* biotype B is thought to have driven the emergence of a diversity of new-tomato infecting begomovirus species via local evolution in Brazil. All these virus species are unique to Brazil and have emerged from the large and genetically diverse reservoir of indigenous begomoviruses infecting non-cultivated plants (Rocha et al. 2013). This hypothesis is supported by the fact that some of these tomato-infecting viruses (ToRMV, ToCMoV, and ToYSV) have been found in ubiquitous, non-cultivated species such as *Nicandra physaloides*, *Solanum nigrum*, and *Datura stramonium* and that begomoviruses originally found in common non-cultivated species, such as *Sida mottle virus* (SiMoV) and *Sida micrantha mosaic virus* (SimMV), have also now also been found infecting tomato crops in the field (Inoue-Nagata et al. 2016)

There have been numerous other studies regarding weeds hosting begomoviruses and whiteflies in different ecological areas around the world (Alabi et al. 2016; Bedford et al. 1998; George et al. 2014; Kashina et al. 2002; Mubin et al. 2012; Rajeshwari et al. 2004; Tahir et al. 2015).

However, the specific weed species, their abundance and growth stages varies in different environments and thus whitefly populations, and frequency of viral presence in weed hosts, may also differ. Furthermore, many weed plants found to be infected by begomoviruses, exhibit no disease symptoms and are therefore difficult to identify. To date, weeds hosting begomoviruses in South Africa and the relationship between viruses and whiteflies feeding on infected weeds have not been thoroughly studied. Only two previous studies have reported on the host range of a South African begomovirus, namely ToCSV, in *S. lycopersicum, D. stramonium*, several *Nicotiana* species and *Phaseolus vulgaris* cultivars and a number of weeds growing around tomato fields, including *Cleome spp., D. stramonium, D. ferox, C. carinatum, Amaranthus hybridus*, *Alternanthera pungens, S. cordifolia, S. rhombifolia* (Pietersen et al. 2008; Esterhuizen, 2012). These findings are of great significance, as it shows that a weed-infecting begomovirus in South Africa do indeed infect crops and visa-versa. Knowledge on which weeds host begomoviruses and are fed upon by *B. tabac*i may provide useful information for integrated management of begomovirus infection and whitefly infestation in crop production regions in South Africa.

1.4.3 Plant identification methods

The identification of an unknown plant specimen involves determining which already known plant species that plant is similar or identical to and then assigning a plant specimen to a particular taxonomic group, and ultimately to a species. The skill to identify an unknown plant is a most valuable asset and an important part of the study of systematics. People with this level of expertise and understanding of plant morphology are normally found in botanical gardens, herbaria, museums, colleges and universities. The process of identification normally involves (1) expert determination, (2) recognition, (3) comparison, and (4) the use of keys and similar devices (Sneath and Sokal; 1973). A key is a device in which a series of choices between contrasting statements are followed, until the correct name is found by the process of elimination. Keys play an integral part in a flora, allowing for proper identification of families, species, and infraspesific taxa (Jin et al. 2015). Keying specimens however takes skill and considerable practice and, although of great reliability, this method presents problems of requiring the valuable time of experts and creating delays for identification. For the layperson, when trying to identify an unknown plant from a particular area, it is normally common practice to refer to books, the latest floras and check list accounting for the plants of that region. Although this is also a reliable method, it may be very time consuming or virtually impossible due the quality of the plant manual or database used, the quality of the specimens available and the observation skill of the user. Another important factor that may hamper proper identification is the lack of suitable materials for comparison, as is the case when trying to identify weeds during their earlier growth stages in a particular agricultural setting. In all the identification resources, plant families are delineated to a large degree by their flower structures, because they display the greatest number of defining traits and these are the characteristics normally represented in the identification keys, making identification of plants at all their life stages particularly problematic.

A novel identification methodology, called DNA barcoding, has the potential to alleviate the above mentioned challenges posed to plant identification. DNA barcoding is a method for identifying living organisms to species level. This approach was first used in animals, where a portion of the mitochondrial cytochrome oxidase 1 (mtCO1) gene was used as a barcode for animal species identification (Herbet et al. 2003). This method makes use of a short (<1000 bp), standardised region of the genome (a “barcode”) that evolves fast enough to differ between closely related species. When a DNA barcode sequence is retrieved, an algorithm is used to compare it to a public community resource of reference DNA sequences or barcodes from identified museum/herbarium voucher specimens, thus enabling the unknown specimen to be identified (Hollingsworth et al. 2011). For plant identification, this approach has several benefits over the existing system of morphological identification and is applicable to all plant life stages (identifying

plant leaves even when flowers or fruit are not available) (CBOL Plant Working Group, 2009; Hollingsworth et al. 2011).

In recent years DNA barcoding has grown substantially, and worldwide efforts coordinated by the Consortium for the Barcode of Life (CBOL) are now being put into retrieving barcode sequences from all organisms (CBOL, 2009). Although several barcoding regions have been developed and are continually being evaluated in plants over the past decade, worldwide effort on DNA Barcoding has led to a degree of standardization in the process of barcoding of plants, and two regions, *matK* and *rbcLa*, are currently being used as barcodes for land plants (CBOL, 2009). The use of *matK* and *rbcLa* as a plant barcode was based on the ease of recovery of the *rbcLa* region and the discriminatory power of the *matK* region. The virtually universal presence and key function in photosynthesis (it codes for the enzyme ribulose bisphosphate carboxylase (RuBisCo), as well as its ease of amplification has made the *rbcL* gene a useful backbone to the barcode dataset, despite it having only modest discriminatory power. One section of its DNA sequence is very variable between species, making it ideal for DNA barcoding. In contrast, *matK* is one of the most rapidly evolving coding sections of the plastid genome (Hilu and Liang, 1997; Hollingsworth et al. 2011; Wicke and Quandt, 2009), and is perhaps the closest plant analogue to the CO1 animal barcode. Unfortunately, *matK* can be difficult to amplify by PCR using existing primer sets, mainly in non-angiosperms. To date, the two-marker plastid barcodes give better discrimination than single marker barcodes, but no other 2-marker or multi-marker plastid barcode have given appreciably greater species resolution than the *rbcL+matK* combination (CBOL, 2009). The use of DNA barcoding for plant identification has found wide applicability amongst professions (i.e. taxonomists, ecologists, conservationists, foresters, agriculturalists, forensic scientists, customs and quarantine officers) (Candresse et al.,2014; Costion et al. 2011). Numerous studies in the last decade have reported the used DNA barcoding as plant identification tool for a range of applications, for example in ecological forensics (Kress et al. 2005; Staats et al. 2016; Yan et al. 2013); for the identification of processed plant material (Fields *et al*., 2015; Galimberti et al. 2013; Mankga et al. 2013) or the identification of protected species in trade (Goncalves et al. 2015; Hartvig et al. 2015), to name a few.

1.5 Section D - Viral metagenomics studies

1.5.1 Traditional virus identification methods

Viruses form a major class of biological entities encompassing diverse environments ranging from algae in marine ecosystems to soil, plant, human and animal systems. When examining the extend of viral diversity on earth, the 2014 database release from the ICTV classified a total of 3,286 known viral species within 7 different orders, 104 families and 505 genera

(http://www.ictvonline.org/virustaxonomy.asp). This huge extent of viral diversity greatly complicates viral identification. In viral diagnostics, the ability to provide a fast, inexpensive and reliable diagnostic method for any given viral pathogen is critical in efforts to fight and control the diseases caused by these ubiquitous pathogens. The detection and identification of known and new viruses have traditionally relied on routine methods, including electron microscopy and biological assays (indicator plants) as a broad screen for plant virus detection. The field of viral diagnostics have however been revolutionized by the development of firstly serology based techniques, such as enzyme-linked immunosorbent assay (ELISA), which used monoclonal or polyclonal antibodies for detection of viruses, and later by nucleic acid based techniques like polymerase chain reaction (PCR) and its variants (Boonham et al. 2014; Kreuze et al. 2009; Prabha et al. 2013).

Both ELISA and PCR based detection methods have improved our ability to detect viral infections over the last 40 years, in terms of sensitivity, specificity and reproducibility (López et al. 2009; Prabha et al. 2013) and both methods are applicable to broad based virus detection. However, the application of these techniques suffer from the drawback that they are largely restricted to known and well characterized viral agents for which serological reagents and/or sequence information are available. The diagnosis of unknown or poorly characterized viral agents still pose very complex challenges that are only partially met by the use of polyvalent serological or molecular assays using degenerate primers. As more and more viruses sequence information become available, these degenerate primers suffer from what has been called 'consensus decay' (Prabha et al. 2013; Zheng et al. 2008). Virus detection is further complicated by mixed infections or virus complexes that play a role in the etiology of particular viral diseases (Prosser et al. 2007). Conventional techniques have limited capability to identify all the virus variants in these plants.

All the above mentioned routinely used diagnostic techniques are very specific and therefore require prior knowledge of the infective agent(s), either its sequence information or detecting antibodies. Given these drawbacks, various deep sequencing methods, based on next generation sequencing (NGS) technologies (Section 1.5.3) have been used for improved detection and characterization of new or unusual viruses affecting diverse plants (Boonham et al. 2014; Dinsdale et al. 2008; Kehoe et al. 2014; Kreuze et al. 2014; Marozova et al. 2008; Muthukumar et al. 2009; Ng et al. 2009; Prabha et al. 2014; Roossinck et al. 2015; Shendure and Ji, 2008; Shokralla et al. 2012). What makes high-throughput, NGS technology particularly suited to the field of plant virus diagnostics is the fact that it allows direct detection, identification and discovery of all viruses in diverse plant species, in an unbiased manner without antibodies or prior knowledge of the virus sequences, as will be discussed in the following section (Bexfield et al. 2011; Roossinck, 2011, Roossinck et al. 2015).

1.5.2 Viral metagenomics approach

While traditional virus detection methods suffer several drawbacks, especially when trying to identify “unknown” or “new” viruses (Bexfield et al. 2011; Roossinck, 2011, Roossinck et al. 2015), viral metagenomics approaches have recently developed into a powerful technique for new and unbiased viral discovery. Metagenomics is the analysis of microbial communities in environmental samples through sequencing, particularly NGS techniques. Viral metagenomics is the study of virus communities (Edwards and Rohwer, 2005), whereby viral particles are purified and sequenced using shotgun sequencing methods (Rasario and Breitbart, 2011). This approach has been successful in identifying novel viruses that infect hosts from all the major taxonomic groups of life: e.g. vertebrates (Phan et al. 2011; Siebrasse et al. 2012), arthropods (Tokarz et al. 2014), plants (Geering et al. 2014), fungi (Marzano and Domier, 2015), bacteria (Deng et al. 2014; Holmfeldt et al. 2013) and archaea (Mochizuki et al. 2012). These viral metagenomic studies have shown that we have vastly underestimated the immense diversity of viruses that occur in our environment. In addition to identifying numerous novel viruses in a wide range of environmental samples, the majority of sequences found in these studies have no detectable similarity to sequences in public databases like GenBank and therefore remain unassigned (Stobbe and Roossinck, 2014; Roossinck 2012b).

The general methodology used in metagenomics studies, include the following steps: isolation of nucleic acids (DNA and/or RNA), random fragmentation by restriction digestion or shearing, ligation to common adapters; size fractionation using agarose electrophoresis, library construction, (paired-end) sequencing, and post-sequencing analysis that involves bioinformatics analysis to reassemble the original nucleic acid sequences and identification of the donor organism(s). One disadvantage of this shotgun sequencing approach is that it cannot distinguish between target (for example viral), host (animal or plant genomic, plastid or mitochondrial nucleic acids) and contaminating nucleic acids, complicating bioinformatics analysis of the data and lowering the number of target sequences obtained. This is particularly a problem in viral metagenomics studies. Since viruses contain both RNA and DNA genomes, and do not have universally conserved genes (as is the case with 16S rRNA gene for bacteria and internal transcribed spacer (ITS) in fungi used for prokaryotic or eukaryotic microbial metagenomics studies), the concept of viral metagenomics was not accepted by many virologists at first (Allender et al. 2001; Marz et al. 2014; Roossinck, 2011, Roossinck et al. 2015). Recent viral metagenomics studies have however come up with methods that are not sequence-specific and that enhance or enrich viral sequences in a sample, which is then sequenced directly or after an amplification step (Kreuze et al. 2009; Roossinck, 2011, Krishnamurthy et al. 2017). Four main classes of nucleic-acids have been targeted by such analyses: (1) total RNA or DNA (Hall *et al*. 2014; Wylie et al. 2015); (2) virion-associated nucleic acids (VANA) purified from virus-like particles (Boukari et al.

2017; Palanga et al. 2016; Roossinck et al. 2015); (3) double-stranded RNAs (dsRNA) (Marston et al. 2013); and (4) virus- derived small interfering RNAs (siRNAs) (Donaire et al. 2009; Kreuze et al. 2009). All of these methods have strengths and weaknesses (reviewed in a number of recent papers) (Howe and Chain, 2015) depending on specific objective or target of the study in case (Halary et al. 2016).

When targeting viruses within the Geminiviridae family, the use of rolling circle amplification (RCA) have provided a powerful new tools for the identifying of known or novel viral species. Since geminiviruses and their associated satellite molecules have circular ssDNA genomes, they are ideal templates for RCA, which are amplified or enriched in a sequence independent manner resulting in a large proportion of the NGS reads and eventually assembled contigs to be virus-derived. (Haible et al. 2006; Idris et al. 2014; Ng et al. 2009). RCA in combination with a high throughput next generation sequencing technology is a technique referred to as "circomics", which has very successfully been applied for both rapid viral diagnosis and identification (Wyant et al. 2011) and viral metagenomics based identification of geminiviruses (or other circular viral genomes) in diverse environmental samples (Prabha et al. 2013; Roossinck, 2011; Roossinck et al. 2015).

Another important aspect of the viral metagenomics approach that needs to be mentioned, is the post-sequencing bioinformatics analysis steps. The output from NGS technologies are gigabases of raw sequence data that require extensive computational analysis using various algorithms and applications to infer biological significance. The large amounts of data obtained from large-scale NGS experiments has triggered the development of numerous bioinformatics tools and method for efficient analysis, interpretation, and visualization of NGS data (Miller et al. 2010; Magi et al. 2010; Rose et al. 2016; Wang et al. 2015). There are many bioinformatics tools that have been developed to specifically support virus detection from NGS data at various points in the analysis pipeline. The general bioinformatics pipeline include the following steps (summarized in Figure 1.5.1 below): (1) quality analysis and removal of adaptors and low-quality sequences (using tools such as FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/); (2) subtraction of the host genomic components through alignment with one or more reference genomes using an aligner or specialist tool (approaches reviewed in Daly et al. 2015) (only possible for hosts with sequenced genomes); (3) reconstruction of sequencing (raw) reads into full length genes and genomes (called contigs) by means of either reference-based mapping or *de novo* assembly (Thomas et al. 2012). Consequently, viral identification is achieved typically by similarity searches on the nucleotide level against a reference database using an alignment algorithm, such as the Basic Local Alignment Search Tool (BLASTn). Although Blastn can detect homology in divergent sequences, this approach is however limited by the relatively few validated viral sequences deposited in public databases, the high diversity within viral families which can obscure relatedness and the huge number of unknown and diverse viruses that have no sequence

similarity currently with the reference sequences in these databases. It is therefore normally followed by less stringency protein alignments using a translated amino acid alignment algorithm (Blastx) for identification of novel viruses with divergent genomes.

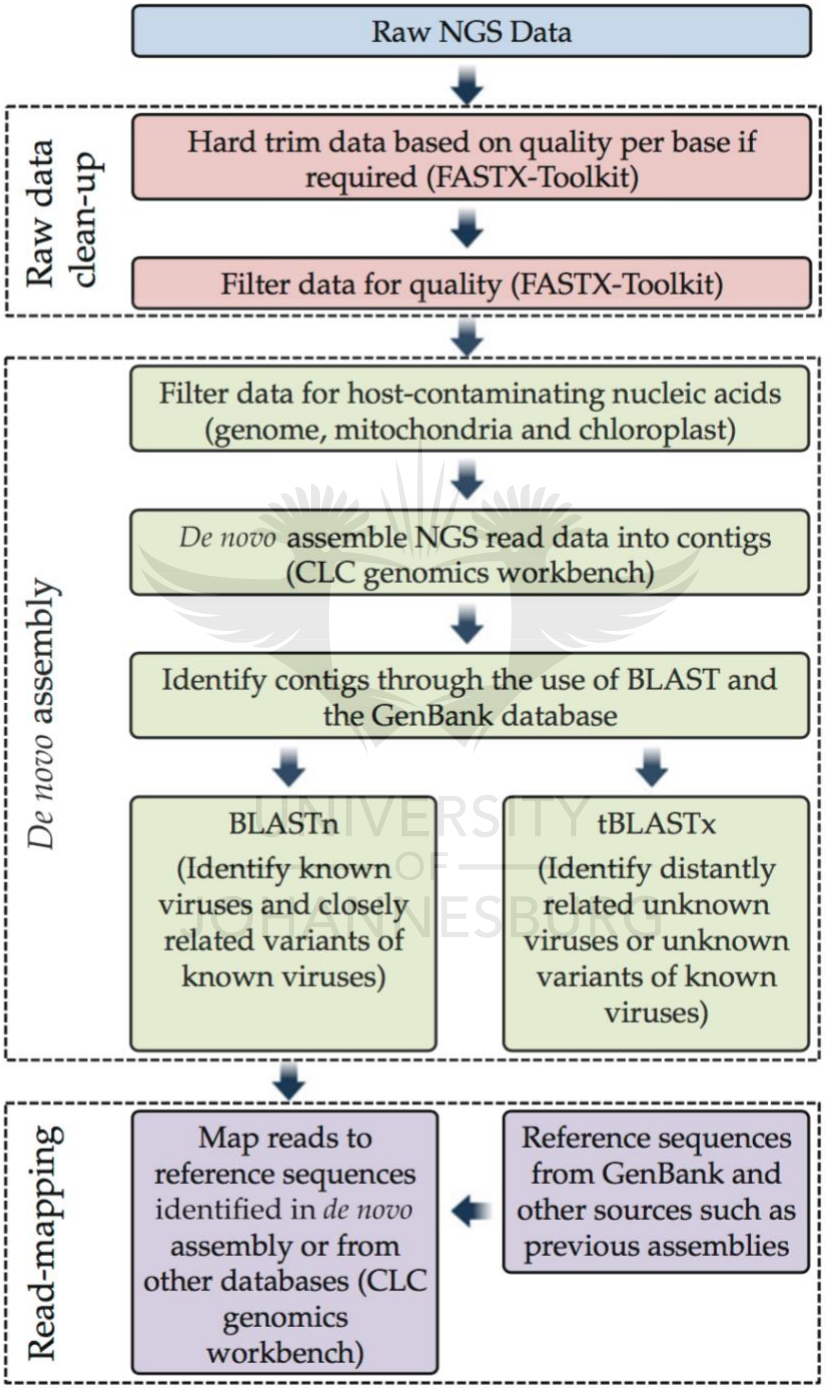


Figure 1.2: A summary of the general bioinformatics workflow analysis of NGS data in a viral metagenomics study (adopted from Burger and Maree, 2015).

1.5.3 Next-generation sequencing technologies

Since the fundamental discovery of the structure of DNA (Waston and Crick, 1953), methods were developed to determine the precise order of nucleotides within a DNA molecule as a routine assay, i.e. DNA sequencing. The first foundational approaches were the Maxam and Gilbert's technique (Maxam and Gilbert's, 1977) and more widely used, Sanger sequencing (Sanger et al. 1977). In Sanger sequencing, the target DNA is amplified by PCR multiple times, producing fragments of different lengths that incorporate a fluorescent "chain terminator" nucleotide that mark the ends of the fragments and allow the sequence to be determined upon separation by capillary electrophoresis. In Sanger sequencing, regions of DNA up to about 900 base pairs in length are routinely sequenced and this technique is still considered the 'gold standard' in terms of both read length and sequencing accuracy. Beginning in 2005, the traditional Sanger-based approach to DNA sequencing has however experienced revolutionary changes with the development of several HTS or NGS platforms that provide an unbiased means to examine billions of templates of DNA and RNA. They are able to generate three to four orders of magnitude more sequence and are considerably less expensive than the original Sanger method.

These new HTS or NGS platforms are based on various implementations of cyclic-array sequencing that involve a dense array of DNA features that are sequenced by iterative cycles of enzymatic manipulation and imaging-based data collection. Table 1.3 provides a summary of commercially launched sequencing platforms up to date. The most widely used being Illumina's Genome Analyzer (http://www. illumina.com), ABI's Solid (http://www.454.com/) and Roche/454. These platforms are quite diverse in sequencing biochemistry as well as in how the array is generated but their work flows are conceptually very similar (Adams et al. 2009; Kreuze et al. 2014, Kehoe et al. 2014; Marozova et al. 2008; Shokralla et al. 2012). In summary, they are characterized by differences in the following three criteria of operation: (1) single-molecule detection per reaction, well, or sensor (Pacific Biosciences and Oxford Nanopore platforms, http://www.pacificbio sciences.com/) or detection of clonally amplified DNA (Illumina, Ion Torrent (http://www.iontorrent.com/), and Roche/454 platforms); (2) mode of detection of nucleotides to make sequencing base calls, i.e. detection of fluorescently modified nucleotides (Illumina, Pacific Biosciences), detection of light via pyrosequencing (Roche 454) or nonoptical detection, by the detection of pH changes during PCR via a solid-state nanopore sensor (Ion Torrent) or the measurement of translocation of DNA through a nanopore sensor (Oxford Nanopore), and lastly (3) the use of a polymerase (Illumina, Ion Torrent, Pacific Biosciences, and Roche 454 platforms) or ligation process (Applied Biosystems SOLiD platform) to drive a sequencing-by-synthesis reaction in which the products of the reaction are measured to produce sequencing data, or direct

measurement of DNA molecules (Oxford Nanopore platform) (Shokralla et al. 2012; Mardis, 2013; Levy and Myers, 2016). These variations in operating procedures amongst the different NGS platforms result in a spectrum of capabilities and specifications that lead to different strengths and weaknesses among the platforms. The two most common specifications used to compare these sequencing platforms are the number of reads produced in a given instrument run and the length of those reads and this provides a useful picture of comparative output. Figure 1.3 shows a graph of the commercially available instruments in terms of read length and depth (http://flxlexblog.wordpress.com).

Over the past decade, the above mentioned high throughput NGS techniques has proven to be powerful tools in viral metagenomics based approaches for virus identification, diagnosis of disease associated viruses and the discovery of novel viruses in an unbiased manner, without requiring antibodies or prior knowledge of the viral genomic sequence (Zheng et al. 2017). Not only do these methods dramatically reduce the time and cost of sequencing for a fraction of the price of sanger sequencing, it also allows parallel sequencing that increases the data output and there is no requirement for initial cloning steps that results in time and cost savings. The replacement of sequence-specific primers (used in PCR variations) with universal adaptors makes NGS suitable for metagenomics projects and has the ability to reveal all the genetic material in a sample an unbiased manner.

Table 1.3 Summary of second-generation sequencing platforms Modified from Levy and Meyers, 2016; Shokralla et al. 2012.

Platform	**Amplification**	**Detection**	**Chemistry**	**Company**
Illumina	Clonal	Optical	Sequence by synthesis	Illumina
Oxford Nanopore	Single molecule	Nanopore	Nanopore	Oxford
Pacific Bioscience System	Single molecule	Optical	Sequence by synthesis	Pacific Bioscience
Ion Torrent	Clonal	Solid state	Sequence by synthesis	Thermofisher
454 FLX System	Clonal	Optical	Sequence by synthesis	Roche Applied Science
Helicos Heliscope	Single molecule	Optical	Sequence by synthesis	Heliscose Bioscience

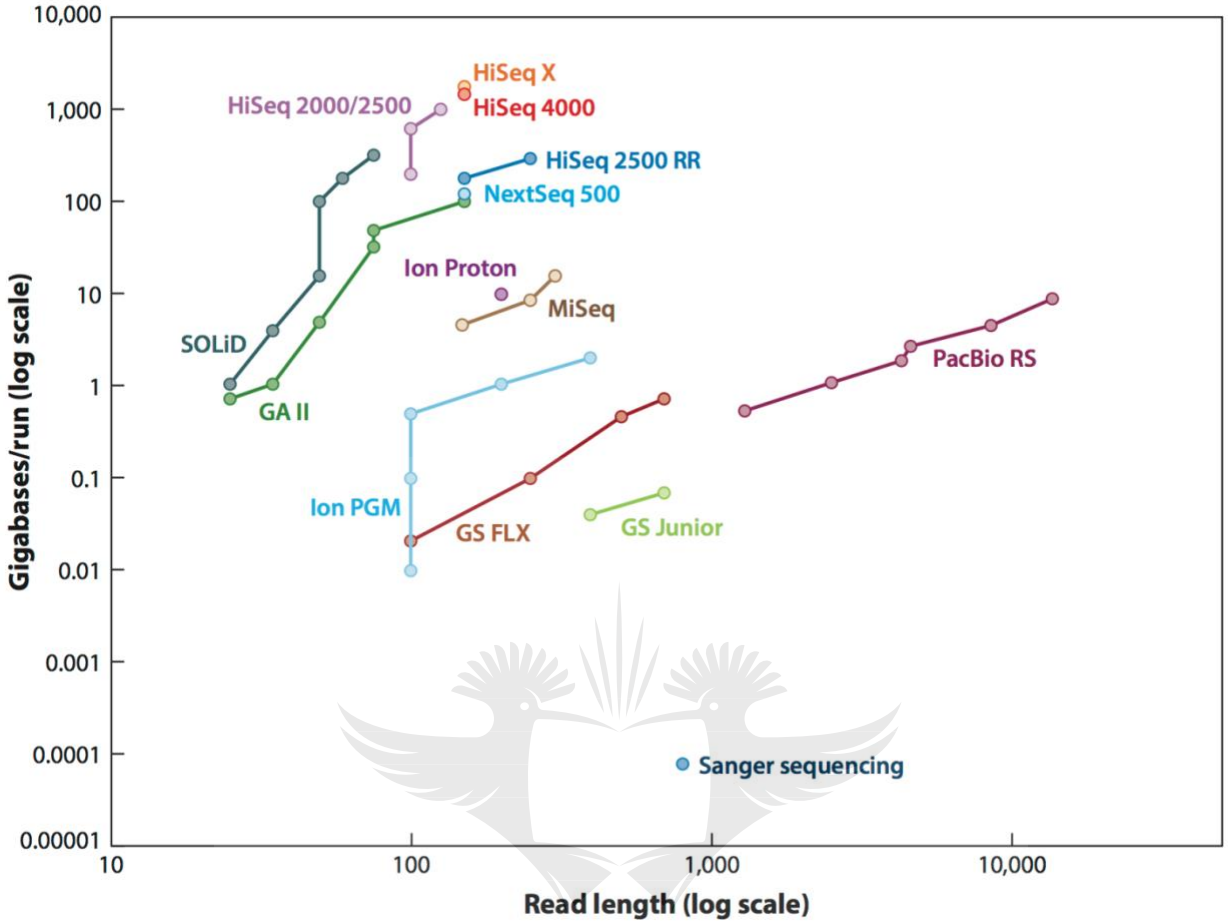


Figure 1.3: Comparison of read length and output of NGS platforms. SOLiD is an Applied Biosystems platform; Ion PGM and Ion Proton are Ion Torrent platforms; GA II, HiSeq, NextSeq, and MiSeq are Illumina platforms; GS FLX and GS Junior are Roche 454 platforms; and PacBio RS is a Pacific Biosciences platform. (Adapted from Levy and Myers, 2016 and http://dx.doi.org/10.6084/m9.figshare.100940)

1.5.4 Technical overview of Illumina sequencing

Illumina technology, previously called Solexa (www.illumina.com), is characterized by short sequence read length (~100 bp) and its sequence by synthesis mechanism using bridge amplification. All the enzymatic processes and imaging steps of the Illumina technology occur in a flow cell. Depending on the specific Illumina platform the flow cell may be separated into 1 (miSeq), 2 (HiSeq2500) or 8 (HiSeq2000, HiSeq2500) separate lanes (Knief, 2014; Magi et al. 2010; Mardis et al. 2008; Quail et al. 2012).

Three main steps are involved in Illumina sequencing 1) Library preparation, 2) Cluster generation and 3) Sequencing. In library preparation DNA is fragmented into individual molecules, followed by ligation of these fragments with Illumina specific adaptors. Forward and Reverse oligonucleotides which are complementary to the adapter sequences are introduced during the library preparation on the surface of the flow cell lanes. Cluster generation occurs on a flow cell after the initial library molecules are copied and removed. The fragments attached to the flow cell are used to generate cluster of identical template molecule using isorthermal amplification. This occurs through cycle alternations of three specific buffers that mediate the denaturation, annealing and extension steps. During these steps a bridge structure is formed when the 3' end of copied library molecules hybridize to the complementary oligo on the flow cell. The next step is the removal of one strand of dsDNA fragment using the cleavable site in the surface oligo, and blocks all the 3' end with ddNTPs to prevent open 3' ends to act as sequence primer site on adjacent library molecules (Bentley et al. 2009; Knief, 2014; Kozich et al. 2013; Marozova et al. 2008).

Sequencing is done by sequence by synthesis approach, where all the four nucleotides are added simultaneously to the flow cell. During these cycles, fluorescently labelled nucleotides are incorporated into the growing DNA chain. Each of these nucleotides has a single different fluorescent label which allows identification and act as reversible terminator to prevent multiple extensions. After imaging, the fluorescent group is removed, the reversible terminator is deactivated and the template strand is ready for the next incorporation cycle. The sequence is read by following the fluorescent signal per extension step for each cluster. The cycle time differs between different recently available Illumina platforms. This is caused by the imaging in the flow cell, therefore reduced imaging time speeds up the sequencing time. This was implemented in HiSeq2500 platform by giving and option to decrease the total surface area to be imaged (Mardis, 2008; Metzker, 2010; Shendure and Ji, 2008; Shokralla et al. 2012; Turncatti et al. 2008).

CHAPTER 2

Introduction

Geminiviruses are plant-infecting viruses (Rojas et al. 2005) that cause severe economic losses to agricultural production worldwide (Harisson, 1985). They are classified under the family *Geminiviridae,* which is the largest and the most economically important group of plant infecting viruses (Alabi et al. 2016; Brown et al. 2015). Viruses belonging to the family *Geminiviridae* are distinct in having genomes of circular, single-stranded DNA contained in twinned isohedral capsules. Their genome can be either monopartite (comprising a single ssDNA component called DNA-A) or bipartite (comprising two ssDNA components called DNA-A and DNA-B) (Harrison, 1985; Idris et al. 2005). The largest and the most diverse genus in the family is the *Begomovirus* genus. Begomoviruses are transmitted exclusively by the whitefly, *B. tabaci* in a persistent, circulative manner and infect a wide range of eudicots hosts (Mansoor et al. 2003; Moriones and Navas-Castillo, 2000; Varsani et al. 2017) and non-cultivated plants around the world (Alabi et al. 2016; da Silva et al. 2011; Graham et al. 2010; García-Andrés et al. 2006; Höfer et al. 1997; Kashina et al. 2002; Tahir et al. 2010; Wyant et al. 2011). The symptoms of a plant infected by a begomovirus display upward curling of leaf, yellowing of leaf veins, foliar crinkling, mosaic/striations, reduction of leaflet area, which results to no flower then no yield (Idris et al. 2005; Varsani et al. 2017).

Geminiviruses cause server economic losses worldwide and are a serious problem in South Africa where, to date, they have been reported to threaten production of cassava, tomato, sweet potato and bean crops (Berrie et al. 2001; Esterhuizen et al. 2012). They are also a serious problem in Africa, where they pose a risk to food security. They threaten production of two of the continent's staple food crops, namely cassava and maize (Bernado et al. 2013; Martin and Shepherd, 2009). The management of the diseases caused by these viruses are challenging and requires an integrated "pest" and "vector" management strategy (Antignus, 2007). Such control strategies require detailed information on the complex vector/virus/host pathposystems interactions, specifically aiming at understanding the relevant factors that contribute to the emergence of these diseases. Information on viral diversity, distribution and their host range are especially important for the development of disease control methods, such as resistant cultivars and cultural control methods aimed at preventing ongoing losses.

In recent years, identification methods for geminiviruses have largely relied upon molecular methods, such as Southern blotting with degenerate, genus specific probes or PCR with specific or degenerate primers. These techniques suffer from several significant drawbacks, especially when aiming to identify 'unknown' or 'new' viruses (Bexfield et al. 2011; Roossinck, 2011, Roossinck et al. 2015). Such specific assays require prior knowledge of the infective agent. However, data is frequently incomplete or even totally absent, especially for many of the more obscure or novel hosts, such as ornamentals or weeds. Recently technologies such as RCA

(Haible et al. 2006) and next generation, high-throughput, parallel sequencing platforms have provided powerful new tools in identifying (new) viral agents (Idris et al. 2014; Mardis, 2013; Roossinck, 2011). One such promising approach is the use of viral metagenomics (VM). Viral metagenomics is a molecular technique that involves purifying DNA or RNA from samples and shotgun sequencing the nucleic acids (Delwart, 2007; Edwards and Rohwer, 2005; Rasario and Breitbart, 2011). This method circumvents the biased nature of classic viral identification methods, allows for characterisation of the complete viral community, including viruses that are too divergent to be detected by PCR assays based on known viral sequences and has revolutionized the exploration of viral diversity (Breitbart et al. 2002; Dinsdale et al. 2008; Muthukumar et al. 2009; Ng et al. 2009).

In spite of the above mentioned advances in viral detection methods, current knowledge of plant virus diversity is still biased towards agents of visible and economically important diseases. Little is known about viruses that have not caused major diseases in crops, or viruses from indigenous or non-cultivated plant species (weeds). Numerous weed species or indigenous plants growing alongside cultivated crops have been reported to harbour begomovirus infection. This supports the notion that weed species act as natural hosts or begomovirus reservoirs, keeping the virus in the agroecosystem during the non-crop seasons and supporting mixed infections that lead to recombination, diversity, and subsequent evolution of these viruses (Ambrozevicius et al. 2002; Azhar et al. 2011; García-Andrés, et al. 2006; George et al. 2014). Discovery of these plant viruses are hindered by the traditional approach of sampling individual symptomatic plants, as many weeds or indigenous plants are non-symptomatic hosts of begomoviruses. Infection by these essentially unknown or poorly studied plant viruses have the potential to further complicate disease management in future and should therefore also be a focus of study.

When studying indigenous or non-cultivated plant species (weeds), accurate identification of the plant species can be problematic, since it requires thorough knowledge of the flora inhabiting certain areas and normally depends on the Floras and checklists published by taxonomists. However, the use of plant identification keys requires specialist understanding of plant morphology, and can be challenging for the untrained. A novel methodology known as DNA barcoding, in which short standardised gene regions are used for species identification, has the potential to mitigate the challenges posed by morphological identification of plant species (Hollingsworth et al. 2011). This approach has several benefits over the existing system of morphological identification and is applicable to all plant life stages (identification even when flowers or fruit are unavailable). World-wide effort on DNA barcoding has, to date, led to a degree of standardization in the process of barcoding, and two regions, *matK* and *rbcLa*, are currently being used as standard barcodes for land plants (CBOL, 2009; Hollingsworth et al. 2011).

The overall aim of this project was therefore to characterize the circular viral (DNA) diversity infecting indigenous or non-cultivated plant species (weeds) in various cropping systems in South Africa, using a metagenomics approach. RCA in combination with a high throughput next generation sequencing technology, a technique referred to as "circomics", was used in this study to allow rapid viral diagnosis and identification (Wyant et al. 2011). This information on viral diversity and host range will present an indication of weed species that act as reservoirs of geminivirus diversity that may play a role in viral evolution in South Africa. It will also allow for the discovery of viruses from indigenous plants that have not been detected by traditional approaches, based on sampling of symptomatic crop plants.

CHAPTER 3

Material and Methods

3.1 Sample collection and DNA isolation

Weed samples (non-cultivated plants) were collected from various vegetable cropping regions in Limpopo, Mpumalanga and Kwa-Zulu Natal provinces of South Africa (Figure 3.1) between February and April in 2014 and again in 2015. Weed samples were collected from and around vegetable crops, such as tomatoes, cucurbits, sweet potatoes, peppers and beans, particularly crops where high numbers of whiteflies (*B. tabaci*) were observed or in and around tomato crops showing begomovirus-like symptoms to increase the likelihood of finding whitefly-transmitted begomoviruses. From each plant, several leaves were collected and placed into a small plastic bag with silica gel (Merck) for later DNA isolation from the desiccated plant material. A voucher specimen was prepared by pressing the plant material. After the plant was dried out, it was mounted on a sheet and additional photographs were taken, in line with the Barcode of life data system (Ratnasingham and Herbet, 2007).

Plant DNA was isolated from 0.5 g of silica dried plant material. DNA extraction was done using the Macherey-Nigel Nucleospin plant DNA extraction kit following the manufactures protocol or by using the CTAB method described by Doyle and Doyle (1987). Ten ml of extraction buffer (10x CTAB) containing 80 µl of beta mercaptoethanol in 50ml blue Falcon tubes were preheated at 65 ˚C for 20 min. Leaf material was grounded (with sand/ using liquid nitrogen) using mortar and pestles until powder and a small portion of buffer were added to the grounded material until a slurry was obtained. The remainder of the buffer was added to the slurry and poured into the blue cap tube and placed immediately in a water bath at 65 ˚C for 20 min. An equal volume of 10 ml of SEVAG (Chloroform and Amyl alcohol) was added to the slurry and the tube was shaken in an orbital shaker (453 x g) for 60 min. The tubes were centrifuged at 3219 x g at 25 ˚C for 20 min and aqueous (top) phase contained the DNA was transferred to a clean blue cap tubes. Thereafter, DNA was precipitated with 100 % ethanol and stored at -20 ˚C. The precipitate was centrifuged at 1181 x g for 5 min. The pellets were collected and washed with a 3 ml of 70% ethanol and dislodged pellet to facilitate washing for 30 min. The washing step was repeated twice. The DNA was then centrifuged at 1811 x g for 5 min. The supernatant was poured off to allow the alcohol to evaporate overnight. The DNA was re-suspended in water and stored at -20 ˚C. As a final step, the DNA was again purified using QIAquick silica columns (Qiagen Inc.). The extracted DNA was separated on a 1% agarose gel (SeaKem) stained with 1 µg/mL ethidium bromide (EtBr) (ThermoFisher Scientific) and viewed under UV light.

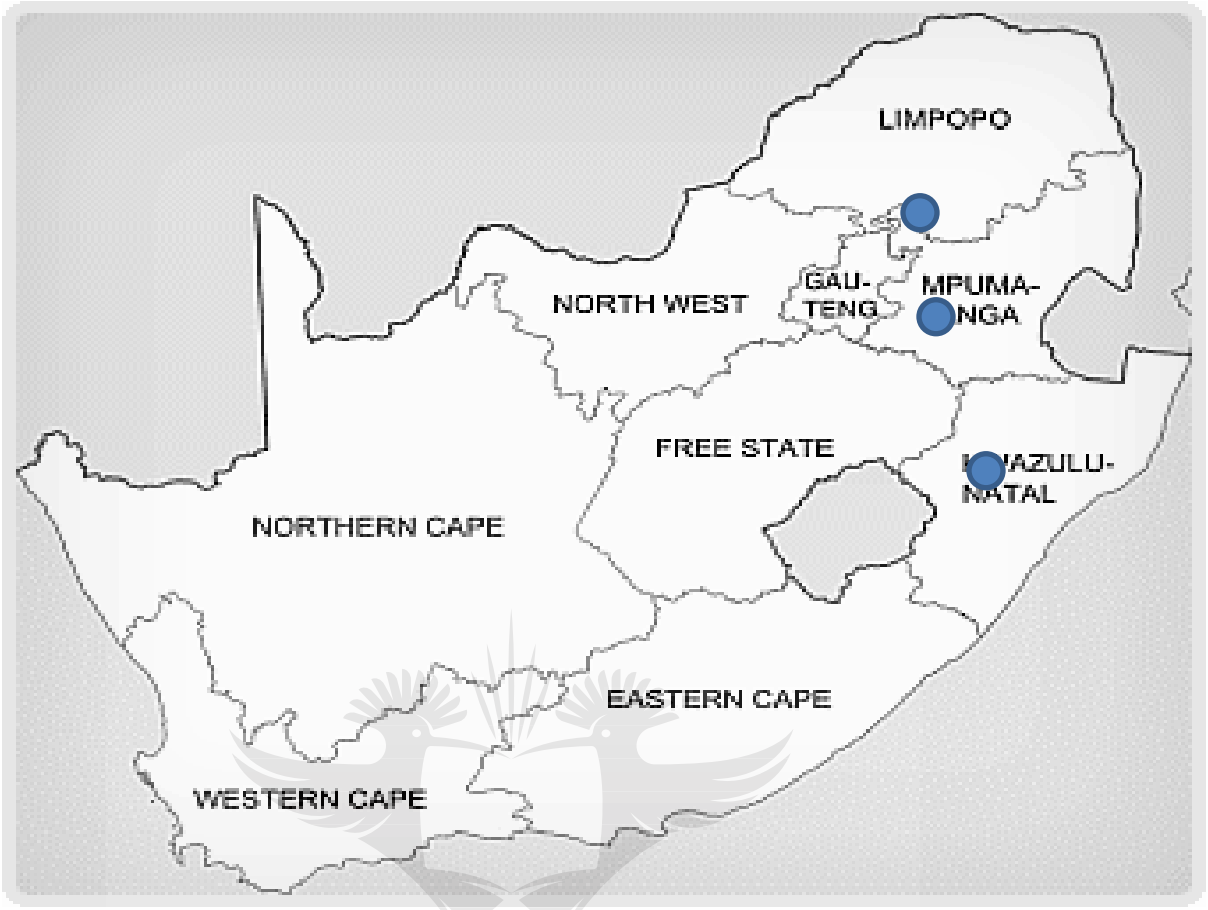


Figure 3.1. Geographical map of South Africa showing farming regions sampled during this study (blue circles). 1: Limpopo (Mooketsi, Trichardsdal, Tzaneen…); 2: Mpumalanga (Malelane, Koomatipoort); 3: Kwa-Zulu Natal (Peacevale, Mkomaas, Oribi Gorge,St Faiths and Ketoridge area).

3.2 Begomovirus detection using PCR

All plant samples were screened for begomovirus infection by PCR, using universal begomovirus primers, TY1(+): 5'-GCCCATGTA(T/C)CG(A/G)AAGCC-3' and TY2(−): 5'-GG(A/G)TTAGA(A/G)GCATG(A/C)G-TAC-3' (Table 3.1 Accotto et al. 2000). The primer pair amplifies a 580-bp DNA fragment comprising the V1 gene (coat protein, CP) from the DNA-A component of begomoviruses. PCR was performed using Deep Vent DNA polymerase (New England BioLabs), using an Eppendorf thermal cycler. Each PCR was carried out in 25 µL volumes and contained a final concentration of 12.5 µl One taq ready mix, 1 µl forward primer, 1 µl reverse primer, 9.4 µl of water and 1-2 µl of total DNA extracted from the collected samples. The cycling parameters were as follows: initial denaturation at 94 °C for 2 min, then 35 cycles of denaturation at 94 °C for 20 sec, annealing at 55 °C for 30 sec and elongation at 72 °C for 40 sec, followed by a final elongation step at 72 °C for 10 min. Amplified fragments were separated

on a 1% agarose (SeaKem) gel stained with 1 µg/mL ethidium bromide and visualized under UV light.

Selected TY-PCR positive samples (positive for begomovirus infection) were further characterised using RCA, a method which enabled a full length viral genome amplification of circular DNA viruses and their respective satellites. Viral diversity was determined by restriction digestion of RCA products using *Hpall* (RCA-RFLP). Circular viral DNA was amplified by using the TempliPhi-kit (GE Healthcare) following the manufacture's protocol. The RCA product (2.5 µl) was digested with *Hpall* (Fementas) at 37 °C for 2 h and separated on a 3% agarose gel stained with ethidium bromide and viewed under UV light. The RCA products were then used as templates for NGS sequencing (Pool 1) as described below.

3.3 Virus and satellite detection using RCA

Because the use of the TY primers to screen the plant material for typical begomoviral infection (3.2) could limit the detection of unknown or more diverse viral or satellite molecules, all the TY-negatives samples were also characterized using RCA, a method that would enable amplification of any unknown circular viral genomic material or circular virus associated satellites. Circular viral DNA was amplified using the TempliPhi-Kit (GE Healthcare) following the manufacturer's protocol as described above. The unknown RCA amplified products in these samples were again characterized by RCA-RFLP using *Hpa*II digestion as described above. The RCA products were then used as templates for NGS sequencing (Pool 2) as described below.

3.4 Virus and satellite characterization using NGS

To ensure characterization of all possible circular viral DNA genomes and satellite molecules, selected samples were sequenced using an RCA/NGS approach, a method that enables unbiased characterization of viruses with circular genomes because it does not rely on any prior sequence information (or primer selection). Two separate rounds of NGS were carried out during this study, as outlined below: After sample collection in 2014, selected TY-PCR-positive and RCA/RFLP-positive samples were pooled together as pool-1, and selected TY-PCR-negative and RCA/RFLP-positive samples were pooled as pool-2 (Figure 3.2). Pool-1 contained thirteen samples and pool-2 contained twenty two different plant DNA samples, in total. The RCA amplified products from the two pooled samples were directly used as templates for next generation sequencing in round-1 (in two separate MiSeq runs), at the ARC bioinformatics platform at Onderstepoort, Pretoria. Samples collected in 2015 were sequenced in round-2. The 2015 samples were analysed the same as samples collected in 2014, but all TY-PCR positive and RCA/RFLP-positive samples were sequenced individually (not pooled). A total of 40 individual, RCA amplified plant samples were sequenced in round-2. For the preparation of the

Illumina libraries, all the DNA samples were prepared using the Nextera DNA sample preparation kit (Illumina) and sequenced using an Illumina MiSeq at the ARC Biotechnology platform in Pretoria, South Africa, using 4 Truseq SBS chemistry with read lengths of paired-ends (2 x 125 bp).

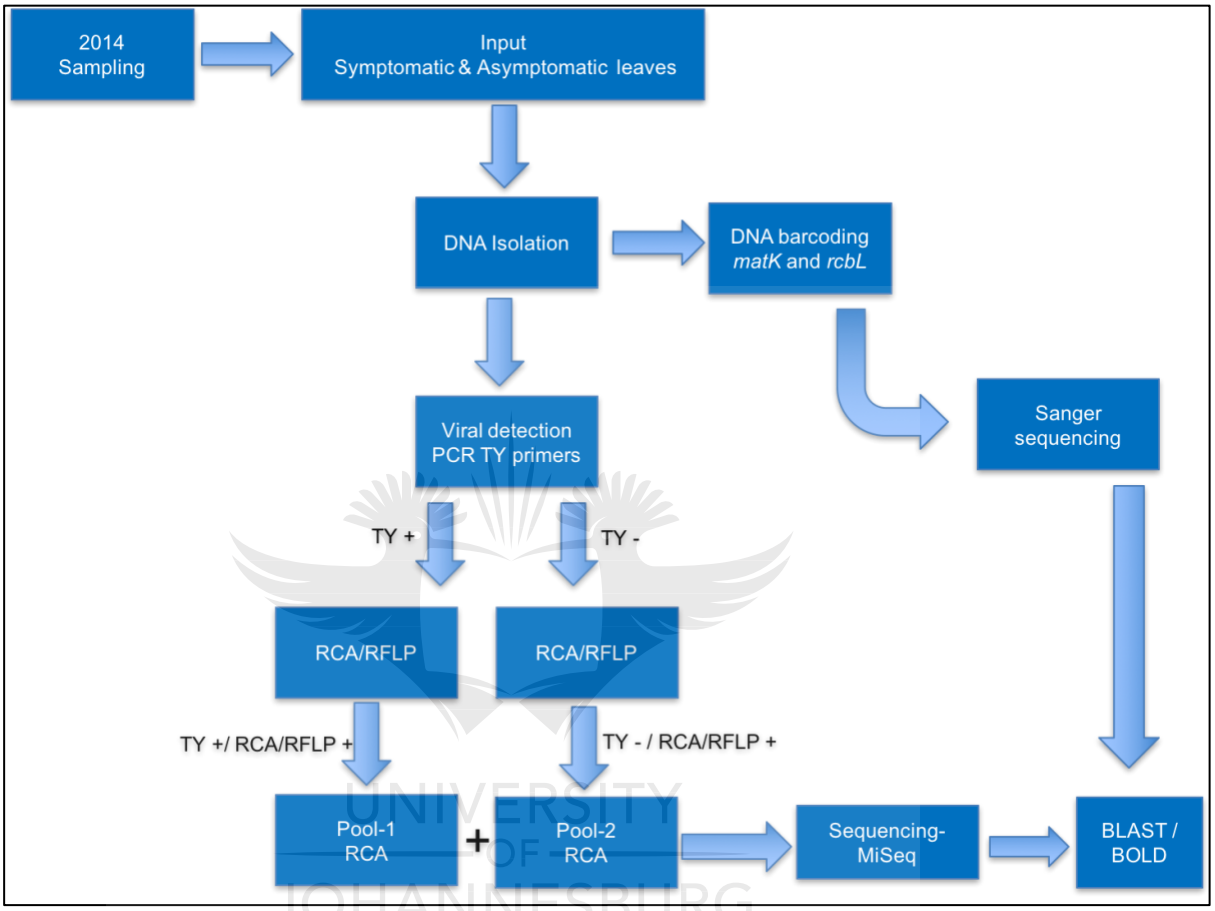


Figure 3.2. Diagram showing the metagenomics strategy used for screening and sequencing of the 2014 collected plant samples. All the plant samples were screened for begomovirus infection using the TY primers, and both samples that tested positive (TY+) and negative (TY-) using these primers were subjected to RCA/RFLP analysis. RCA amplified DNA from samples that were TY+ and produced a distinctive RCA/RFLP pattern (RCA/RFLP +) were combined in Pool-1 and RCA from samples that were TY- but RCA/RFLP +) were combined in pool-2. Pool-1 and pool-2 were combined and subjected to NGS sequencing using the Illumina MiSEq platform, followed by data analysis and homology searches using BLAST. Identification of the host plant was achieved using DNA barcoding by amplification and sanger sequencing, followed by homology searches using BLAST and BOLD.

3.5 Illumina sequence data analysis

For the 2014 sample set, two RCA-Illumina paired-end datasets were generated from the MiSeq platform for both the Pooled-1 and Pooled-2 samples, and for the 2015 samples, 40 RCA-Illumina paired-end dataset were generated (total of 42 individual libraries). All the datasets were analysed using CLC Genomics version 10. Raw reads were imported as paired-end (distance of 180 - 250 nucleotides) and quality control for each dataset was performed using the QC sequencing report embedded in the software as well as FastQC function (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). For the RCA data sets, a quality threshold of 30 and a Phred score of 33 were selected for trimming options. Reads were filtered by removal of low quality sequences (limit of 0.05), removal of ambiguous nucleotides (maximum of 2 nucleotides allowed) and removal of adapter sequences (Nextera V2 transposase 1 forward/reverse: GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG; Nextera v2 transposase 2 forward/reverse: TCGTCGGCAGCGTCAGATCTGTATAAGAGACAG). Each RCA data set was then assembled into contigs using the *de novo* assembly tool. The *de novo* assemblies were performed with the following settings: automatic word size (20), bubble size (50), minimum contig length of 200 nucleotides, length fraction (0.5), similarity (0.8), mismatch cost of 2, insertion cost of 3, and deletion cost of 3. The *de novo*-generated contigs were then submitted for multiBLASTn searches against the nucleotide collection at the NCBI portal (http://www.ncbi.nlm.nih.gov/) via the CLC interface. The full-length reference sequences of the viruses detected in the BLASTn searches were retrieved and used in subsequent reference-guided assemblies for selected samples. Alignments were imported into Bioedit (Tom Hall, Isis pharmaceuticals, Inc. 1997-2004) for viewing and editing. Pairwise sequence comparisons were computed with the MUSCLE algorithm (Edgar, 2004) implemented in SDT v 1.0 (Muhire et al. 2014). Phylogenetic analysis based on the assembled genome sequences was performed in Mega 7 (http://www.megasoftware.net/) using the Neighbour-joining (NJ) or Maximum likelihood method with 1000 bootstrap replications.

Table 3.1 Primers used in viral and satellite genome detection (PCR) and sequencing in this study.

Primer	Primer sequence (5'-3')	Expected product size (nt)
TY-1	GCCCATGTA(T/C)CG(A/G)AAGCC	500
TY-2	GG(A/G)TTAGA(A/G)GCATG(A/C)GTAC	
DNA101	CTGCAGATAATGTAGCTTACCAG	~1300
DNA102	CTGCAGATCCTCCACGTGTATAG	
UNI101	AAGCTTGCGACTATTGTATGAAAGAGG	~1300
UNI102	AAGCTTCGTCTGTCTTACGAGCTCGCTG	
Alpha1F	CTATACATCGGGTGGAACCG	~1300
Alpha15	AACCATCCCCTCCTGAAAAG	
1.52CF	CTCTGCAATAGCGCGTACCTGG	~1300
1.5CR	GGAAAGGCCATGCAGTCACC	
PBL1v2040	GCCTCTGCAGCARTGRTCKATCTTCATACA	600
PCRc1	CTAGCTGCAGCATATTTACRARWATGCCA	
Beta 01	GGTACCACTACGCTACGCAGCAGCC	600-700 & ~1350
Beta 02	GGTACCTACCCTCCCAGGGGTACAC	
DNA-AX1F	CCGGGTGTCCGTACTATGTT	∞
DNA-AX1R	CAGGTCCCCGGAGAGTATGA	∞
DNA-AX2F	CACCCATGCATCTAACCCAGT	∞
DNA-AX2R	GAACGTATCTTCGCACCACC	∞
DNA-AX3F	GGTGGTGCAAGATACGTTC	∞
DNA-BW01F	CCGATTCGTCCGTACTATGTATTAGG	∞
DNA-BW01R	GGACAACTGTAAGTGTACCCTTGTACC	∞
DNA-BW03R	GCGTGCAGCAGCGTATTGTT	∞
DNA-BW04F	TCCCCACGTTATCCTGTGTG	∞
DNA-BW04R	AAGTTCCGATCAACGCCTCC	∞
BetasatS1-2F	GGCTGGTGACTTTGTCCTTGT	∞
BetasatS1-2R	CTATCACCAGTAGGACCTCCG	∞
BetasatNGS3F	GGCTGGTGACTTTGTCCTTGT	∞
BetasatNGS3R	GTATCACCAGTAGGACCTCCG	∞

Primers were designed with the web-based software program Primer3 (http://frodo.wi.mit.edu /primer3/), integrated DNA technologies (IDT) scitools (http://scitools.idtdna.com/scitools/, applications/ OligoAnalyzer) analysis was used to eliminate primers with stem loop structures or that self-annealed and all primers were synthesized by Inqaba Biotec, South Africa. ∞ Primers were used for sequencing. The primers labelled DNA-A and DNA-B were designed to specifically

amplify the DNA-A and DNA-B components respectively, in sample 242 and was used to primer walk the full genome sequence (~2700bp) in 900bp stretches. The BetasatS1 primers were designed to specifically amplify the Beta-satellite in sample 10 and BetasatNGS for the Beta-satellite in sample 30.

3.6 Virus and satellite detection and confirmation by PCR amplification and sequencing.

Data analysis revealed the presence of novel begomoviral DNA-A, DNA-B genomes, putative alpha and beta-satellite molecules, as well as completely novel, putative viral genomes. In order to confirm the presence of these viral contigs in the 2015 samples, and in addition to determine their respective plant hosts in the 2014 pooled NGS samples (pool-1 and pool-2), PCR amplification and sequencing were performed on selected samples. Based on the viral contigs assembled during *De novo* assembly, several primer pairs were selected from literature or designed (Table 3.1) to detect their presence. To detect the alpha and beta satellites in the individual weed samples, PCR was performed using the following Beta-satellite specific primers, Beta 01/Beta 02, DNA101/DNA102, UN101/Un102, and alpha-satellite specific primers, Alpha1F/Alpha1R primers, respectively (Table 3.1 Briddon et al. 2002). The primers bind to a region conserved for all begomovirus associated DNA satellites and amplifies band of 700-1350 bp. The PCR was performed with Deep Vent DNA polymerase (New England BioLabs), using an Eppendorf thermal cycler. The PCR was carried out in a 25 µl volume, containing 12.5 µl One taq ready mix, 1 µl forward primer, 1 µl reverse primer, 9.4 µl of water and 0.5-1 µl RCA amplified viral concatamers. The reaction was carried out using the following cycling conditions at 94 ˚C for 3 min, then 35 cycles of denaturation at 94 ˚C for 1 min, annealing at 62 ˚C for 1 min and elongation at 72 ˚C for 1.5 min, followed by final elongation step at 72 ˚C for 10 min. Amplified fragments were separated on a 1% agarose (SeaKem) gel stained with 1 µg/mL ethidium bromide (ThermoFisher Scientific) and visualized under UV light. The complete beta satellite and alpha satellites of selected isolates were sequenced directly, by primer walking at Inqaba Biotech using Bigdye cycle sequencing kit.

In order to detect the DNA-A and DNA-B viral contigs identified by NGS, the bipartite viral genomes in the individual samples were amplified using DNA-AX1F/DNA-AX1R primers and DNA-BW01F/DNA-BW01R primers, respectively (Table 3.1).The primer pairs amplifies 2500bp. The PCR was performed with Deep Vent DNA polymerase (New England BioLabs) using Eppendorf thermal cycler. The PCR was carried out in a 25 µl volume, containing 12.5 µl One taq ready mix, 1 µl forward primer, 1 µl reverse primer, 9.4 µl of water and 0.5-1 µl RCA amplified viral concatamers. The reaction was carried out using the following cycling conditions at 94 ˚C for 3 min, then 35 cycles of denaturation at 94 ˚C for 1 min, annealing at 60 ˚C for 1 min and

elongation at 72 ˚C for 1 min, followed by final elongation step at 72 ˚C for 10 min. The complete DNA-A and DNA-B components of selected viral isolates were sequenced by primer walking (Table 3.1) using BigDye cycle sequencing kit at Inqaba Biotech.

3.7 Plant identification using DNA Barcoding of the *rbcLa* and *matK* locus.

To identify the plants infected by begomoviruses or any of the other detected viral isolates, PCR amplification was done using two sets of barcoding primers, namely *rbcLa* and *matK* (Table 3.1) (CBOL, 2009). The *rbcLa*F and *rbcLa*R primer pair amplifies ~580 bp of ribulose 1,5 biphosphate carboxylase (Bafeel et al. 2011). The matK_390F and matK_1326R primer pair amplifies ~930 bp of Maturase K (Coénoud et al. 2002). The PCR was done at the African Centre for DNA Barcoding (ACDB), University of Johannesburg using Epperndof thermal cycler. The PCR was carried out in a 25 µl volume, containing 12.5 One taq ready mix (New England BioLabs), 0.3 µl forward primer, 0.3 µl reverse primer, 0.8 µl BSA, 9.4 µl of water and 2 µl DNA. A volume of 0.5 µl dimethyl sulfoxide (DMSO) was added to the *matK* PCR reactions. The reaction for *rbLa* was carried out using the following cycling conditions: 94 ˚C for 3 min, then 28 cycles of denaturation at 94 ˚C for 1 min, annealing at 48 ˚C for 1 min and elongation at 72 ˚C for 1 min, followed by final elongation step at 72 ˚C for 7 min. The reaction for *matK* was carried out using the following cycling conditions: 94 ˚C for 1 min, then 35 cycles of denaturation at 94 ˚C for 30 sec, annealing at 50 ˚C for 1 min and elongation at 72 ˚C for 1 min, followed by final elongation step at 72 ˚C for 5 min. The amplified DNA fragments were separated in 1% agarose gel stained with EtBr and viewed under UV light. The PCR product was purified and used as template for cycle sequencing. The cycle sequencing reaction was performed using BigDye Terminator Cycle Sequencing Kit, using 3730XL DNA Analyser (Applied Biosystems) available in Botany Department at the University of Johannesburg. The reaction was carried out in a 10 µl volume, containing 0.3 µl BigDye, 5X sequencing buffer, 0.3 µl forward or reverse primer, 0.5 µl DMSO (for matK), 6.9 µl water and 2 µl DNA. The reaction was carried out using the following cycling conditions at 95 ˚C for 10 min, then 28 cycles, annealing at 50 ˚C for 5 sec and elongation at 60 ˚C for 4 min.

Table 3.2 Primers used for DNA Barcoding in this study (Bafeel *et al*., 2011; Cuénoud *et al*., 2002).

Primer	Primer sequence (5'-3')	Expected product size (nt)
rbcLaF	ATTCACCACAAACAGACTAAAGC	~580
rbcLaR	GTAAAATCAAGTCCACCRCG	
matK-KIM3F	CGTACAGTACTTTTGTGTTTACGAG	~930
matK-KIM1R	ACCCAGTCCATCTGGAAATCTTGGTTC	
matK390F	CGATCTATTCATTCAATATTTC	~930
matK1326R	TCTAGCACACACGAAAGTCGCGAACT	

CHAPTER 4

Results

4.1 Sample collection and virus detection

This study was conducted between 2014 and 2015 in farming areas in the provinces of Limpopo, Mpumalanga and Kwa-Zulu Natal. The study aimed to determine the diversity of circular DNA viruses, particularly geminiviruses, associated with indigenous plants growing in different cropping systems. A total number of 230 indigenous plants, which were growing in and alongside cultivated areas, were collected. The plants were at various growth stages. Some plants were growing around the border of crop lands and were more mature plants, but plants growing inside or alongside the cultivated crop land were normally immature, and without flowering structures (important for plant identification). The majority of the leaves showed no symptoms of begomovirus virus infection, only a few plants showed symptoms of yellowing, stunting and leave cupping. Total DNA was isolated from all the collected samples (labeled OHM0003 - OHM0228) and screened for begomovirus infection by TY-PCR and RCA/RFLP, as described above (Section 3.2 and 3.3). A total of 55 weed/indigenous plant samples (24%) tested positive for possible begomovirus infection using universal begomovirus primers TY1 and TY2 (termed TY positive (TY+). A total of175 plants tested negative using universal begomovirus primers (TY negative, (TY-)), but were RCA/RFLP positive (RCA/RFLP+), i.e. they did contain circular amplicons of possible viral origin and were included in the subsequent sequencing analysis. A total of 58 plant samples, (36 TY+ and 22 TY-/RCA/RFLP+) were included in the NGS analysis to determine the diversity of viral genomic components associated with the collected plant samples (Table 4.1).

Table 4.1. Samples collected in three Provinces in South Africa during the summer months of 2014 and 2015.

Number of samples per province	Provinces			Total
	Limpopo	Mpumalanga	Kwa-Zulu Natal	
Samples tested	86	111	33	230
TY+ samples	20	23	12	55
TY+ / RCA-RFLP+	7	9	4	20
TY - /RCA/RFLP+	20	14	11	5
Samples included in NGS	22	16	4	42

4.2 RCA/RFLP

All the collected samples were screened by RCA/RFLP to determine the diversity of circular genomic components present in the host plant material. RCA is an unbiased method that enables amplification of any circular DNA components in the isolated DNA. Restriction digestion using *Hpa*II then allowed analysis of the genetic diversity (represented by similar or diverse banding patterns between individual samples) to be determined. The RCA-RFLP profiles for a select number of samples (representative of multiple RCA/RFLP analysis) are indicated in Fig 4.1. Most samples produced a single, high molecular weight band (>5000 bp) and this band remained undigested after incubation with the restriction enzyme, *Hpa*II. Some samples produced a characteristic RCA/RFLP, with multiple bands of various sizes (<5000 bp), possibly indicative of viral genomic components. These samples were considered to be RCA/RFLP positive and were included in the subsequent NGS analysis to determine if the bands had a possible viral origin.

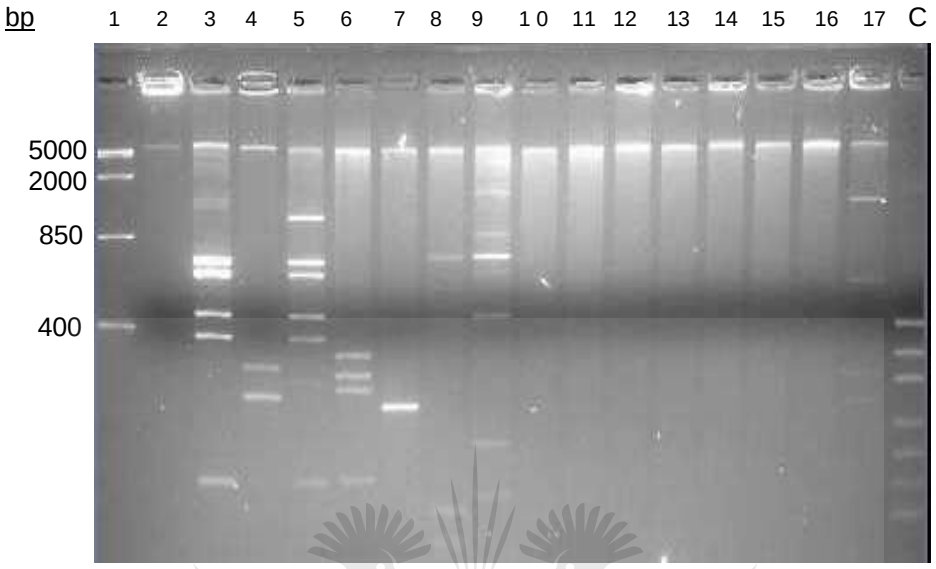


Figure 4.1. Rolling circle amplification-restriction fragment length polymorphism (RCA-RFLP) profiles from selected indigenous plants. Lane (1) GeneRuler Express marker 400-5000 bp, (2) OHM0062, (3) OHM0063, (4) OHM0066, (5) OHM0067, (6) OHM0069, (7) OHM0070, (8) OHM0072, (9) OHM0077, (10) OHM0079, (11) OHM0080, (12) OHM0087, (13) OHM0089, (14) OHM0090, (15) OHM0091, (16) OHM0092, (17) OHM0094, (C) positive control.

4.3 Virus characterization

Data analysis of the both the combined (2014) and individual (2015) plant NGS datasets were carried out to identify genomic sequences of possible viral origin. After *de novo* assembly of each dataset, comparison of the resulting contigs to the full nucleotide collection at GenBank database using multiBLASTn searches, identified a large number of contigs (89) with high sequence similarity to ssDNA viruses (Table 4.2 and 4.3). More specifically, a number of contigs had high sequence similarity to begomoviral DNA-A and DNA-B genomes, putative alpha and beta-satellite molecules, and genomes from the recently described family of *Genomoviridae* (Table 4.2). For both the combined and individual NGS data sets, the genome of a total of 12 ssDNA viruses or ssDNA virus associated molecules (satellites) were identified, 10 of which were verified through PCR, and were assembled to complete genomes completed (Table 4.2). For each of these identified contigs, the full-length reference sequences of the related viruses detected in the BLASTn searches, were retrieved and used in subsequent phylogenetic analysis and also for reference-guided assembly of the full genomes from the raw reads. Further data analysis, identification and phylogenetic analysis of the identified viral contigs will be discussed below

separately, according to (i) the NGS data set they were identified in, and (ii) according to the type of ssDNA virus or virus-associated molecule that was identified within the sample.

4.3.1 Detection of viral contigs in the Pool-1 and Pool-2 NGS data set.

As explained before, samples collected in 2014, and found to be TY PCR-positive and RCA/RFLP-positive, were pooled together as pool-1, and selected TY PCR-negative and RCA/RFLP-positive samples were pooled as pool-2. Pool-1 and pool-2 contained 13 and 22 different plant samples, respectively. Initial data analysis of the pool-1 dataset, revealed 21 contigs with BLASTn similarity to virus-associated genomes (Table 4.2). More specifically, 17 contigs had high sequence similarity (84% - 99%) to the DNA-A component of known begomoviruses with one contig having a 100% sequence similarity to a begomovirus-associated alphasatellite. One contig had 96% sequence similarity to a begomovirus-associated betasatellite and two contigs had 94% and 100% sequence similarity to two Genomoviruses previously identified from Dragonflies (Ng et al. 2011; Rosario et al. 2012). Initial data analysis of the pool-2 dataset revealed eight contigs with blast similarity to virus-associated genomes. Surprisingly, four contigs in pool-2 had high sequence similarity (86% - 91%) to the DNA-A component of known begomoviruses, despite testing negative in the initial TY-PCR (Table 4.2). Lastly, one contig had 92% sequence similarity to a begomovirus-associated betasatellite and three contigs had between 93% and 100% sequence similarity to a circular DNA molecule associated with *Clerodenron Golden mosaic China virus* (CGMV) (Zaffalon et al. 2012). In order to identify the individual plant samples from which these viral genomes in pool-1 and pool-2 originated, the plant samples that were pooled in pool-1 and 2 were subsequently screened for the presence of these viral genomes using PCR as explained in section 3.1.3. The individual plant samples were then included in the second round of NGS (2015), and further identification and phylogenetic analysis of these viral genomes will be explained in the sections below, according to the plant sample they were found in.

Table 4.2. Contig number, sequence identity and percentage genome coverage of a selection of the recovered ssDNA genomes of weed-infecting viruses and their associated ssDNA molecules, found in Pool-1 and Pool-2 by NGS.

Contig number[@]	Length (nt)	Closest sequence similarity [#]			% of full genome recovered
		Description	Accession	% Sequence similarity[#]	
1.8	2111	*Tomato curly stunt virus*	AF261885	98.8	77
1.9	2080	*Tomato curly stunt virus*	AF261885	98.5	69
1.15	2590	*Tomato curly stunt virus*	AF261885	99.3	90
1.37	2698	*Tomato curly stunt virus*	AF261885	96.9	96
1.311	2383	*Tomato curly stunt virus*	AF261885	98.6	84
1.16	2591	*Tomato curly stunt Lanseria virus**	-	96.4	93
1.19	2618	*Tomato curly stunt Mooketsi Virus**	-	88.0	94
1.52	2165	*Dragonfly-associated circular virus 3*	JX185428	61.0	98
1.1236	2198	*Dragonfly-associated circular virus 1*	JX185430	64.0	98
2.55	1372	*Cotton leaf curl Gezira alphasatellite*	HM446369	82.0	99
1.136	1308	*Tomato leaf curl Togo betasatellite*	HQ586965	96.7	94
2.435	1317	*Tomato leaf curl Togo betasatellite*	HQ586965	92.2	94
2.8	974	Circular DNA molecule associated with CGMV China[$]	FN658712	94	100

[#] Blast similarity or based on muscle alignment and similarity matrix calculation using SDT
* *Tomato curly stunt Lanseria virus* and *Tomato curly stunt Mooketsi virus* are two begomoviruses, previously identified in South Africa (Esterhuizen, 2012).
[$] *Clerodenron Golden mosaic China virus* (CGMV)
[@] The number 1 and 2 at the beginning of the contig number indicates the pool-1 or pool-2 origin of the contig.

Table 4.3 Information on South African begomoviruses and other viral genomic components (DNA-A) identified in this study.

Isolate	**Collection date**	**Province/district**	**Host**	**% identity to closest blast relative ToCSV AF261885**	**Acronym**
Tomato curly stunt virus (monopartite begomovirus)					
S16 SAMooketsi2014	01/2014	Limpopo, Mooketsi	*Malvastrum coromandelianum*		ToCSV-[ZA:Mks:2014]
S77SAKomatipoort2014	02/2014	Mpumalanga, Komatipoort	*Acalypha indica*	ToCSV, AF261885, 92%	ToCSV-[ZA:Kmp:2014]
S66 SAKomatipoort2014	02/2014	Mpumalanga, Komatipoort	*S. lycopersicum*	ToCSV, AF261885, 96%	ToCSV-[ZA:Kmp:2014]
S69 SAKomatipoort2014	02/2014	Mpumalanga, Komatipoort	*S. lycopersicum*		ToCSV-[ZA:Kmp:2014]
Malvastrum curly stunt virus [#](monopartite begomovirus)					
S10 SAMooketsi2014	01/2014	Limpopo, Mooketsi	*Malvastrum coromandelianum*	ToCSV, AF261885, 88,4%	MalCSV-[ZA:Mkt:2014]
Sida corlifolia golden mosaic virus (*bipartite begomovirus*)					
S30 SAMooketsi2011	01/2011	Limpopo, Mooketsi	*Sida corlifolia*	CoYVV, KX101212, 72%	SiCYMV-[ZA:Mkt:2014]
Phaseolus vulgaris begomovirus (bipartite begomovirus)					
S242 SA_Cato ridge_2015	2/2015	KwaZulu-Natal, Cato ridge	*Phaseolus vulgaris*	*Soybean chlorotic blotch virus, 70,3%*	*PaVBV* --[ZA:CaR:2015]
Sida corlifolia yellow mosaic betasatellite (begomovirus)					
S30 SAMooketsi2011	02/2011	Limpopo, Mooketsi	*Sida corlifolia*	ChaYMB, KT454829, 74,6%	SiCYMB-[ZA:Mkt:08])
Tomato leaf curl Togo betasatellite (begomovirus)					
S10 SAMooketsi2014-beta	01/2014	Limpopo, Mooketsi	*Malvastrum coromandelianum*	ToLCTGB, KT382329, 78%	ToLCTGB-[ZA:Mkt:2014]
***Cynoglossum offinale* associated circular DNA virus** (Genomovirus)					
S8 SA_Mooketsi2014_DATE	02/2014	Limpopo, Mooketsi	*Cynoglossum offinale*	ND	COasCV-[ZA:Mkt:2014]

[#] New species proposal

4.3.2 Detection of monopartite begomoviral genomes in individual plant samples.

Sample 66 and 69 were collected from tomato plants in Komatipoort (Mpumalanga) in February 2014 and sample 16 and 77 from weeds growing alongside tomato crops in Mooketsi (Limpopo) in January 2014 (Table 4.3). The plant material collected in this study was identified by DNA barcoding, as described in section 4.3.6 below. Samples 66 and 69 were confirmed as *Solanum lycopersicum* and samples 16 and 77 were identified as *Malvastrum coromandelianum* and *Acalypha indica*, respectively. One viral contig (~2.7 kb) was recovered from each of these MiSeq datasets and was found to represent a begomovirus, since it showed typical begomovirus DNA-A genome organization (Six ORFs: V1, V2, C1, C2, C3 and C4 and a 259 nt intergenic region (IR) containing the invariant nonanucleotide motif (TAATATTAC) (Figure 4.2). The complete nucleotide sequences of the four recovered begomovirus isolates ranged between 2758 - 2765 nt in size (Table 4.3). Based on a SDT pairwise alignment, the four genome sequences shared 92-96% nucleotide identity with ToCSV-[ZA-Ond-98] (AF261885) isolates previously found in tomato plants in South Africa. All four isolates shared >91% identity to ToCSV-[ZA-Ond-98] and according to the 91% species demarcation criteria (Brown et al. 2015), the viral isolate from sample 77 should be regarded as a strain (92% nucleotide identity) of ToCSV-[ZA-Ond-98] (<94% nucleotide identity) and the viral isolates from samples 16, 66 and 69 (94-99% nucleotide identity) should be regarded as isolates of ToCSV-ZA. All the newly identified ToCSV isolates were given descriptors indicating the location and date of collection (Table 4.3). Phylogenetic analyses grouped the four DNA-A nucleotide genomes of the tomato and weed-infecting begomoviruses within the SAI subclade, along with all other ToCSV isolates. Interestingly, the two ToCSV genomes recovered from weed species, *M. coromandelianum* and *A. indica* formed a separate subgroup, separate from the tomato-infecting ToCSV isolates (Fig 4.3).

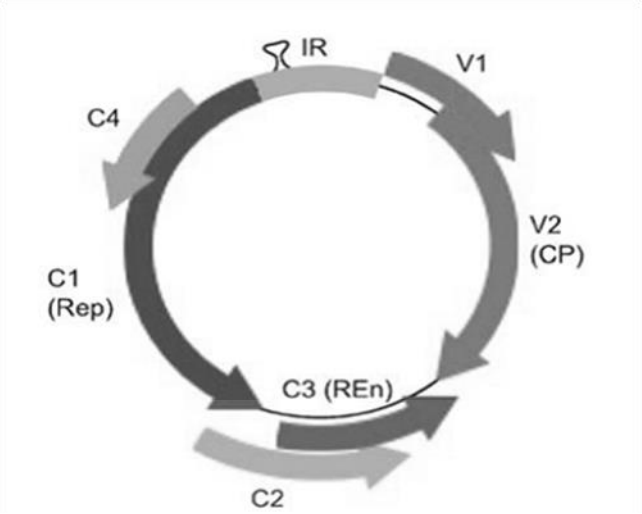


Figure 4.2. Schematic representation of the genome organization of the tomato-infecting begomovirus isolates sequenced in this study. The single-stranded virion DNA genome consisted of 2766-2800 nucleotides. All viral isolates had a genome organization similar to other monopartite begomoviruses containing six ORFs (direction indicated by arrows) namely: two virion-sense (V) ORFs, V1 and V2 and four complementary-sense strand (C) ORFs, C1, C2, C3 and C4. A 259 nt intergenic region (IR) contains the nonanucleotide motif TAATATTAC. The conserved inverted repeat flanking the nonanucleotide sequence is symbolised by a stem-loop. V1 encodes the capsid protein (CP), V2 a movement protein, C1 the replication initiator protein (Rep), C2 a transcriptional activator protein (TrAP), C3 a replication enhancer protein (REn), and C4 a symptom and movement determinant.

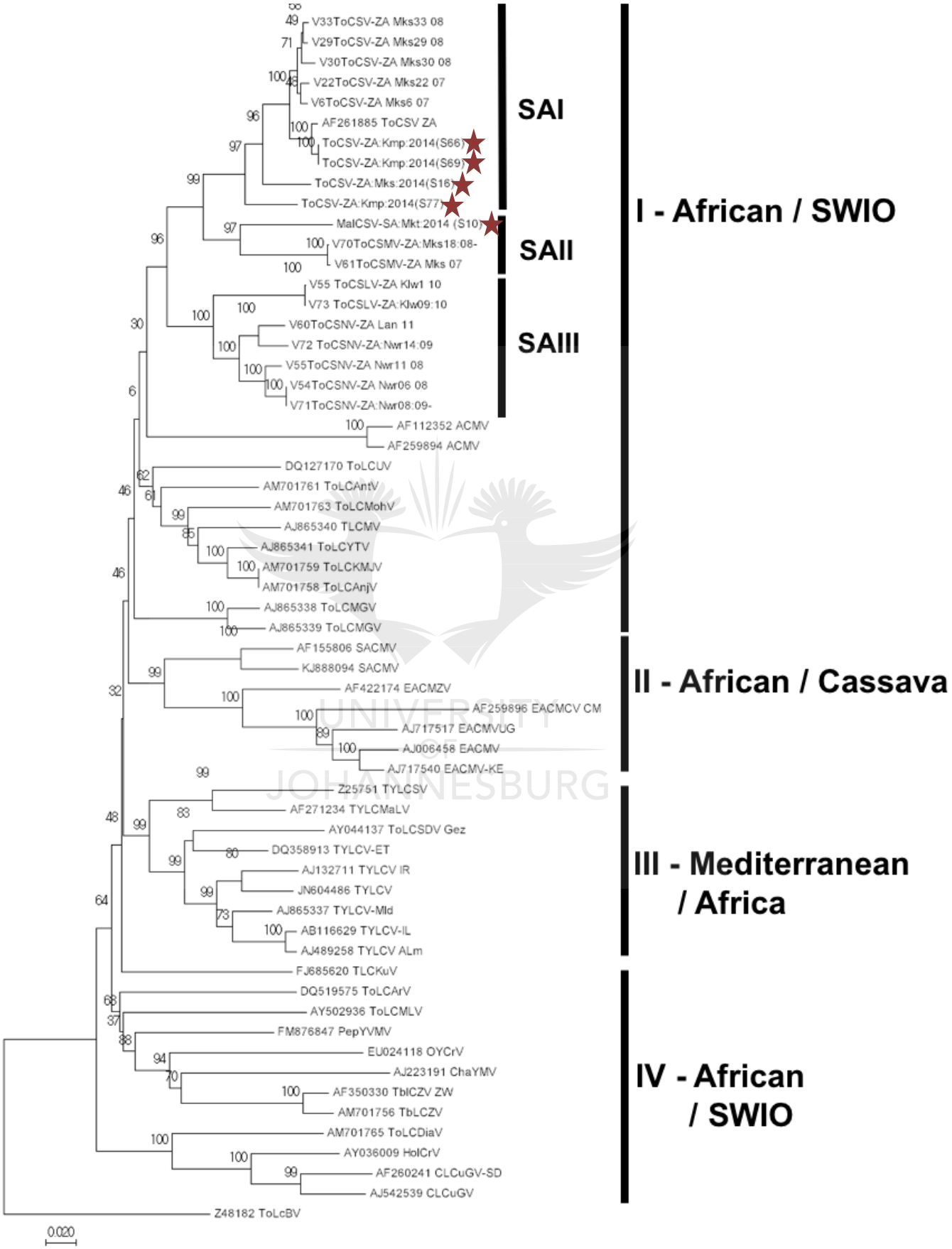


Figure 4.3. Phylogenetic tree showing the relationships between the newly identified South African begomoviruses and other begomoviruses from around the world. The tree was inferred using the Neighbor-Joining method (Saitou and Nei, 1987) using the MEGA7 program (Kumar et al. 2016). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (2000 replicates) are shown next to the branches (Felsenstein, 1985). The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Maximum Composite Likelihood method (Tamura et al. 2004) and are in the units of the number of base substitutions per site. Begomoviruses used for the phylogenetic tree are indicated by NCBI accession numbers on the figure.

Sample 10 was collected in the Mooketsi Valley (Limpopo) in January 2014. It was identified as *Malvastrum coromandelianum* by DNA barcoding, as described in section 4.3.6 below. One viral contig (~2.7 kb) recovered from sample 10 MiSeq dataset was found to represent a begomovirus, since it showed typical begomovirus DNA-A genome organization (Six ORFs: V1, V2, C1, C2, C3 and C4 and a 259 nt intergenic region (IR) containing the invariant nonanucleotide motif (TAATATTAC) (Fig 4.2). The complete nucleotide sequence of the begomovirus isolate recovered from the NGS data was determined to be 2767 bp in size. Based on a SDT pairwise alignment with their closest known relatives downloaded from NCBI, the viral contig shared the highest nucleotide sequence identity (88.4%) to *Tomato curly stunt Mooketsi virus*-[ZA:Mooketsi:2007] (ToCSMV-[ZA:Mks:07] (Table 4.3) found in tomato plants in Mooketsi in South Africa (Esterhuizen et al. 2012). According to the begomovirus species demarcation threshold of 91% (Brown et al. 2015), the begomovirus found in S10 represent an a new begomovirus species, for which we propose the name *Malvastrum curly stunt virus* - [South Africa:Mooketsi:2014] (MalCSV-[ZA:Mkt:2014]. Phylogenetic analysis grouped MalCSV-[ZA:Mkt:2014] within the SAII subclade, within the larger African/ South West Indian Ocean (SWIO) cluster, along with two isolates of ToCSMV (Fig 4.3). No DNA-B molecule was identified in the sample 10 dataset and repeated attempts to detect a DNA-B component failed using the 'universal' B-component-specific PCR primers (Table 3.1, primer table in methods). As discussed in section 4.1.3.3 below on beta-satellite detection, ToLCTGB was identified in the sample 10 dataset and successfully detected using the satDNA-specific PCR primers. This suggests that MalCSV is a monopartite virus that may require the associated ToLCTGB betasatellite for infection, but this remains to be investigated by construction and agroinoculation experiments with agroinfectious clones of this viral isolate and the associated satellite molecule.

4.3.3 Detection of begomoviral betasatellite genomes in individual plant samples.

Sample 10 was collected in the Mooketsi Valley (Limpopo) in January 2014 and was identified as *Malvastrum coromandelianum* by DNA barcoding, as described in section 4.3.6 below. One contig (~1.3 kb), recovered from the sample 10 Miseq dataset was found to represent a betasatellite,

since it showed typical genome organization of begomovirus-associated betasatellites (A-rich region, stem-loop structure (TAATATTAC seq), satellite conserved region and Beta-C1 gene (117 amino acids) in the complementary-sense strand) (Table 4.3). Based on a SDT pairwise alignment, the satellite shared the highest nucleotide sequence identity (78%) to *Tomato leaf curl Togo betasatellite* (ToLCTGB) (KT382329), found in tomato in Togo, Africa. The complete nucleotide sequence of the beta-satellite recovered from the NGS data was determined to be 1316 bp in size. According to the betasatellite species demarcation threshold of 78% (Briddon et al. 2008), the beta-satellite found in sample 10 represent an isolate of *Tomato leaf curl Togo betasatellite* (ToLCTGB), for which we propose the addition of the following designation ToLCTGB - [South Africa:Mooketsi:2014] (ToLCTGB-[ZA:Mkt:2014]). Phylogenetic analysis confirmed that ToLCTGB-[ZA:Mkt:2014] formed a cluster with other ToLCTGB sequences from Africa, more specifically from West Africa (Ghana and Togo) (Fig 4.4).

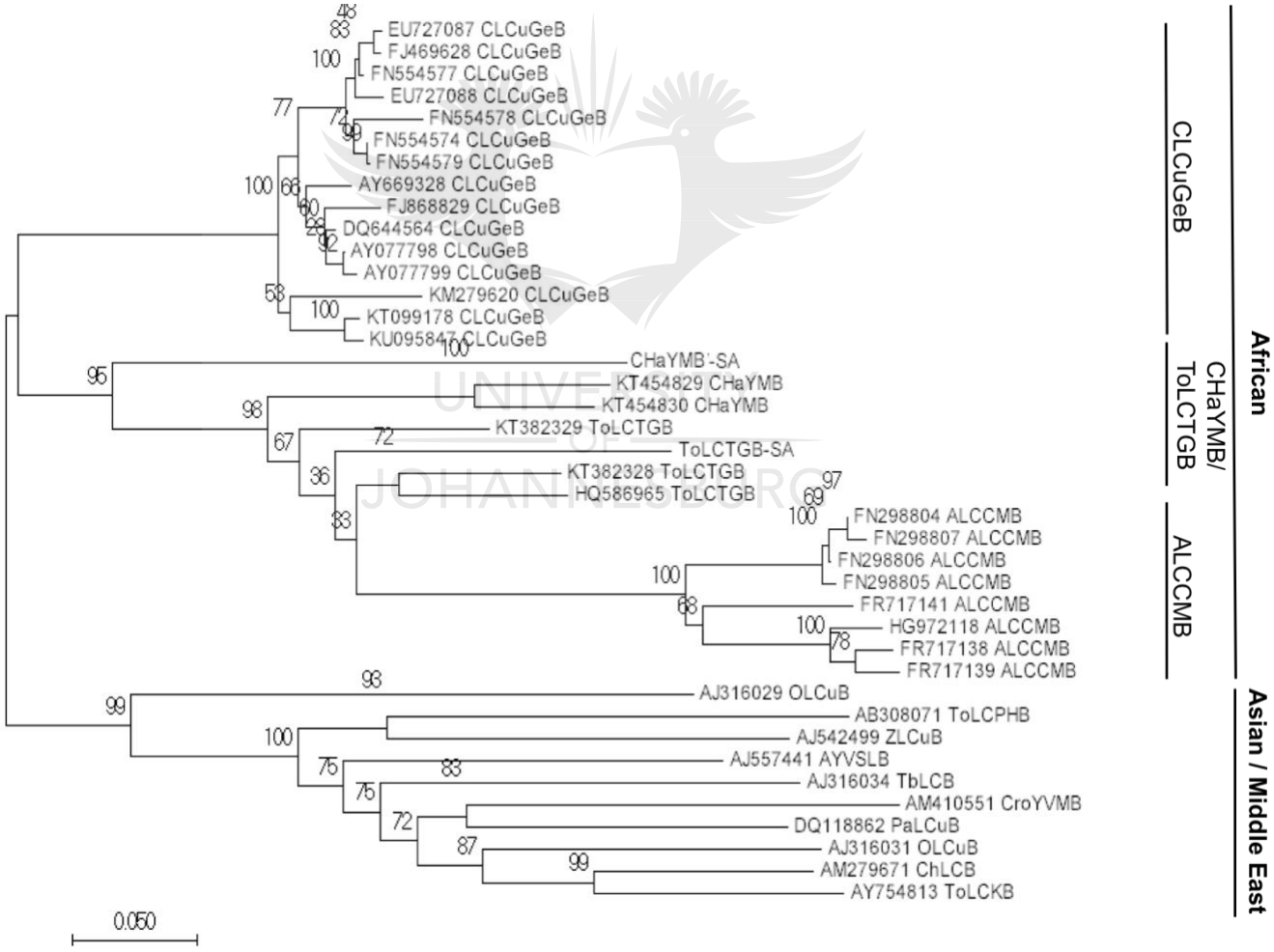


Figure 4.4. Phylogenetic tree showing the relationships between the two newly identified South African betasatellites and other betasatellites from around the world. The tree was inferred using

the Neighbor-Joining method (Saitou and Nei, 1987) using the MEGA 7 program (Kumar et al. 2016). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (2000 replicates) are shown next to the branches (Felsenstein, 1985). The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Maximum Composite Likelihood method (Tamura et al. 2004) and are in the units of the number of base substitutions per site.

Sample 30 was collected in the Mooketsi Valley (Limpopo) in 2011. Sample 30 was identified as *Sida corlifolia* by DNA barcoding, as described in section 4.3.6 below. The *S. cordifolia* plant sample 30 showed bright yellow mosaic symptoms on its upper leaves (Fig 4.5). One contig (~1.3 kb), recovered from the sample was also found to represent a beta-satellite, with typical beta-satellite genome organization (stem-loop structure (TAATATTAC seq), beta-C1 gene (117 amino acids)). According to the SDT based pairwise alignment it shared the highest nucleotide sequence identity (74.6%) to *Chayote yellow mosaic Benin betasatellite* (ChaYMB) (KT454829; KT454830), found in *Momordica charantia* (Bitter melon) plants showing yellow mosaic symptoms in Benin in 2014 (Table 4.3). The complete nucleotide sequence of this betasatellite was determined to be 1342 bp in size. According to the betasatellite species demarcation threshold of 78% (Briddon et al 2008), the betasatellite found in sample 30.1 represent a new betasatellite, for which the name Sida corlifolia yellow mosaic betasatellite – [South Africa:Mooketsi:2008] (SiCYMB-[ZA:Mkt:08]) is proposed. Phylogenetic analysis grouped SiCYMB-[ZA:Mkt:08] with other CHaYMB isolates reported (Figure 4.4). Interestingly, the clade formed by the SiCYMB and ChaYMB isolates are sister to the clade formed by the ToLCTGB isolates and this close phylogenetic relationship is perhaps reflecting their close geographic relationship.

Figure 4.5. Plant sample 30 was collected in the Mooketsi Valley (Limpopo) in 2011 and was identified as *Sida corlifolia* by DNA barcoding. The plant showed bright yellow mosaic symptoms on its upper leaves and was found to contain an isolate of the bipartite virus, *Sida corlifolia golden mosaic virus* - [South Africa:Mooketsi:2014] (SiCYMV-[ZA:Mkt:2014] (4.3.4), as well as *Sida corlifolia yellow mosaic betasatellite* (4.3.3).

4.3.4 Detection of bipartite begomoviral genomes in individual plant samples

Sample 242 was collected in Cato ridge in KwaZulu-Natal summer 2015. Sample 242 was identified as *Phaseolus vulgaris* (bean) by DNA barcoding, as described in section 4.3.6 below. Two viral contigs (~2.8 kb) were recovered from sample 242 MiSeq dataset and were found to represent a bipartite begomovirus, since the one genomic component showed a genome organization typical of a bipartite begomovirus DNA-A (five ORFs: V1, C1, C2, C3 and C4 and a 171 nt intergenic region (IR) containing the invariant nonanucleotide motif (TAATATTAC) and the second genomic component, showed a typical DNA-B genome organization (two major ORFs: BC1 on the complementary- sense strand and BV1 on the virion-sense strand and a 230 nt common region (CR) shared with the DNA-A within the IR). Most importantly, the DNA-A molecule lacked the AV2 ORF that is typically present in Old World begomoviruses (Stanley et al. 2005). The complete nucleotide sequence of the begomovirus isolate recovered from the NGS data was determined to be 2715 bp and 2811 bp in size, for the DNA-A and DNA-B respectively. Nucleotide-sequence comparisons with the closest relatives downloaded from NCBI, showed that both the DNA-A and DNA-B components shared the highest nucleotide sequence identity (70.3% and 33%) to the DNA-A and DNA-B component of an isolate of SbCBV, respectively (Table 4.4). Because the DNA-B component shared such low sequence similarity across its full genome

sequence, the BV1 and BC1 ORFs were also investigated and found to share a higher sequence similarity at both the nucleotide and amino acid level, with the respective ORF's from the DNA-B component of SbCBV (37,3 – 57,6%) (Table 4.5). According to the begomovirus species demarcation threshold of 91% (Brown et al. 2015), the bipartite begomovirus found in S242 represents a new begomovirus species for which we propose the name Phaseolus vulgaris begomovirus, pending further sequence confirmation and investigation to determine its biological significance (host range, symptomology, transmission vector). Phylogenetic analysis grouped *Phaseolus vulgaris begomovirus* DNA-A in a cluster with other bipartite viruses, and specifically in a cluster with *Soybean chlorotic blotch virus* (Figure 4.6).

Table 4.4 Percentage identities for complete genome and selected open reading frame nucleotide and amino acid sequences of *Paseolus vulgarus* begomovirus [ZA:CaR:2015] DNA-A with selected begomoviruses.

Begomoviruses*	**Genome**	**V1**		**C1**	
		nt	**aa**	**nt**	**aa**
Soybean chlorotic blotch virus	70.3	69,7	76,8	72,8	75,8
Hedyotis uncinella yellow mosaic virus	68,28	69	69,9	71,7	71,5
Tomato leaf curl Ghana virus	67,3	67,3	67	70	72

Table 4.5 Percentage identities for complete genome and selected open reading frame nucleotide and amino acid sequences of *Paseolus vulgarus* begomovirus [ZA:CaR:2015] DNA-B with selected begomoviruses.

Begomoviruses*	**Genome**	**BV1**		**BC1**	
		nt	**aa**	**nt**	**aa**
Soybean chlorotic blotch virus	33	37,3	53,7	50,3	57,6
East African cassava mosaic virus	29	38,8	36,9	51,4	57,6
East African cassava mosaic Cameroon virus	29	38,6	37,7	50,2	55,3

Sample 30 was collected in the Mooketsi (Limpopo) in 2011. This plant sample was identified as *Sida corlifolia* by DNA barcoding, as described in section 4.1.3.6 below. Some of the plant leaves, especially the apical leaves, had a bright yellow and green mosaic coloration, but no stunting or cupping was observed (Fig 4.5). Two begomoviral contigs (~2.8 kb) were recovered from the sample 30 MiSeq dataset and was found to represent a bipartite begomovirus, since the one genomic component showed a genome organization typical of a bipartite begomovirus DNA-A (Five ORFs: V1, C1, C2, C3 and C4 and a 176 nt intergenic region (IR) containing the invariant nonanucleotide motif (TAATATTAC) and the second genomic component, showed a typical DNA-B genome organization (two major ORFs: BC1 on the complementary- sense strand and BV1 on the virion-sense strand and a 169 nt common region (CR) shared with the DNA-A within the IR). The DNA-A molecule also lacked the AV2 ORF that is typically present in Old World begomoviruses (Stanley et al. 2005). Each CR contained two identical iterons upstream of the AC1 TATA box, as well as identical stem–loop sequences that included the conserved TAATATTAC nonanucleotide sequence. The complete nucleotide sequence of the begomovirus isolate recovered from the NGS data was determined to be 2658 bp and 2644 bp in size, for the DNA-A and DNA-B respectively. Nucleotide-sequence comparisons with the closest relatives downloaded from NCBI showed that the DNA-A component shared the highest nucleotide sequence identity (72%) to an isolate of CoYVV (KX101212) (Table 4.6), whereas the full genome of the DNA-B molecule had no significant similarity when the full genome was compared using BLAST. When the individual DNA-B ORFs were compared, the BC1 and BV1 showed the highest nucleotide sequence identity (62,3% and 49,6%) to the corresponding ORFs from an isolate of *Corchorus yellow vein virus* DNA-B (CoYVV) (KX101212) (Table 4.7). Similar to CoYVV, the DNA-A component lacked a AV2 ORF, but contained the PWRTNAGT motif in the N-terminus of the deduced CP sequence encoded by AV1 of CoYVV (Harrison et al., 2002). According to the begomovirus species demarcation threshold of 91% for a DNA-A component (Brown et al. 2015), the bipartite begomovirus found in this sample therefore represent a new bipartite begomovirus species, for which we provisionally propose the name Sida corlifolia golden mosaic virus - [South Africa:Mooketsi:2014] (SiCYMV-[ZA:Mkt:2014], according to the host species and symptoms observed in the sampled plant and pending further investigation to determine its biological significance (host range, symptomology, transmission vector). Phylogenetic analysis grouped SiCYMV-[ZA:Mkt:2014] DNA-A within in a basal position to the cluster containing CoYVV, *Corchorus yellow vein Vietnam virus* and CoGMV. Interestingly, this group of viruses cluster together with other bipartite viruses from the OW, separate from related bipartite viruses found in the NW (Fig 4.6).

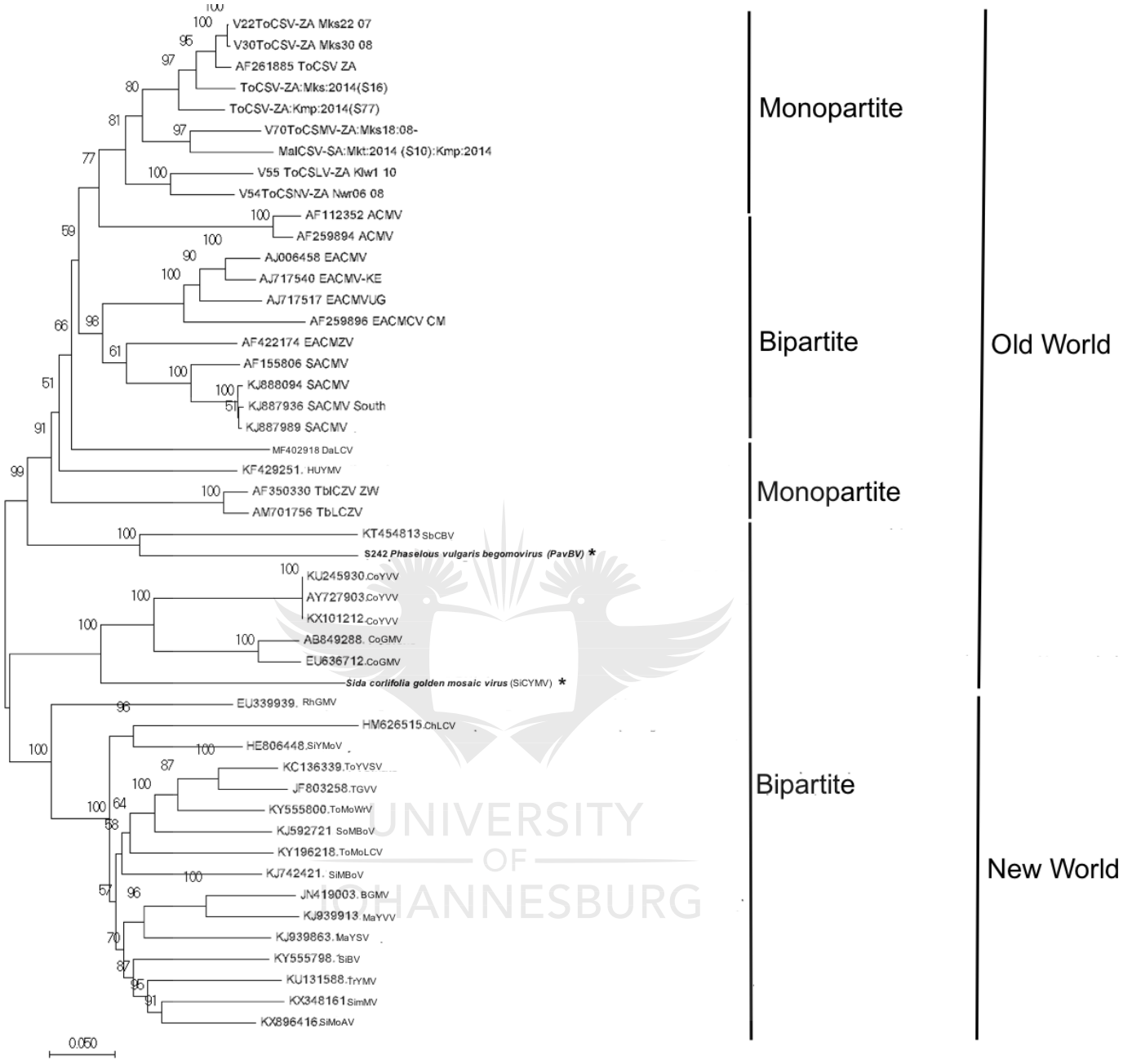


Figure 4.6. Phylogenetic tree showing the relationships between the DNA-A component of two newly identified South African bipartite viruses (in bold, with *) and begomoviruses from around the world. The tree was inferred using the Neighbor-Joining method (Saitou and Nei, 1987) using the MEGA 7 program (Kumar et al. 2016). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (2000 replicates) are shown next to the branches (Felsenstein, 1985). The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Maximum Composite Likelihood method (Tamura et al, 2004) and are in the units of the number of base substitutions per site. Betasatellites used for the phylogenetic tree are indicated by NCBI accession numbers in the figure.

Table 4.6 Percentage identities for complete genome and selected open reading frame nucleotide and amino acid sequences of ***Sida corlifolia golden mosaic virus*** [ZA:Mks:08] (SiCYMV) **DNA-B** with selected begomoviruses.

Begomoviruses*	**Genome**	**BV1**		**BC1**	
		nt	**aa**	**nt**	**aa**
Corchorus golden mosaic virus	41	48	53	53	62
Corchorus yellow vein virus	40	49.6	52,7	57,7	62,3
Euphorbia mosaic virus	37	46	50,8	54,3	59

Table 4.7 Percentage identities for complete genome and selected open reading frame nucleotide and amino acid sequences of ***Sida corlifolia golden mosaic virus*** [ZA:Mks:08] (SiCYMV) **DNA-A** with selected begomoviruses.

Begomoviruses*	**Genome**	**V1**		**C1**	
		nt	**aa**	**nt**	**aa**
Corchorus yellow vein virus	60.49	62	71,3	75	78
Corchorus golden mosaic virus	58,2	61	69,3	74	77

4.3.5 Detection of a genomovirus genome in individual plant samples.

Sample 8 was collected in Mooketsi Valley (Limpopo) in February 2014. This plant sample was identified as *Cynoglossum officinale* by DNA barcoding, as described in section 4.1.3.6 below. One viral contigs (2182 nt) were recovered from the sample 8 MiSeq dataset (An identical contig was also recovered from the pool-1 dataset). Blast analysis of the assembled sequence contig revealed similarities (76% identity; 68% coverage) to the sequence of *Bemisia*-associated genomovirus AdDF (KY230613), a virus from the family *Genomoviridae*, genus *Gemybolavirus* (Varsani and Kuprovic, 2017). The putative ORFs of this sequence were detected using ORF finder (http;//www.ncbi.nlm.hih.gov/projects/gorf/.). Like other genomoviruses, the recovered viral contig had two large ORFs, one in the virion-sense and one in the complementary-sense strand (Fig4.7), as well as an origin of replication (TAATATTAT) similar to geminiviruses, and has an intergenic region (IR) comprising 193 nt. The proteins encoded by the two ORFs share significant amino acid sequence identity with putative capsid (CP) and Rep proteins of *Genomoviridae* like

viruses. The first ORF (sense) is 864 nt-long and encodes a 287 aa-long putative coat protein (CP) (Fig 4.7) with 100% coverage and 57% aa identity with the CP from *Bemisia*-associated genomovirus AdDF (KY230613) (JX185428, 7e-90). The complementary strand ORF is 1122 nt-long, with a putative intron of 170 nt and encodes a Rep protein that is 316 aa long and shared 68% aa identity with the Rep from *Dragonfly associated gemyduguivirus-1* (formerly known as *Dragonfly-associated circular virus-3*) (JX185428) (39% coverage, 9e-53) (Table 4.3). All typical genomoviridae amino acid motifs were identified in the predicted aa sequence of this ORF: motif I (LLTYAQ), motif II (THYHA), GRS domain (RIFDIDGYHPNILRGR), motif III (YACK), Walker A (GDSRTGKT), Walker B (VFDDM), and motif C (WCNN). Based on the analysis of distribution of the pairwise identities across genomes, CPs and Reps, a threshold of 78% for species demarcation has previously been proposed (Varsani and Kuprovic, 2017). Despite the relative low overall genome % identity with other known viruses, the viral genome recovered from *Cynoglossum officinale*, has a typical genome organization and genetic relationship to other genomoviruses, indicating it should be considered new members of this family. The virus described here is proposed as ‘‘*Cynoglossum offinale* associated circular DNA virus (COasCV)”. Phylogenetic analysis of the deduced amino acid sequence of Rep for the new virus was performed after alignment with representative genomovirus-like sequences (Fig 4.8). The COasCV genomovirus is closely related to the majority of gemycircularvirus-like sequences, including the type species *Sclerotinia gemycircularvirus* 1. The COasCV genome clustered close to *Bemisia-associated genomovirus* AdDF (KY230613) and *Dragonfly associated gemyduguivirus* 1, but clearly in a separate branch, suggesting that it may be classified in a distinct genus.

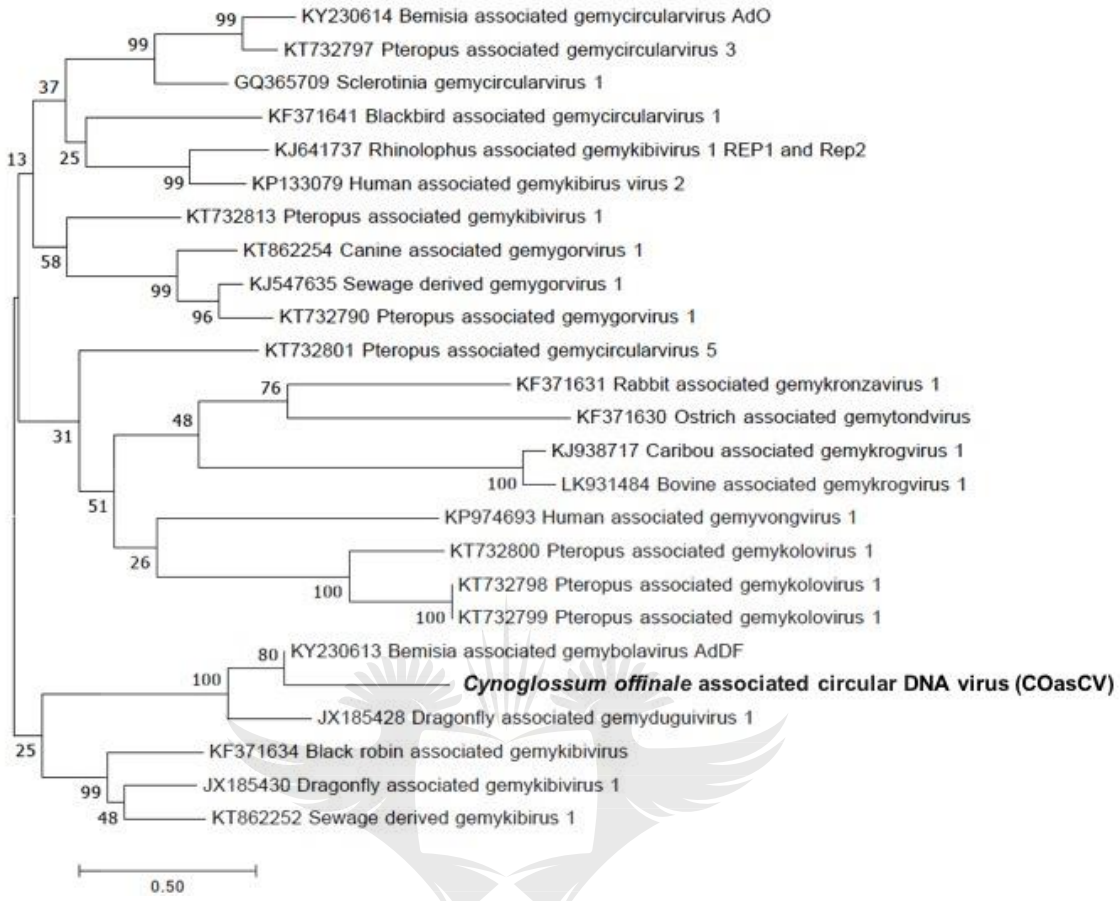


Figure 4.7. Phylogenetic analysis by Maximum Likelihood method of replication-associated proteins from the genomovirus identified in sample 8 (in bold) and representative sequences of genomoviruses with their new designations, as according to the International Committee on Taxonomy of Viruses (ICTV). Genera are also indicated, according to ICTV. Bootstrap values above 50% are shown at the branch nodes (1,000 replications). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. Bar: substitutions per site. Evolutionary analyses were conducted in MEGA7.

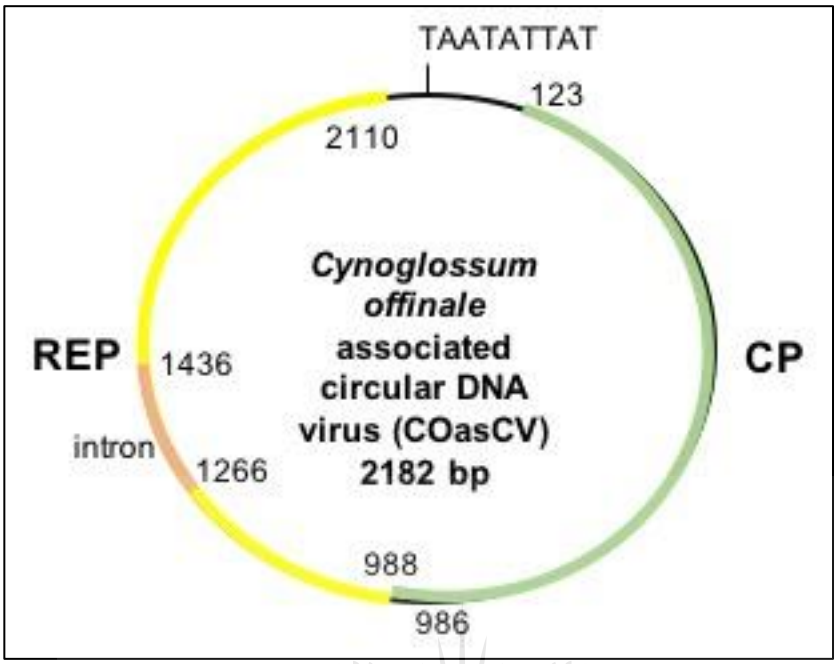


Figure 4.8. Diagram depicting the proposed genome organization of *Cynoglossum offinale* associated circular DNA virus (COasCV) identified in this study. The two open reading frames, encoding the putative CP (sense) and Rep (antisense) proteins, intron and ori are indicated. **CP**: Coat protein; **Rep**: replication-associated protein; **ori**: origin of replication.

4.3.6 Plant identification by DNA Barcoding

In this study, 230 plant samples were collected, at various growth stages and screened for DNA-viral infection. All the plants that tested positive in the viral screening and identification procedure outlined above, were further identified to the genus and species level. Identification of the plant species was firstly performed morphologically, using the prepared herbarium specimens, and secondly, the identity was confirmed molecularly by DNA barcoding, using the chloroplast *rbcLa* and *matK* genes. PCR amplification of both the rbcl and *matK* gene was generally successful with all the candidate loci. While only one primer set was used to amplify the *rbcL* locus, four different primer sets were used to amplify the *matK* sequences across all the species tested in this study. PCR success rate varied across different loci and weed species. The *rbcL* was the only locus successfully amplified in all tested weed species using only one primer set. High rates of PCR success were also apparent at *matK.* The sequences generated were then compared to reference barcode database constructed on BOLD (Barcode of Life Data Systems) and BLAST (Basic Alignment Search Tool). BLAST and recently developed BOLD searches were applied to the produced sequences using the available online databases. The results from BLAST and BOLD queries are summarized in Table 4.8. The species level identification for all samples were in agreement between the two databases used (BLAST and BOLD), and were also in agreement

with the morphological identification provided by examining the prepared Herbarium specimens for each sample.

Table 4.8. Voucher name, Barcoding identification, E-value and Max ID of all weed samples infected with begomoviruses.

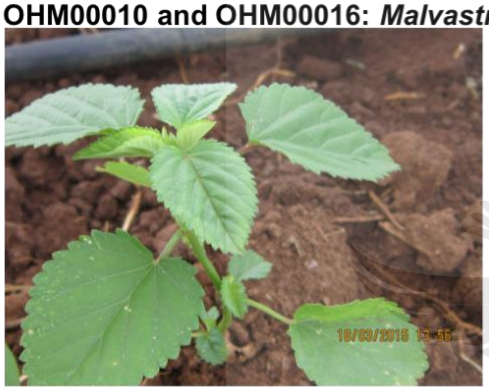
	rbcLa			matK		
Voucher name	**Barcoding identification**	**E value**	**max ID**	**Barcoding identification**	**E value**	**Max ID**
OHM0008	*Cynoglossum officinale*	0	98.18%	*Cynoglossium officinale*	0	96.63%
OHM0010	*Malvastrum coromendalianum*	0	98.88%	*Malvastrum coromendalianum*	0	99.30%
OHM0016	*Malvastrum coromendalianum*	0	99.54%	*Malvastrum coromendalianum*	0	99.26%
OHM0018	*Abutilon indicum*	0	99.58%	*Abutilon indicum*	0	100%
OHM0066	*Solanum lycopersicum*	0	100%	*Solanum lycopersicum*	0	100%
OHM0069	*Solanum lycopersicum*	0	100%	*Solanum lycopersicum*	0	100%
OHM0077	*Acalypha indica*	0	99.28%	*Acalypha indica*	0	98.18%
OHM0242	*Phaseolus vulgaris*	0	99.81%	*Phaseolus vulgaris*	0	96.28%
OHM0030	*Sida cordifolia*		99.71%	*Sida cordifolia*	0	96.82%

OHM0008: *Cynoglossum officinale*

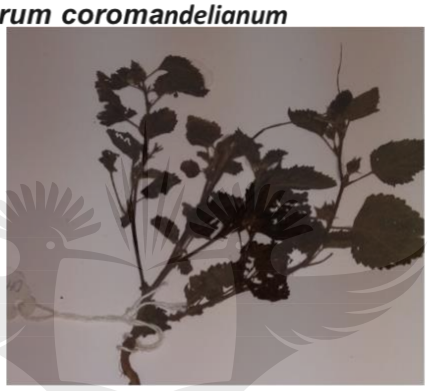
Herbarium specimen

Cynoglossum officinale 190506a.jpg

OHM00010 and OHM00016: *Malvastrum coromandelianum*



Field sample



Herbarium specimen

By Franz Xaver
https://commons.wikimedia.org/w/index.php?curid=28868737

Figure 4.9. Photographs depicting the sampled indigenous plants / weeds (OHM0008 and OHM00010 and OHM00016) at the time of collection (field sample), the prepared herbarium specimens or as a fully grown, or mature plants showing identifiable characteristics (pictures sourced from the world wide web as indicated on the diagram). Sample number and barcoding results (plant identification) are indicated at the top of each picture.

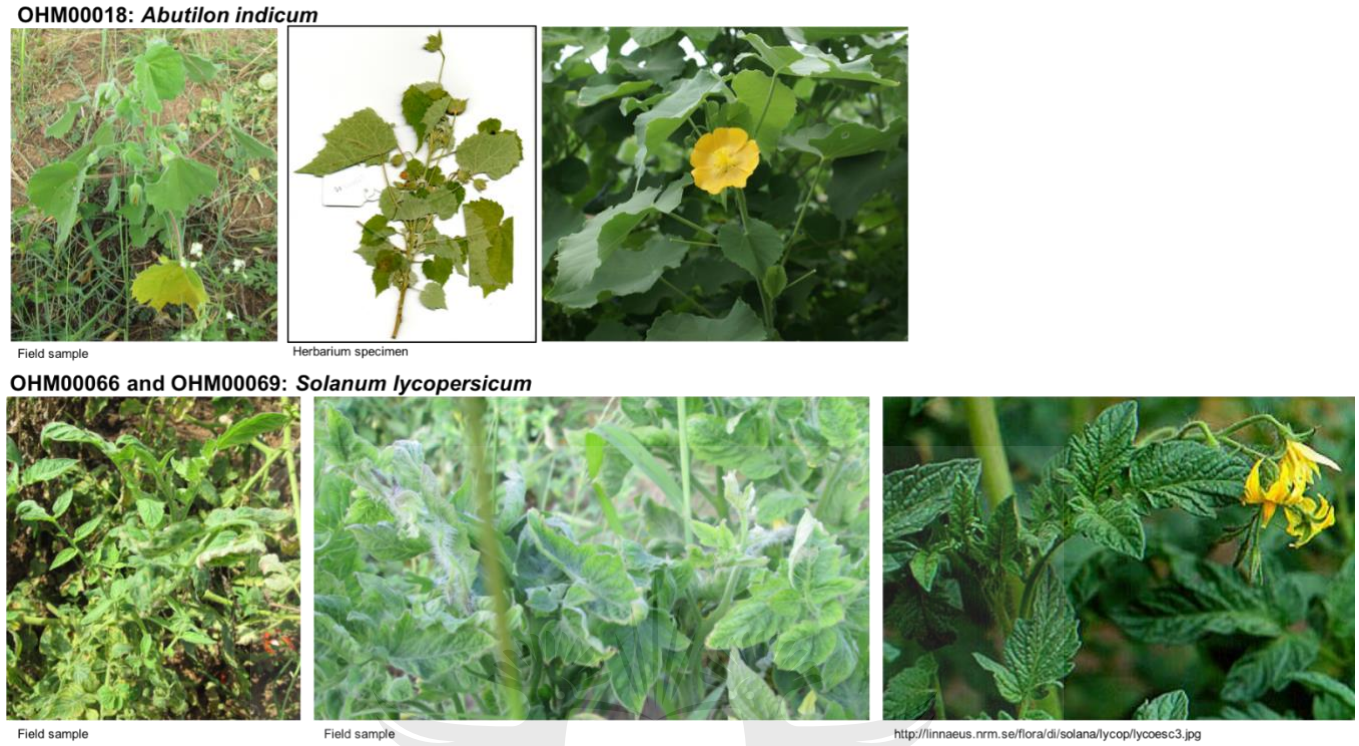


Figure 4.10. Photographs depicting the sampled indigenous plants / weeds (OHM00018, OHM00066, OHM00069, OHM242) at the time of collection (field sample), the prepared herbarium specimens or as a fully grown, mature plants showing identifiable characteristics (pictures sourced from the world wide web as indicated on the diagram). Sample number and barcoding results (plant identification) are indicated at the top of each picture.

Figure 4.11. Photographs depicting the sampled indigenous plants / weeds (OHM242 and OHM0077) at the time of collection (field sample), the prepared herbarium specimens or as a fully grown, mature plants showing identifiable characteristics (pictures sourced from the world wide web as indicated on the diagram). Sample number and barcoding results (plant identification) are indicated at the top of each picture.

CHAPTER 5

Discussion

With the aim to determine the (DNA) viral diversity infecting indigenous or non-cultivated plant species (weeds), a survey was carried out between 2014 and 2015 in three Provinces (Limpopo, Mphumalanga and KwaZulu-Natal) of South Africa. Samples were collected from a total of 230 plants, growing alongside cultivated crops and screened for viral infection. These plant samples were collected in and around vegetable crops, such as tomatoes, cucurbits, sweet potatoes, peppers, beans, which according to literature, are known to become infected by Geminiviruses. Fields from which samples were collected, were also specifically chosen, where high numbers of whiteflies (*B. tabaci*) were observed, to increase the likelihood of finding whitefly-transmitted begomoviruses. Plants were mostly sampled at random and not only if they showed symptoms of possible viral infection (i.e chlorosis, curling, stunting, etc), weed plants found to harbour *B. tabaci*, were preferentially sampled.

The first step in the screening process, involved total DNA isolation from each plant sampled. This total DNA served both as a source of genomic plant DNA for plant identification by DNA barcoding, and also as DNA source to screen the plants for the presence of any circular DNA viruses, particularly Geminiviruses that include whitefly transmitted-begomoviruses. The viral screening strategy used in this study, involved the initial screening of the plant samples by PCR, using universal/degenerate begomovirus primers (TY1/2; Accotto et al. 2000) that targets begomovirus coat protein (CP), a highly conserved protein within this family and generally used for provisional begomovirus identification (Brown, 2000). It was found that 24% of the non-cultivated or indigenous plants sampled tested positive for possible begomovirus infection. This is not surprising, as begomoviruses are vectored by *B. tabaci*, a whitefly vector that has a very wide host range and feeds on an estimated 600 plant species (Berlinger, 1986). The presence of the broadly polyphagous Middle East - Asia minor 1 (MEAM1) cryptic species of *B. tabaci*, or more commonly known B biotype in South Africa, and specifically in these three provinces, has previously been shown (Esterhuizen et al. 2012a), and would vector these begomoviruses both amongst indigenous plant species, and between cultivated crop plants and indigenous plants growing in cropping areas. It has previously been shown around the world (Bedford et al. 1998; Ambrozevicius et al. 2002; García-Andrés, *et al.* 2006, Varsani et al. 2008; Azhar et al. 2011), and also in South Africa, that tomato infecting begomoviruses are commonly detected in weeds growing in and around the cultivated areas (Esterhuizen et al. 2012b).

Taking into consideration that the above mentioned initial screening by PCR would only reveal the presence of begomoviruses, the full viral diversity of these samples were further explored using RCA, RCA/RFLP and the results extended by NGS. RCA is a method which enables unbiased, full length viral genome amplification, of any circular DNA viruses and their respective

satellites, without any prior information about the virus sequences and as such, will allow identification of novel circular DNA viruses or satellites. The use of the RCA technique, in combination with RFLP in this study, was found to be a rapid, reproducible method to screen a large number of plant samples for viral biodiversity. In this study, a total of 45 weed samples were TY negative, but provided an unique RCA/RFLP profile, that were then further explored by NGS. The one drawback of this strategy, to screen plants for novel viral infection, is the fact that the RCA reaction amplifies any circular DNA, including bacterial or plant mitochondrial or plastid DNA and therefore provide a lot of 'false positives', i.e. samples that provide an RCA/RFLP profile but are not infected by circular DNA viruses. For example, a large number of plants were included in the 2015 NGS experiment, but although these plants provided an RCA/RFLP profile, no contigs of viral origin could be detected upon data analysis and the majority of the NGS reads obtained were from a plant origin. This should be kept in mind, as the high cost of NGS sequencing on individual plant samples can be prohibitive to using this strategy for viral discovery. Indeed, to overcome this drawback, and reduce the cost of NGS experiments, some RCA samples in this study were pooled and sequenced in one NGS run (NGS 2014, Pool-1 and 2). It was found that this method is still an effective method for viral discovery, as numerous viral contigs were detected in both Pool-1 and 2; the only drawback with the pooling-strategy was that for most viral contigs, the full viral genome could not be assembled (can be overcome by increasing the sequencing depth) and, post-NGS, it requires a lengthy follow up procedure, to trace the original plant samples, where the detected viral molecules originated from. Nevertheless, the RCA/RFLP/NGS strategy did drove useful, as it did allow the detection and characterization of both previously known viruses (four strains or variants of ToCSV), and previously unknown, novel virus species/molecules (one monopartite and two bipartite begomoviruses and two betasatellite molecules) (Table 4.2). Lastly, RCA allowed for the direct sequencing (NGS or sanger sequencing) of the viral genomes without cloning and provided high fidelity results with the same quality as conventional cloning and sequencing, but with significantly reduced effort and costs.

Samples collected in 2014, were screened by TY PCR, RCA/RFLP and then divided and pooled into two groups for characterization by NGS, i.e. pool-1, that contained 13 plant samples that provided a positive TY PCR result and a RCA/RFLP profile predictive of possible DNA viral molecules, and pool-2, that contained 22 plant samples that provided a negative TY PCR result but did provide a RCA/RFLP profile. Data analysis of the NGS data set generated from the pooled RCA products from pool-1 and 2, identified 29 contigs with blast similarity to virus-associated genomes (Table 4.2). De novo sequence assembly from the combined data sets, allowed re-assembly of seven begomovirus genomes, two genomovirus genomes, one aphasatellite and one betasatellite, but for none of these detected viral molecules, could the complete/full genome sequence be recovered (69 – 99% of full genome recovered). This is one of the drawbacks of

using a pooled-sample strategy for NGS, as the sequencing depth has to be sufficiently high to allow complete sequencing and coverage of each of the viral genomes. Another problem to the strategy of genome assembly from the combined data set, is the ability for the computational algorithm to correctly allocate individual viral reads from the combined data set, to the viral genome that they originated from, particularly when two or more viral genomes in the pooled sample have very high sequence similarity (Simmonds et al. 2017). When two or more viral genomes have high sequence similarity, some of the viral reads will be mis-assigned and a chimeric viral genome will be constructed during De novo assembly, that did not exist in the original plants sampled. Therefore, when using NGS for viral discovery, one always has to confirm the resultant viral genomes, by going back to the original samples, amplifying the viral genomes using PCR, followed by sequencing using primer walking and Sanger sequencing. In this study, this was found to be a difficult, time-consuming and a costly exercise. For the genomes recovered from the pooled samples, six begomovirus genomes were retraced to their original plant samples, the two genomovirus genomes were traced to two plant samples and two betasatellite genomes were retraced to two original plant samples (and connected with its associated helper virus, ie. sample 10 and 30), but both the alpha-satellite and the circular DNA molecules associated with CGMV-China, could not be retraced.

Previous studies investigating tomato-infecting begomoviruses in South Africa have indicated the widespread presence, and indeed the dominance of ToCSV throughout tomato cropping regions in South Africa, and also partially in the areas surveyed in this study (Limpopo, Mpumalanga and KwaZulu-Natal provinces) (Pietersen et al. 2008; Esterhuizen et al. 2012b; Moodley et al. 2018). This begomovirus has also been shown to have a wide host range, infecting various weed species, in addition to tomato and bean crops (Pietersen et al. 2008; Esterhuizen et al. 2012b). It is therefore not surprising that four ToCSV genomes were recovered in this study (Table 4.3), two from tomato plants and two from weeds species, i.e *Malvastrum coromandelianum* and *Acalypha indica*. The discovery of ToCSV in these two weed hosts represents a further new natural host report for this virus, in addition to the host reported before (*S. lycopersicum, Phaseolus vulgaris* cultivars, *Datura stramonium*, *D. ferox,* several *Nicotiana* species, *Cleome spp., C. carinatum, Amaranthus hybridus*, *Alternanthera pungens, Sida cordifolia, S. rhombifolia* (Pietersen et al. 2008; Esterhuizen et al. 2012b). These weed hosts are found widespread in South Africa, often surrounding areas of crop cultivation, and should be targeted for removal as they are acting as begomovirus reservoirs, maintaining these virus in the agro-ecosystem, and this may lead to the generation of new species/strains by recombination and component exchange. The four monopartite ToCSV isolates recovered from sample 66, 69, 16 and 77, shared between 92-96% nucleotide identity with ToCSV-[ZA-Ond-98] (AF261885), and are therefore considered as different isolates (Sample 16, 66 and 69; 94-99% nucleotide identity), or strain (sample 77, 92%

nucleotide identity) of ToCSV-ZA (Table 4.3). As expected, phylogenetic analysis grouped the four ToCSV isolates within the SAI subclade, along with all other ToCSV isolates, reflecting their origin in southern Africa. Interestingly, the two ToCSV genomes recovered from weed species, *M. coromandelianum* and *A. indica* formed a separate subgroup, apart from the tomato-infecting ToCSV isolates (Figure 4.3), perhaps reflecting the genetic adaption of these two viral species to the different host species they were found in.

One plant sample, collected in January 2014 in the Mooketsi Valley (Limpopo) tested positive for TY- PCR and produced and RCA/RFLP profile indicative of a monopartite begomovirus (~2.7kbp), with extra bands possibly indicating the presence of a satellite molecule. The RCA-DNA from this sample was sequenced as part of the pooled-1 samples and then later, sequenced again as sample 10, in the individual NGS sequencing run in 2015. From the combined pool-1 data set, one viral contig (1.19, Table 4.2) of 2618 bp was recovered, with high sequence similarity (88%) to *Tomato curly stunt Mooketsi virus*-[ZA:Mooketsi:2007] (ToCSMV-[ZA:Mks:07]. Upon sequencing of sample 10 in the individual sequencing run in 2015, the full viral genome of this viral isolate was recovered (2767 bp), and was found to represent a begomovirus, with typical begomovirus-DNA-A genome organization (Fig 4.2). Based on the SDT pairwise alignment, it was found to have 88,4% sequence similarity to ToCSMV-[ZA:Mks:07], and according to the begomovirus species demarcation threshold of 91% (Brown et al. 2015), the begomovirus found in sample 10 represent an a new begomovirus species. The sample 10 host plant were subsequently identified by DNA barcoding (discussed below) as *Malvastrum coromandelianum*. Consequently, the name Malvastrum curly stunt virus - [South Africa:Mooketsi:2014] (MalCSV-[ZA:Mkt:2014], was proposed for this newly identified begomovirus (Table 4.3). Based on current evidence, i.e. typical monopartite genome organization, absence of a DNA-B genome in either the pool-1 NGS dataset (2014), or the sample 10 dataset (2015) and the inability to detect a DNA-B component by PCR, MalCSV is most likely a true monopartite begomovirus, but this must be further investigated.

Phylogenetic analysis grouped MalCSV-[ZA:Mkt:2014] within the SAII subclade, within the larger African/ South West Indian Ocean (SWIO) cluster, along with two isolates of ToCSMV (Fig 4.3), suggesting that this virus is most probably indigenous to southern Africa and has not been introduced from other regions. The high sequence similarity and close phylogenetic relationship between the tomato-infecting ToCSMV and the weed-infecting MalCSV virus, further supports the idea that weed species harbour diverse virus species, that supports mixed infections that lead to recombination, diversity, and subsequent evolution of these viruses. (Bedford et al*.* 1998; Ambrozevicius et al*.* 2002; García-Andrés, et al*.* 2006, Varsani et al*.* 2008; Azhar et al*.* 2011). *Malvastrum coromandelianum* is reported to have originated from tropical America and is

therefore considered as an invasive species in South Africa (www.invasives.org.za) and is reported to be particularly problematic in cropping areas in Limpopo, Mpumalanga, Eastern Cape and Gauteng Provinces. *Malvastrum coromandelianum* plants have also previously been reported as host for begomoviruses, particularly in China. In addition to the isolates of MalCSV and a ToCSV found in this host in this study, *M. coromandelianum* has previously been reported as host for three other begomoviruses, including *Tomato yellow leaf curl China virus* (TYLCCNV), *Malvastrum yellow vein virus* (MaYVV) and *Malvastrum yellow vein Yunnan virus* (MaYVYV) (Zhou et al. 2003; Jiang & Zhou, 2005; Liu et al. 2009). Most of these viruses were also associated with a betasatellite molecule (Zhou et al. 2003; Jiang and Zhou, 2005; Liu et al. 2009). This was also found to be the case for MalCSV-[ZA:Mkt:2014] found in sample 10, as will be discussed below.

The RCA/RFLP profile for sample 10 suggested the presence of an associated satellite molecule and indeed, a beta-satellite molecule was detected in both the pooled-1 combined data set and sample 10 NGS run in 2015. In the pool-1 data set, only 99% of the full genome was recovered (1308 bp), but in the sample 10 dataset, the full genome (1316 bp) of a begomovirus-assocated beta-satellite molecule was obtained upon de novo assembly. Based on SDT pairwise alignment, the beta-satellite shared 78% nucleotide sequence identity to *Tomato leaf curl Togo betasatellite* (ToLCTGB) (KT382329), found in a tomato plant in 2006 in Togo (Kon and Gilbertson, 2012). According to the 78% beta-satellite species demarcation threshold (Briddon et al. 2008), the beta-satellite found in sample 10 represent an isolate of *Tomato leaf curl Togo betasatellite* (ToLCTGB), for which the addition of the following designation, ToLCTGB - [South Africa:Mooketsi:2014] (ToLCTGB-[ZA:Mkt:2014]), was suggested. Phylogenetic analysis confirmed that ToLCTGB-[ZA:Mkt:2014] formed a cluster with other ToLCTGB sequences from Africa, more specifically from West Africa (Ghana and Togo) (Fig 4.4). The association between the newly identified, weed infecting begomovirus MalCSV-[ZA:Mkt:2014] and the beta-satellite ToLCTGB-[ZA:Mkt:2014], previously found in tomato plants, but here, isolated from a weed, should be further investigated. As previously reported, beta-satellites require the presence of the helper begomovirus for replication, insect transmission, encapsidation, movement in plants and the satellite molecule can play a role in symptom development (Idris et al. 2011; Mubin et al. 2010). In this study, no typical begomovirus symptoms were observed on the plant sample 10, but this could be due to the plant growth stage upon sampling and should be further investigated by large scale sampling of this plant species in that area of the Limpopo Province. Furthermore, it has previously been shown that ToLCTGB is not required for infectivity or symptom induction by its helper begomovirus in tomato plants, but that the satellite does increase disease severity, in a host-dependent manner (increased symptom severity in *Nicotiana* and D*. stramonium*) (Kon and Gilbertson, 2012).

One weed plant sample (sample 30), collected in 2011 in the Mooketsi Valley (Limpopo) tested positive for TY-PCR and produced and RCA/RFLP profile indicative of a bipartite begomovirus (bands adding up to ~5,6kbp), with extra bands possibly indicating the presence of a satellite molecule. The RCA-DNA from this sample was sequenced in the individual NGS sequencing run in 2015 and after de novo assembly, two begomoviral contigs (~2.8 kb) were recovered, found after data analysis to represent a DNA-A and a DNA-B begomoviral genomic component. Based on the SDT pairwise alignment of these bipartite genomic components, the DNA-A molecule was found to have 72% sequence similarity to the DNA-A component of *Corchorus yellow vein virus* (KX101212) (Table 4.6) and the individual DNA-B ORFs (BC1 and BV1) showed the highest nucleotide sequence identity (62,3% and 49,6%) to the corresponding ORFs from *Corchorus yellow vein virus* DNA-B (CoYVV) (KX101212) (Table 4.7). According to the begomovirus species demarcation threshold of 91% for a DNA-A component (Brown et al. 2015), the begomovirus found in sample 30 represent a new bipartite begomovirus species. The sample 30 plant was identified by DNA barcoding as *Sida corlifolia*, and most importantly, upon collection, this weed sample was found to display symptoms typically associated with viral infection, i.e. bright yellow and green mosaic coloration of the apical leaves (Fig 4.5). Based on this observation, the name Sida corlifolia golden mosaic virus - [South Africa:Mooketsi:2014] (SiCGMV-[ZA:Mkt:2014] was proposed for this newly identified bipartite begomovirus. Further sampling of this host species should be carried out, particularly in the Limpopo province, to determine the extent to which *Sida* plants are infected by this viral species. Furthermore, future research should include the construction of infectious clones of the DNA-A and DNA-B components, to determine the biological significance (host range, symptomology, transmission vector) of this bipartite begomovirus, and also to confirm that the SiCGMV DNA-A and DNA-B components represent a biologically functional unit from the same begomovirus (i.e. that the SiCGMV DNA-A molecule support episomal replication of the DNA-B molecule).

Phylogenetic analysis of the DNA-A component of SiCGMV-[ZA:Mkt:2014] clustered SiCGMV in a basal position with two isolates of so called Corchoviruses, namely *Corchorus yellow vein Vietnam virus* (CoYVV),and *Corchorus golden mosaic virus* (CoGMV) reported from India and Vietnam. These two Chorcoviruses form an interesting subgroup of whitefly transmitted viruses, similar to Legumoviruses (range of legume-infecting viruses from India and Southeast Asia) and Sweepoviruses (viruses isolated from I*pomoea* species, originating from America, Asia and Europe). These three subgroups of viruses have previously been shown to be genetically distinct, and basal to all other begomoviruses, likely due to distinct evolutionary histories or genetic isolation in their host species (Briddon et al. 2010). When comparing their genetic features, typical begomoviruses have typically been subdivided into NW and OW members, where NW members

have a bipartite genome, with both components needed for infectivity (Stanley et al. 2005) and the majority of OW begomoviruses have monopartite genomes, and most of these interact with ssDNA satellites (alpha and betasatellites) (Briddon et al. 2000; Saunders et al. 2000; Zhou et al. 2003). In addition, all OW begomoviruses have an extra AV2 ORF in DNA-A that is not present in NW begomoviruses (Rybicki, 1994; Stanley et al. 2005) and NW begomoviruses have an N-terminal PWRsMaGT motif in the coat protein (CP) encoded by AV1, which is absent from OW begomoviruses (Harrison et al. 2002). Lastly, in most OW begomoviruses, there are two iterons upstream of the AC1 TATA box, with a complementary iteron downstream, but the downstream iteron is lacking in most NW begomoviruses (Arguello-Astorga et al. 1994). In this study, these unique 'NW-like' genetic features, i.e. absence of a AV2 ORF and presence of PWRTNAGT motif in AV1, have also been found in the DNA-A component of SiCGMV-[ZA:Mkt:2014], similar to CoYVV and CoGMV. The identification of another OW virus (SiCGMV) with NW-like features, this time in Africa, in addition to previously identified Corchoviruses from the East, is surprising and it also gives additional support to the idea that NW begomoviruses may have originated in the OW and were subsequently disseminated to the New World, and that Corchoviruses may be a remnant of such New World begomoviruses that once populated the OW (Ha et al. 2008; Hassan et al. 2015).

The RCA/RFLP profile for sample 30 suggested the presence of an associated satellite molecule and indeed, a beta-satellite molecule was detected in sample 30 NGS run in 2015. De novo assembly recovered the full genome (1342 bp) of a begomovirus assocated betasatellite molecule and based on SDT pairwise alignment, this beta-satellite shared 74,6% nucleotide sequence identity to *Chayote yellow mosaic Benin betasatellite* (ChaYMB) (KT454829; KT454830), found in *Momordica charantia* (Bitter melon) plants showing yellow mosaic symptoms in Benin in 2014 (Table 4.3). According to the 78% beta-satellite species demarcation threshold (Briddon et al. 2008), the beta-satellite found in sample 30 represent a new betasatellite, for which the name Sida corlifolia yellow mosaic betasatellite – [South Africa:Mooketsi:2008] (SiCYMB-[ZA:Mkt:08]) was proposed. Phylogenetic analysis grouped SiCYMB-[ZA:Mkt:08] with other CHaYMB isolates (Figure 4.4). Interestingly, the clade formed by the SiCYMB and ChaYMB isolates are sister to the clade formed by the ToLCTGB, found in another weed plant (sample 10, this study) in the Mooketsi valley (Limpopo), and this close phylogenetic relationship is reflecting their close geographic relationship and possibly, shared ancestry.

Many monopartite and some bipartite begomoviruses found in the OW are often associated with satellite DNA components (alpha- and betasatellites), which may or may not contribute to the induction of typical disease symptoms (Mansoor et al. 2006; Briddon et al. 2010). Most

begomoviruses occurring in the NW have bipartite genomes and do not require satellite DNAs for induction of disease symptoms (Nawaz-ul-Rehman et al. 2009). From literature, we know that betasatellites depend on their helper begomoviruses for replication and spread within the host plant, and they exhibit a remarkable level of flexibility in their capacity to be recognized and replicated by Rep proteins of monopartite begomoviruses and, in a some cases, by bipartite begomoviruses (Briddon et al. 2010). Their role in diseases caused by these helper begomoviruses is complex and can range from being required for development of typical symptoms to enhancing symptom development (synergism) or not being required at all. This study provided the first reports of beta-satellite molecules in South Africa. The first betasatellite (sample 10), called (ToLCTGB-[ZA:Mkt:2014]), was found in association with a newly identified monopartite begomovirus (MalCSV-[ZA:Mkt:2014]) in *M. coromandelianum* host, and the second betasatellite is a newly identified betasatellite called SiCYMB-[ZA:Mkt:08] (sample 30), that was found in association with a newly identified bipartite begomovirus (SiCGMV-[ZA:Mkt:2014]), in *S. corlifolia* weed host. Both satellite molecules were found in weed hosts, frequently fed on by *B. tabaci* and well known for being begomoviral hosts. Currently, the role of these satellites in pathogenicity and the nature of their relationships with their helper viruses (i.e., are these stable relationships or reassortants) remains to be investigated. When taking into consideration the diversity of begomoviruses and satellite components reported over the last few years, around the world and also in parts of Africa (Brown et al. 2015), the finding of betasatellite molecules in association with mono and bipartite begomoviruses in South Africa is not surprising. It however highlights the urgent need for more in depth monitoring and characterization of begomovirus diversity among crops and weed across South Africa.

A second bipartite virus was identified from plant sample 242 , which was collected in 2015 in Cato Ridge (KwaZulu-Natal) The sample tested positive for TY-PCR and produced and RCA/RFLP profile with bands adding up to ~5,6kbp, thus possibly indicating the presence of a bipartite begomovirus. The RCA-DNA from this sample was sequenced in the individual NGS sequencing run in 2015 and de novo assembly recovered two begomoviral contigs (~2.8 kb) that represents a DNA-A and a DNA-B begomoviral genomic components. Based on nucleotide-sequence comparisons with the closest relatives downloaded from NCBI, the DNA-A and DNA-B components shared the highest nucleotide sequence identity (70.3% and 33%) to the DNA-A and DNA-B component of an isolate of *Soybean chlorotic blotch virus* (SbCBV), respectively (Table 4.4). Because the DNA-B component shared such low sequence similarity across its full genome sequence, the individual ORFs (BC1 and BV1) were analysed and showed the highest nucleotide and amino acid sequence identity to the corresponding ORFs from the DNA-B component of SbCBV (37,3 – 57,6%) (Table 4.5). According to the begomovirus species demarcation threshold of 91% (Brown et al. 2015), the bipartite begomovirus found in S242 represent an a new

begomovirus species, for which the name Phaseolus vulgaris begomovirus was proposed, as plant sample 242 was identified as *Phaseolus vulgaris* (bean) by DNA barcoding. At the time of collection, no symptoms were observed on the bean plant, but again, this could be due to the (early) stage of development, therefore no symptom description could be incorporated into the proposed viral name. Further investigation into the biological characteristics of this bipartite begomovirus are therefore urgently required.

Phaseolus vulgaris, also known as the common bean or green bean, is part of the legume family (Fabaceae). It is an annual legume that is grown worldwide and is a vital source of protein in many developing countries where it is the second most important source of dietary protein and starch after maize. An evolutionary distinct subgroup of bipartite begomoviruses have been reported to naturally infecting common bean and non-cultivated legume hosts around the world, including *Bean calico mosaic virus* BCaMV), *Bean chlorosis virus* (BChV), *Bean chlorotic mosaic virus* (BChMV), *Bean dwarf mosaic virus* (BDMV), *Bean golden mosaic virus* (BGMV), *Bean golden yellow mosaic virus* (BGYMV), *Bean yellow mosaic Mexico virus* (BYMMxV)*, Macroptilium yellow spot virus* (MaYSV), *Sida micrantha mosaic virus* (SimMV), *Mungbean yellow mosaic India virus* (MYMIV) and *Mungbean yellow mosaic virus* (MYMV) (Fauquet et al. 2008). Some species, such as BGMV, has been important bean pathogens, causing significant yield losses (Morales, 2006), but despite their economic importance, few studies have focused on begomovirus populations infecting legume crops and non-cultivated legume hosts (Faria and Maxwell, 1999; Lima et al. 2013). iill

Up untill this study, only three bipartite legumovirus species have been identified in Africa (Nigeria and Uganda), including SbCBV, *Soybean mild mottle virus* (SbMMV) and *Desmodium mottle virus* (DesMoV) (Alabi et al. 2010; Mollel et al. 2017). Furthermore, the only geminiviruses previously reported to infect bean plants specifically in South Africa, include *Bean yellow dwarf virus* (BeYDV) and *Bean yellow dwarf virus-mild* (BeYDV-m), both viruses belonging to the Mastrevirus genus, and found to significantly affect French bean production in the Northern Province and Mpumalanga around 1997 (Liu et al. 1997; Halley-Scott et al. 2007). The *Phaseolus vulgaris begomovirus* identified in sample 242 represent a new bipartite legumovirus species for Africa and groups phylogenetically in a cluster with other OW bipartite begomoviruses, and specifically in a cluster with the legumovirus SbCBV (Fig 4.6). In addition to this close phylogenetic relationship, both SbCBV and Phaseolus vulgaris begomovirus identified in sample 242 does not possess the characteristic N-terminal nuclear localizing signal (NLS) 7-PWRsMaGT unique to the CP of NW begomoviruses, they lack the virus-sense AV2, a signature gene present in 'Old World' begomoviruses, but unlike SbCBV and other 'legumoviruses', no ORF AC5 could be detected for *Phaseolus vulgaris begomovirus*. Another legumovirus, namely SbMMV, has also been reported

to lack ORF AC5 (Olufemi et al. 2010). Unlike in other begomoviruses, AC5 is reported to be conserved in 'legumoviruses', but its function is still unclear (Ilyas et al. 2009). Since soybean is an introduced crop in South Africa, the identification of a distinct legumovirus naturally infecting bean crops in South Africa, suggests the occurrence of 'legumoviruses' in indigenous plants in the country and underscores the need for an in depth look at the diversity of begomoviruses infecting bean crops and non-cultivated legume hosts in South Africa.

Plant sample 8, was collected in February in 2014 in the Mooketsi Valley (Limpopo) and although it tested negative in the TY- PCR, it produced a RCA/RFLP profile and was subsequently sequenced as part of the pooled-2 samples and then later, sequenced again as sample 8, in the individual NGS sequencing run in 2015. One contig (2182 nt) was recovered from the both NGS sequencing runs and blast analysis revealed a high sequence similarity (76% identity) to the sequence of *Bemisia*-associated genomovirus AdDF (KY230613), a virus from the family *Genomoviridae*, genus *Gemybolavirus*. Further analysis of the putative viral contig revealed a typical geminivirus origin of replication, IR region, and two large ORFs that includes a putative CP and Rep protein, similar to those found in *Genomoviridae* like viruses. Based on the 78% for species demarcation rule for genomoviruses (Varsani and Kuprovic, 2017), and taking into account the 76% nucleotide sequence similarity of the viral contig recovered from sample 8 to *Bemisia*-associated genomovirus AdDF, and the presence of a genome organization typical of genomoviruses (Figure 4.7), the name *Cynoglossum offinale* associated circular DNA virus (COasCV) was proposed for this viral isolate. Phylogenetic analysis of the deduced amino acid and Rep protein sequences of COasCV also clustered with other gemycircularvirus-like sequences, including *Bemisia*-associated genomovirus AdDF (KY230613) and *Dragonfly associated gemyduguivirus-1*, but clearly in a separate branch, suggesting that it may be classified in a distinct genus in this family. It should be noted that COasCV was identified from the plant *C. offinale* by DNA barcoding, but at the time of sampling, the plant did not display any disease symptoms. *Cynoglossum offinale* has also previously been reported as a begomovirus host and infection by *Alfalfa mosaic virus* (ALMV) induced typical begomoviral symptoms, including stunting, interveinal chlorotic spots, and reduction of leaf lamina (Bellardi et al. 2002).

In the last decade, viral metagenomic studies using high-throughput sequencing methods, similar to the study presented here, has revolutionized and enhanced our perception of virus diversity in the world around us. Many novel groups of viruses have been discovered over the past decade, including viruses with small, moderately-sized, and even large genomes. One such viral group include the newly formed family *Genomoviridae*, comprising single-stranded small circular DNA viruses (Kupovic et al. 2016; Varsani and Kuprovic, 2017). Although this family of viruses has recently been established, it is rapidly growing (>121 viral genomes recovered up to January

2017) and already include nine different genera (Varsani and Kuprovic, 2017). Members of this virus family have been discovered from a variety of different organisms and environmental samples. The founding member of the family include a plant (*Sclerotinia sclerotiorum*) pathogenic fungus-infecting virus, namely *Sclerotinia gemycircularvirus-1* (SsHADV-1) (Yu et al. 2010, 2013). A total of seven genomovirus species have been described in association with different insect species, including five species within the *Gemycircularvirus* genus, four of which were identified in dragonflies and one from a mosquitoes and two species from the *Gemyduguivirus* and *Gemykibivirus* genera were found in association with dragonflies (Ng et al. 2011; Rosario et al. 2012).

Traditionally, viral classification by the ICTV have always been guided by biological features (i.e host range, replication cycle, structure and properties of virus particles), in addition to genomic sequence information. Investigations into the biological features of newly identified viral species have not been able to keep up with the tremendous pace by which new viral species are currently being identified and added to the global virome, particularly by high-throughput sequencing and metagenomic studies (Simmonds et al. 2017). Most viruses are known today solely from sequence data, available from NCBI, and have not been characterized experimentally. This is also the case with the recently formed *Genomoviridae* family. The only *Genomoviridae* member for which some biological information has been established includes SsHADV-1, which has been shown to infect and confer hypovirulence to the plant pathogenic fungus *S. sclerotiorum* (Yu et al. 2010; 2013) and is also able to infect a mycophagous insect (*Lycoriella ingenua*), which acts as a transmission vector (Liu et al. 2016). The other more than 120 possible members in this virus family, have been recovered and are represented by sequence information only, from diverse environments and hosts. The identification of *Cynoglossum offinale* associated circular DNA virus (COasCV) by this metagenomic study is no different. As for all other sequenced-based virus species, there is a great need for biological characterization of COasCV, to determine the true host, vectors and possible economic impact of this newly identified viral species. But in reality, unless a virus is suspected to be pathogenic and influence the local/global economy, it is unlikely that many of these viruses, including COasCV, will receive the same level of experimental characterization as other important pathogens. Even in the absence of biological data, the genetic sequence of COasCV remains an important resource, that will assist with determining the evolutionary relationship of South African, and global viral isolates, information that are important to understand and predict the evolution of agriculturally pathogenic viruses.

As most of the collected plants in this study were volunteer plants in the agricultural environment and at an early growth stage, flowering or fruit structures that are important features for morphological identification were absent. Plant identification was therefore accomplished by DNA

barcoding, a molecular method that makes use of short, agreed-upon region of a genome, as a unique species identifier and provides identification of a specimen to species level, even if only a small fragment of plant material is available. In this study the *rbcLa* and *matK* genes were used for the barcoding of pland plants (CBOL, 2009; Hollingsworth et al 2011). These two genes were chosed as plant barcodes based on the ease of recovery of rbcLa region and the discriminatory power of the matK region. When the plant samples were collected, a photo was taken of each plant, a voucher specimen was prepared and leaf material was collected and placed into a small bag with silica gel for later DNA extraction. All the plant samples that contained circular DNA viruses, were identified by PCR amplification and sanger sequencing of the *rbcLa* and *matK* genes. Species identification was accomplished by comparing the sequence data with the reference databases in Genbank, using a BLAST search, and also with the BOLD, to identify the closest match for each *rbcLa* and *matK* haplotype. PCR amplification of both the rbcl and the *matK* was generally successful with all the candidate loci in this study (Table 4.8). While only one primer set was used to amplify the *rbcL* locus, four different primer sets were used to amplify the *matK* sequences across all the species tested in this study. The species level identification for all samples were in agreement between the two databases used (BLAST and BOLD), and were also in agreement with the morphological identification provided by examining the prepared Herbarium specimens for each sample. The DNA barcoding approach allowed for rapid and easy identification of the non-cultivated plants, without any specialized botanical training being required.

The most effective management measures for viral diseases, are an integrated pest management (IPM) strategy, aimed at decreasing the amount of viral inoculum. This can be achieved by a combination of physical, chemical and biological practices, and most importantly, developing resistant crop cultivars (Pico *et al.* 1996). Given the difficulty in controlling begomovirus infection by managing the vector population, and time investment required to breed for resistance, removal of the primary inoculums and preventing the movement of inoculums into healthy crops is an important control measure, that will assist in disease reduction. The widespread presence of viruses, and particularly begomoviruses, in uncultivated/wild hosts has been demonstrated repeatedly around the world (Bedford et al*.* 1998; Ambrozevicius et al. 2002; García-Andrés, et al*.* 2006, Azhar et al*.* 2011; Fiallo-Olivé et al*.* 2012; Prajapat et al. 2014) and the list of (begomo)-viruses reported from wild hosts are ever increasing. This study also demonstrated the presence of three monopartite and one bipartite begomovirus, two associated betasatellite molecules, as well as one genomovirus, in four different uncultivated host species, i.e. *Abutilon indicum, Acalypha indica, Cynoglossium officinale, Malvastrum coromendalianum and Sida corlifolia*. At least one or more reports can be found in literature, that indicate each of these weed species to be begomovirus hosts. These reports show that they harbour one or more begomovirus species,

both mono- and bipartite, and in some cases, alpha- and betasatellites or defective DNA molecules as well (Bellardi et al. 2002; Jiang et al. 2005; Huang et al. 2006; Ha et al. 2008; Liu et al. 2009; Paul et al. 2011; Prajapat et al. 2014; Ferro et al. 2017). *Sida* species in particular, are one of the most abundant natural reservoirs for begomoviruses in several regions around the world (Guo and Zhou, 2006; Ferro et al. 2017). Furthermore, weed species such as these are highly adaptable to changing environmental conditions. As such, they are found widely distributed around the world, and also in South Africa, where they not only serve as an important source of primary inoculum, but are virus reservoirs that support mixed infection, recombination, pseudo-recombination and evolution of new viral species. The additional knowledge about alternative hosts of (begomo)-viruses in South Africa provided by this study, can be used to develop effective management practices, such as the implementation of host-free periods, and the removal or eradication of these viral-source plants in agricultural areas. These measures will assist with the successful control of (begomo)-virus infection in crop plants.

From the results of this study, it is clear that a wide range of cultivated plants, such as tomatoes and beans, and non-cultivated weed plants, both indigenous and introduced species, serve as (largely) symptomless viral reservoirs in South Africa, and that the diversity of small, DNA-genome based viruses are vastly underexplored up to date. From this group of viruses, begomoviruses are known to be the causal agents of important diseases that cause tremendous yield losses in crops worldwide (Palmer and Rybicki, 1998). Three of the five identified begomoviruses in this study, three of them novel virus, are known to adversely affect crop production in South Africa (ToCSV). The OW begomoviruses, Corchoroviruses and Legumoviruses that were identified in this study, had both monopartite and bipartite genomes, and both novel monopartite and bipartite viruses were found in association with a novel betasatellite molecules, providing the first report for both bipartite begomoviruses (non-cassava infecting) and the presence of betasatellite molecules in South Africa. This genetic diversity should be kept in mind when begomovirus resistance genes are incorporated by breeders in crop plants, as the durability of resistance based on a single gene would be questionable. These begomoviruses were identified from both crop and uncultivated plants, focusing attention to these plants as reservoirs for survival in the post-harvest season and remind us that they should be removed as part of cultural control methods. Lastly, a novel plant-infecting virus (COasCV) was identified in South Africa, in association with a the plant host *C. offinale*, and although this genomovirus was identified in an apparent asymptomatic host, it highlights the need for biological characterization of more virus within the genomovirus family, which are largely known by genetic information only, in order to understand their role in viral evolution, as well as possible economic impact that they may have.

In conclusion, the results of this study expanded our understanding of the diversity of small, circular, DNA viruses infecting both crop plants and non-cultivated hosts in three provinces in South Africa. It also highlights the need for more in-depth studies across the agricultural and botanical landscape of South Africa. Although the crop infecting viruses are of possible economic concern, the viral diversity discovered in indigenous plants and non-cultivated weeds are expected to be quite dynamic, and will give a complete picture of the evolution and the shaping of viral diversity in South Africa, and Africa as a whole. The results also emphasize the potential for the emergence of novel begomoviruses by genetic recombination amongst these viral reservoirs. The study further lends support to the use of a metagenomic (NGS)-approaches to explore viral diversity, as it once again proved to be a very powerful, although costly, method for viral discovery. Lastly, most of the collected plants were volunteer plants in the agricultural environment and at an early growth stage, without flowering or fruit structures, that are important for traditional identification methods based on morphology. The use of DNA barcoding for plant identification in this study allowed for the rapid and easy identification of the non-cultivated plants, without any specialized botanical training being required.

CHAPTER 6

References

Abdullahi, I., Winter, S., Atiri, G.I. and Thottappilly, G. (2003) Molecular characterization of whitefly, *Bemisia tabaci* (*Hemiptera: Aleyrodidae*) populations infesting cassava. *Bulletin of Entomological Research* 93:97-106.

Abstract of Agricultural Statistics (2011) Directorate: Statistics and Economic Analysis of the Department of Agriculture, Forestry and Fisheries, http://www.daff.gov.za/docs/statsinfo/abstract_2011.pdf.

Accotto, G.P., Navas-Castillo, J., Noris, E., Moriones, E. and Louro, D. (2000) Typing of tomato yellow leaf curl viruses in Europe. *European Journal of Plant Pathology* 106: 179-186.

Adams, I.P., Glover, R.H., Monger, W.A., Mumford, R., Jackeviciene, E., Navalinskiene, M., Samuitene, M. and Boonham M. (2009) Next generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. *Molecular Plant Pathology* 10:537-545.

Alabi, O., Villegas, C., Gregg, L., and Murray K.D. (2016) Complete nucleotide sequences of a new bipartite begomovirus form *Malvastrum* sp. plants with bright yellow mosaic symptoms in South Texas. *Archives of Virology* 161:1729-1733.

Alavandi, S.V. and Poornima, M (2012) Viral metagenomics: a tool for viral discovery and diversity in aquaculture. *Indian Journal of Virology* 23:88-98.

Albuquerque, L.C., Varsani, A., Fernandes, F.R., Pinheriro, B., Martin, D.P., Ferreira, P.O., Lemos, T.O. and Inoure-Nagata, A.K. (2012) Further characterization of tomato-infecting begomoviruses in Brazil. *Archives of Virology* 157:747-752.

Alexander, H.M., Mauck, K.E., Whitfield, A.E. and Malmstrom, C.M. (2014) Plant-virus interactions and agro-ecological interface. *European Journal of Plant Pathology* 138:529-547.

Allander, T., Emerson, S.U., Engle, R.E., Purcel, R.H. and Bukh J. (2001) A virus discovery method incorporating DNase treatment and its application to the identification two bovine parvovirus species. *Proceedings of the National Acadamy of Science* 98:11609-11614.

Ambrozevicius, L.P., Calgario, R.F., Fontes, E.P.B., Carvalho, M.G and Zerbini, F.M (2002) Genetic diversity of begomovirus infecting tomato and associated weeds in south-eastern Brazil. *Fitopatologia Brasileira* 27: 372-377.

Anderson, P.K., Cunningham, A.A., Patel, N.G., Morales, F.J., Epstein, P.R and Daszak, P. (2004) Emerging infectious disease of plants: Pathogen pollution, climate change and agrotechnology drivers. *TRENDS in Ecology and Evolution* 19:535-544.

Antignus, Y. (2007) The management of *Tomato yellow leaf curl virus* (TYLCV) in greenhouses and the open field, a strategy of manipulation. In: *Tomato yellow leaf curl virus* disease: Management, molecular biology, breeding for resistance. Ed. by Czosnek, H. Springer, Dordrecht, The Netherlands, pp. 263-278.

Ashfaq, M., Hebert, P.D.N., Mizra, M.S., Khan, A.M., Mansoor S., Shah G.S. and Zafa Y (2014) DNA Barcoding of *Bemisia tabaci* complex (*Hemiptera*: *Aleyrodidae*) reveals southerly expansion of the dominant whitefly species on cotton in Pakistan. *PLOS One* 9:14.

Azhar, M.T., Amin, I., Anium, Z.I. and Masoor, S. (2011) *Gossypium robinsonii*, an Australian wild cotton species is an asymptomatic host of the cotton leaf curl disease pathogen complex. *Australasian Plant Disease Notes* 6:7-10.

Bafeel, S.O., Arif, I.A., Bakir, M.A., Khan, M.A., Farhan, A.H.A., Homaidan, A.A.A., Ahamed, A. and Thomas, J. (2011) Comparative evaluation of PCR success with universal maturase K (*matK*) and ribulose-1,5-biphosphate carboxylase oxygenase subunit (*rbcL*) for barcoding of some arid plants. *Plant Omics Journal* 44:195-198.

Barbosa, J.C., Barreto, S.S., Inoue-Nagata, A.C. and Rezende, J.A.M. (2011) Characterization and experimental host range of a Brazilian tomato isolate of *Tomato server rugose virus*. *Journal of Phytophathology* 157: 644-646.

Bedford, I.D., Briddon, R., 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 the Entomological Society of America* 125: 311-325.

Bedford, I.D., Kelly A., Banks, G.K., Briddon, R.W., Cenis, J.L., and Markham, P.G. (1998) *Solanum nigrum*: an indigenous weed reservoir for a tomato yellow leaf curl geminivirus in southern Spain. *European Journal of Plant Pathology* 104:221-222.

Bellardi, M.G., Rubies-Autonell, C. and Pisi, A. (2002) *Cynoglossum officinale*, a New Natural Host of *Alfalfa mosaic virus. Disease Notes* 86: 444.

Bellardi, M.G., Rubies-Autonell, C. and Pisi, A. (2002) *Cynoglossum officinale*, a new natural host of *Alfalfa Mosaic Virus*. *Disease Notes* 86(4):444

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell J.M., Bryant J., Carter R.J., Cheethan R.K., Cox A.J., Ellies D., Flatbush M.R., Gormeley, N.A., Humphray, S.J., Irving, L.J., Karbelashvili, M.S., Kirk, S.M., Li, H., Lui, X., Maisinger, K.S., Murray, L.J., Obradovic, B., Ost, T., Parkinson, M.L., Patt, M.R., Rasolonjatovo, I.M., Reed, M.T., Rigatti, R. and Rodighiero, C. (2009). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456: 53-59.

Bernardo, P., Golden, M., Akram, M., Naimuddin Nadarajan, N., Fernandez, E., Granier, M., Rebelo, A.G., Peterschmitt, M., Martin, D.P. and Roumagnac, P. (2013) Identification and characterisation of a highly divergent geminivirus: evolutionary and taxonomic implications. *Virus Research* 177:35–45.

Bernardo, P., Muhire, B., Francois, S., Deshoux, M., Hartnady, P., Farkas, K., Kraberger, S., Filloux, D., Fernandez, E., Galzi, S., Ferdinand, R., Granier, M., Marais, A., Monge Blasco, P., Candresse, T., Escriu, F., Varsani, A., Harkins, G. W., Martin, D. P. and Roumagnac, P. (2016) Molecular characterization and prevalence of two capulaviruses: *Alfalfa leaf curl virus* from France and *Euphorbia caput-medusae latent virus* from South Africa. *Virology* 493, 142-153.

Berrie, L.C., Palmer, K.E., Rybicki E.P., Hiyadat, S.H., Maxwell, D.P.,and Rey, M.E.C. (1997) A new isolate of *African cassava mosaic virus* in South Africa. *African Journal of Root and Tuber Crops* 2: 49-52.

Berrie, L.C., Rybicki, E.P., andRey, M.E.C. (2001) Complete nucleotide sequence and host range of *South African cassava mosaic virus*: further evidence for recombination amongst begomoviruses. *Journal of General Virology* 82: 53-58.

Berry, S.D., Fondong, V.N., Rey, C., Rogan, D., Fauquet, C.M., and Brown, J.K. (2004) Molecular evidence for five distinct *Bemisia tabaci* (Homoptera: Aleyrodidae) Geographic haplotypes associated with cassava plants in sub-Saharan Africa. *Annals of Entomological Society of America* 97:852-859.

Bethke, J.A., Byrne, J.J., Hodges, G.S., Mckenzie, C.L., and Shatters, Jr. R.G. (2009) First record of the Q biotype of the sweet potato whitefly, *Bemisia tabaci* in Guatemala. *Phytoparasitica* 37: 61-64.

Bexfield, N. and Kellam, P. (2011) Metagenomics and the molecular identification of novel viruses. *The Veterinary Journal* 190:191-198.

Bigarré, L., Salah, M., Granier, M., Frutos, R., Thouvenel, J.-C.,and Peterschmitt, M. (1999). Nucleotide sequence evidence for three distinct sugarcane streak mastreviruses. *Archives of Virology*, 144(12), 2331–2344.

Bisimwa, E., Walangululu, J., and Bragard, C. (2015) Cassava mosaic disease yield loss assessment under various altitude agro-ecosystems in the Sid-Kivu region, Democratic Republic of Congo. *Tropiculture* 33: 101-110.

Boonham, N., Kreuze, J., Winter, S., van der Vlugt, R., Bergevoet, J., Tomlinson, J. and Mumford, R. (2014) Methods in virus diagnostics: From ELISA to next generation sequencing. *Virus Research* 186:20-31.

Boukari, W., Alcala-Briseńo, R.I., Kraberger, S., Fernandez, E., Filloux, D., Dougrois, J., Comstock, J.C.,Lett, J., Martin, D.P., Varsani, A., Roumagnac, P., Polston, J.E. and Rott P.C., (2017) Occurrence of novel Mastrevirus in sugarcane germplasm collections in Florida Guadeloupe and Réunion. *Virology Research* 14:146.

Boykin, L.M., Shatters, Jr. R.G., Rosell, R.C., McKenzie, C. L., Bagnall, R., De Barro, P. and Frohlich, D.R. (2007) Global relationships of Bemisia tabaci (*Hemiptera*: *Aleyrodidae*) revealed using Bayesian analysis of mitochondrial COI DNA sequences. *Molecular Phylogenetics and Evolution* 44:1307-1319.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, H.M., Segall, A.M., Mead, D., Azam, F. and Rohwer, F. (2002) Genomic analysis of uncultured marine viral communities. *Proceedings of the National Acadamy of Science* 99: 14250-14255.

Briddon, R.W., Mansoor, S., Bedford, I.D., Pinner, M.S. and Markham, P.G. (2000) Clones of *cotton leaf curl geminivirus* induce symptoms atypical of cotton leaf curl disease. *Virus Genes* 20:19-26.

Briddon, R.W., Heydarnejad, J., Khosrowfar, F., Mussumi, H., Martin, D.P. and Varsani, A. (2010) *Turnip curly top virus*, a highly divergent geminivirus infecting turnip in Iran. *Virus Research* 152: 169-175.

Briddon, R.W., Mansoor, S., Bedford, I.D., Pinner, M.S., Saunder, S.K., Stanley, J., Malik, K.A. and Markham P.G. (2001) Identification of DNA components required for induction of cotton leaf curl disease. *Virology* 285: 234–243.

Briddon, R.W. and Stanley, J. (2006). Subviral agents associated with plant single-stranded DNA viruses. *Virology* 344:189-210.

Briddon, R.W., Bull, S.E., Mansoor, S., Amin, I. and Markham, P.G. (2002) Universal primers for the PCR–mediated amplification of DNA β: a molecule associated with some monopartite begomovirus. *Molecular Biotechnology* 20: 315-318.

Briddon, R.W. (2002) Diversity of European begomoviruses: identification of a new disease complex. *Bulletin* 32: 1-5.

Briddon, R.W., Bull, S.E., Amin, I., Idris, A.M., Mansoor, S., Bedford, I.D., Dhawan, P., Rishi, N., Siwatch, S. S., Abdel-Salam, A.M., Brown, K.J., Zafar, Y. and Markham P.G (2003) Diversity of DNA β, a satellite molecule associated with some monopartite begomoviruses. *Virology* 312:106-121.

Briddon, R.W., Patil, B.L., Bagewadi, B., Nawaz-ul-Rehman, M.S and Fauquet, C.M. (2010) Distinct evolutionary histories of the DNA-A and DNA-B components of bipartite begomoviruses. *BMC Evolutionary Biology*, 10:97.

Brown, K.J., Idris, A.M., Torres-Jeres, I., Banks, G.K.and Wyatt, S.D. (2001) The core region of the coat protein gene is highly useful for establishing the provisional identification and classification of begomoviruses. *Archives of Virology* 146: 1581-1598.

Brown, J.K., and Idris, A.M. (2005) Genetic differentiation of whitefly *Bemisia tabaci* mitochondrial cytochrome oxidase 1, and phylogeographic concordance with the coat protein of the plant virus genus *Begomovirus*. *Annals of Entomological Society of America* 98:827-837.

Brown, K.J., Zerbini, M.L., Navas-Castillo, J., Moriones, E., Ramos-Sobrinho, R., Silva, J.C.F., Fiallo-Olivé, E., Briddon, R.W., Hernández-Zapeda, C., Idris, A., Malathi, V.G., Martin, D.P.,

Rivera-Bustamante, R., Ueda, S. and Varsani, A. (2015) Revision of *Begomovirus* taxonomy based on pairwise sequence comparisons. *Archives of Virology* 160: 1593-1619.

Bull, S.E., Tsai, W.S., Briddon, R.W., Markham, P.G., Stanley, J. and Green, S.K (2004) Diversity f begomovirus DNA β satellite of non-malveceous plants in east and south east Asia. *Archives of Virology* 149: 1193-1200.

Burdon, J. and Thrall, P. (2008) Pathogen evolution across the agro-economical interface: Implications for management. *Evolution Application* 1:57-65.

Burger, J.T.and Maree, H.J. (2015) Metagenomic next generation sequencing of viruses infecting grapevines. *Methods in Molecular Biology* 1302: 315-330.

Candresse, T., Filloux, D., Muhire, B., Julian, C., Galzi, S., Fort, G., Bernardo, P., Dougrois, J., Fernandez, E., Martin, D.P., Varsani, A. and Roumagnac, P. (2014) Appearances can be deceptive: Revealing a hidden viral infection with deep sequencing in plant quarantine context. *PLoS ONE* 9:1-13.

Carabali, A., Belloti, A.C. and Montoya-Lerma, J. (2010) Biological parameters of *Bemisia tabaci* (Gennadius) biotype B (Hemiptera: Aleyrodidae) on *Jatropha gossypifolia*, commercial (*Manihot esculenta*) and wild cassava (*Manihot flabellifolia* and *M. carthaginensis*) *(Euphorbiaceae*). *Neotropical Entomology* 39:562-567.

Castillo-Urquiza, G.P., Beserra, Jr. J.E.A., Bruckner, F.P., Lima, A.T.M., Varsani, A., Alfenas-Zerbini, P. and Zerbini, F.M. (2008) Six novel begomoviruses infecting tomato and associated weeds in Southern Brazil. *Archives of Virology* 153:1985-1989.

CBOL, Plant Working Group, (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences* 106: 12794-12797.

Chu, D., Zhang, Y., Brown, K.J., Cong, B., Xu, B., Wu, Q. and Zhu, G. (2006) The introduction of the Q Biotype Bemisia tabaci from the Mediterranean region into China on ornamental plants. *Florida Entomology* 89:168-174.

Costion, C., Ford, A., Cross, H., Crayn, D., Harrington, M. and Lowe, A. (2011) Plant DNA barcode can accurately estimate species richness in poorly known floras. *PLoS ONE* 6:1-9.

Creamer, R., Hubble, H. and Lewis, A. (2005) Curtovirus infection of chile pepper in New Mexico. *Plant disease* 89:480-487.

Cuénoud, P., Sovolainen, V., Chatrou, L.W., Grayer, R.J. and Chae, M.W. (2002) Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid *rbcL*, *atpB* and *matK* DNA sequences. *American Journal of Botany* 89:132-144.

Daly, G.M., Leggett, R.M., Rowe,W., Stubbs, S., Wilkinson, M., Ramirez-Gonzalez, R.H., Heeney, J.L. (2015) Host Subtraction, Filtering and Assembly Validations for Novel Viral Discovery Using Next Generation Sequencing Data. *PLoS One*, 10/6: e0129059.

Da Silva, S.J.C., Castillo-Urquiza, G.P., Júnior, B.T.H., Assuncăo, I.P., Lima, G.S.A., Pio-Ribeiro, G., Mizubutin, E.S.G. and Zerbini, F.M (2011) High genetic variability and recombination in a begomovirus population infecting the ubiquitous weed *Cleome affinis* in north eastern Brazil. *Archives of Virology* 156: 2205-2213.

De Barro P.J., Lui S., Boykin L.M. & Dinsdale A.B. (2011) *Bemisia tabaci*: A statement of species status. *Annual Reviews of Entomology* 56:1-510.

Delatte, H., Duyck, P., Triboine, A., David, P., Becker, N., Bonato, O.and Reynaud, B. (2009). Differential invasion success among biotypes: case of *Bemisia tabaci*. *Biological Invasions* 11:1059-10770.

Delwart, E.L. (2007) Viral metagenomics. *Reviews in Medical Virology*, 17:115-131.

Dennehy, T.J., DeGain, B.A., Harpold, V.S., Brown, K.J., Morin, S. and Fabrick, J.A. (2005) New challenges to management of whitefly resistance to insecticides in Arizona. *Vegetable Report*.

Dennehy, T.J., DeGain, B.A., Harpold, V.S., Zaborac, M., Morin, S., Fabrick, J.A., Nichols, R.L., Brown, K.J., Byne, F.J. and Li, X. (2010) Extraordinary resistance to insecticides reveals exotic Q biotype of *Bemisia tabaci* in the New World. *Journal of economic Entomology* 103: 2174-2186.

Deng, L., Ignacio-Espinoza, J. C., Gregory, A.C., Poulos, B.T., Weitz, J.S., Hugenholtz, P. and Sullivan, M.B. (2014). Viral tagging reveals discrete populations in *Synechococcus* viral genome sequence. *Nature* 513:242-245.

De La Rua, P., Simon, B., Cifuentes, D., Martinez-Mora, C. and Cenis, J.L., (2006) New insights into the mitochondrial phylogeny of the whitefly Bemisia tabaci (Hemiptera Aleyrodidae) in Mediterranean basin. *Journal of Zoological Systematics and Evolutional Research* 44, 25-33.

Dinsdale, A., Cook, L., Riginos, C., Buckley, Y.M. and De Barro, P. (2010) Refined global analysis of Bemisia tabaci (Hemiptera: Sternorrhycnha: Aleyrodoidea: Aleyrodidae) mitochondrial cytochrome oxidase 1 to identify species level genetic boundaries. *Annals of Entomological Society of America* 103:196-208.

Dinsdale, E.A., Edwards, R.A., Hall, D., Angly, F., Breitbart, M., Brulc, J.M., Rohwer, F. (25 authors) (2008). Functional metagenomic profiling of nine biomes. *Nature*, 455(7214): 830–830.

Donaire, L., Wang, Y., Gonzalez-Ibeas, D., Mayer, K.F., Aranda, M.A. and Llave, C. (2009) Deep-sequencing of plant viral small RNAs reveals effective and widespread targeting of viral genomes. Virology 392, 203–214.

Drayton, G.M., Teulon, D. A.J., Workman, P.J.and Scott, L.A.W., (2009) The Christmas dispersal of Bemisia tabaci (GENNADIUS) in New Zealand. *Pest on Horticultural Crops* 63: 310-314.

Edwards, R.A., and Rohwer F. (2005) Viral metagenomics. *National Review of Microbiology*, 3: 504-510.

Esterhuizen, L.L., Mabasa, K.G., van Heerden, S.W., Czosnek, H., Brown, K., van Heerden, H. and Rey, M.E.C. (2012a) Genetic identification of members of the *Bemisia tabaci* cryptic species complex from South Africa reveals native and introduced haplotypes. *Journal of Applied Entomology*, 137 (1-2), 122-135.

Esterhuizen, L.L. (2012b) A study of the South African *Tomato curly stunt virus* pathosystem: epidemiology, molecular diversity and resistance. PhD. Thesis. University of Johannesburg, Johanesburg, South Africa.

Fargette, D., Konate, G., Fauquet, C., Muller, E., Peterschmitt, M. and Thresh, J.M. (2006). Molecular ecology and emergence of tropical plant viruses. *Annual Review of Phytopathology* 44: 235-260

Faria, J.C. and Maxwell, D.P. (1999) Variability in geminivirus isolates associated with *Phaseolus* spp. in Brazil. *Phytopathology* 89, 262–268.

Fauquet, C. and Stanley, J. (2003) Geminivirus classification and nomenclature: progress and problems. *Annals of Applied Biology* 142:165-189.

Fauquet, C.M., Bisaro, D.M., Briddon, R.W., Brown, K. J., Harrison, B.D., Rybicki, E.P., Stenger, D.C. and Stanley, J. (2003). Revision of taxonomic criteria for species demarcation in the family *Geminiviridae*, and an updated list of begomovirus species. *Archives of Virology* 148: 405-421.

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

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39:783-791.

Fernandes, J.J., Carvalho, M.G., Andrade, E.C., Bromnonschenkel ,S.H., Fonte,s E.P.B. and Zerbini, F.M. (2006) Biological and molecular properties of *Tomato rugose mosaic virus* (ToRMV), a new tomato-infecting virus begomovirus from Brazil. *Plant Pathology* 55: 513-522.

Ferro, C.G., Silva, J.P., Xavier, C.A.D., Godinho, M.T., Lima, A.T.M., Mar, T.B., Lau D. and Zerbini, F.M. (2017) The ever-increasing diversity of begomoviruses infecting non-cultivated hosts: new species from *Sida* spp. and *Leonurus sibiricus*, plus two New World alphasatellites. *Annals of Applied Biology* 170: 204–218.

Fields, A.T., Abercrombie, D.L., Eng, R., Feldheim, K., Chapman, D.D. (2015) A novel mini-DNA Barcoding assay to identify processed fins from internationally protected shark species. *PLoS ONE* 10:1-10

Frohlich, D.R., Torres-Jeres, I., Bedford, I.D., Markham, P.G. and Brown, K.J. (1999) A phylogeographical analysis of *Bemisia tabaci* species complex based on mitochondrial DNA markers. *Molecular Ecology* 8:1683-1691.

Galimberti, A., Mattia, F.D, Losa, A., Bruni, I., Federici, S., Casiraghi, M., Martellos, S. and Labra, M. (2013) DNA Barcoding as a new tool for food traceability. *Food Research International* 50:55-63.

García-Andrés, S., Monci, F., Navas-Castillo, J. and Moriones, E. (2006) Begomovirus genetic diversity in the native plant reservoir *Solanum nigrum*: evidence for the presence of a new virus species of recombinant nature. *Virology* 350: 433–442.

Geering, A.D.W., Maumus, F., Copetti, D., Choisne ,N., Zwickl ,D.J., Zytnick ,M., McTaggart ,A.R., Scalabrin, S., Vezzulli, S., Wing, R.A., Quesneville, H. and Teycheney, P.(2014) Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution. *Nature Communications* 5: 5269-5279.

George, B., Kumar, R.V. and Chakraborty, S. (2014) Molecular characterization of *Chilli leaf curl virus* and satellite molecules associated with leaf curl disease of *Amaranthus* spp. *Virus Genes* 48: 397-401.

Gonçalves, P.F.M., Oliveire-Marques, A.R., Matsumoto, T.E., Miyaki, C.Y. and (2015) DNA Barcoding identities of illegal parrot trade. *Journal of Heredity* 106.

Graham, A.P., Martin, D.P. and Roye, M.E. (2010) Molecular characterization and phylogeny of two begomoviruses infecting *Malvastrum americanum* in Jamaica: evidence of the contribution of inter-species recombination to the evolution of malvaceous weed-associated begomoviruses from the northern Caribbean. *Virus Genes* 40(2): 256-66.

Guo, X. and Zhou, X. (2006) Molecular characterization of a new Begomovirus infecting *Sida cordifolia* and its associated satellite DNA molecules. *Virus Genes*, 33(3): 279-285

Ha, C., Coomb, S., Revill P., Harding, R., Vu, M., Dale, J. (2008). Molecular characterization of begomoviruses and DNA satellites from Vietnam: additional evidence that the New World geminiviruses were present in the Old World prior to continental separation. *Journal of General Virology*, 89 312–326.

Halary, S., Temmam, S., Raoult, D. and Desnues, C. (2016) Viral metagenomics: are we missing the giants. *Current Opinion in Microbiology* 31:34-43.

Hall, R., Wang, J., Todd, A.K., Bissielo, A.B., Yen, S., Strydom, H., Moore, N.E., Ren, X., Huang, Q.S. and Carter, P.E., Peacy M. (2014) Evolution of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery. *Journal of Virological Methods*. 195:194-204.

Halley-Stott, R. P. Tanzer, F. Martin, D. P. and Rybicki, E. P. (2007) The complete nucleotide sequence of a mild strain of *Bean yellow dwarf virus*. *Archives of Virology*, 152:1237–124

Hartvig I., Czako M., Kjæ E.D., Nelsen L.R., Theilade I. (2015) The use of DNA Barcoding in identification and conservation of Rosewood (*Dalbergia* spp.) *PLoS ONE* 10:9.

Hadjistylli, M., George, K.R. and Gauthier, N. (2015) Frist report of the Sub-Saharan Africa 2 species of the *Bemisia tabaci* complex in the Southern France. *Phytoparasitica* 43: 679-687.

Haible, D. Kober,S. and Jeske, H. (2006) Rolling circle amplification revolutionizes 524 diagnosis and genomics of geminiviruses. *Journal of Virological Methods* 135: 9-16.

Hanley-Bowdoin L., Settlage S.B., Orozco B.M., Nagar S. & Robertson D (1999). *Critical Reviews in Plant Science* 18:71-106.

Hanssen, I.M., Lapidot, M. and Thomma, B.P.H.J. (2010) Emerging Viral Diseases of Tomato Crops. *Molecular Plant Microbe Interaction* 23: 539-548.

Harrison, B.D. (1985) Advances in geminivirus research. *Annual Review of Phytopathology* 23: 55-82.

Harrison, B. D., Swanson, M. M. & Fargette, D. (2002) Begomovirus coat protein: Serology, variation and functions. *Physiological and Molecular Plant Pathology* 60:257-271.

Herbert, P.D.N., Cywinska, A., Ball, S.L. and de Waard J.R. (2003) Biological identifications through DNA barcodes. *The Royal Society* 270:313-321.

Hernández-Zapeda, C., Varsani, A. and Brown, K.J. (2013) Intergeneric recombination between a new, spinach-infecting curtovirus and a new geminivirus belonging to te genus *Becurtovirus*: first new world exemplar. *Archives of Virology* 158: 2245-2254.

Hill, B.G, (1967) A study of *Trialeurodes vaporariorum* and *Bemisia tabaci* on tobacco, with special references to the transmission of the leaf curl disease. M. Sc thesis, University of Pretoria, South Africa.1:127-146.

Hilu, K.W. and Liang, H. (1997) The *matK* gene: sequence variation and application in plant systematics. *American Journal of Botany* 84:830-839.

Hollingsworth, P.M., Graham, S.W.and Little, D.P. (2011) Choosing and Using a Plant DNA Barcode. *PLoS ONE* 6:1-13.

Hogenhout, S.A., Ammar, E., Whitefield, A.A. and Redinbaugh, M.G. (2008) Insect vector interactions with persistently transmitted viruses. *Annual Review of Phytophathology* 46:1-418.

Holmfeldt, K., Solonenko, N., Shah, M., Corrier, K., Riemann, L., VerBerkmoes, C. and Sullivan, M.B. (2013) Twelve previously unknown phage genera are ubiquitous in global oceans. *Proceedings of the National Acadamy of Science* 110:12798-12803.

Horowitz, A.R., Kontsedalov, S., Khasdaan, V. and Ishaaya, I. (2005) Biotypes B and Q of *Bemisia tabaci* and their relevance to neonicotinoid and pyriproxfen resistance. *Archives of Insect Biochemistry and Physiology* 58:216-225.

Hopkins, M., Kailasan, S., Cohen, A., Roux, S., Tucker, K.P., Shevenell, A., Agbandje-Mckenna, M. and Breitbart, M. (2014) Diversity of environmental single-stranded DNA phages revealed by PCR amplification of the partial major capsid protein. *International Society for Microbial Ecology* 8:2093-2103.

Howe, A. and Chain, P.S.G. (2015) Challenges and opportunities in understanding microbial communities with metagenome assembly (accompanied by IPython notebook turorial). *Frontiers in Microbiology* 6:678-681.

Höfer, P., Engel, M., Jeske H. and Frischmut T. (1997) Nucleotide sequence of a new bipartite geminivirus isolated from common weed *Sida rhombifolia* in Costa Rica. *Journal of General Virology* 78:1785-1790.

Idris, A.M., Briddon, R.W., Bull, S.E. and Brown, K.J (2005) Cotton leaf Gezira virus-satellite DNAs represent a divergent, geographically isolated Nile Basin lineage: predictive identification of a satDNA REP-binding motif. *Virus Research* 109: 19-32.

Idris, A.M., Shahid, M.S., Briddon, R.W., Khan, R.J., Zhu, K.J. and Brown, J.K. (2011) An unusual alphastellite associated with monopartite begomoviruses attenuates symptoms and reduces betasatellite accumulation. *Journal of General Virology* 92:706-717.

Idris, A.M., Al-Saleh M., Piatek, M.J., Al-shahwam, I., Ali, A. and Brown, J.K (2014) Viral metagenomics: Analysis of begomoviruse by Illumina high throughput sequences. *Viruses* 3:1219-1236.

Ilyas, M., Qazi, J., Mansoor, S. and Briddon, R.W. (2009) Molecular characterisation and infectivity of a ‘‘Legumovirus’’ (genus Begomovirus: family Geminiviridae) infecting the leguminous weed *Rhynchosia minima* in Pakistan. *Virus Research* 145:279–284.

Inoue_nagata, A.K., Lima, M.F. and Gilbertson R.L. (2016) A review of geminivirus (begomovirus) disease in vegetable and other crops in Brazil: current status and approaches for management. *Horticutura Brasileira* 34:008-018.

Jeske, H. (2009). Geminiviruses. *TT Viruses* 331:185-226.

Jiang, T. and Zhou, X.P. (2005.) Molecular characterization of a distinct begomovirus species and its associated satellite DNA isolated from *Malvastrum coromandelianum* in China. *Virus Genes* 31: 43–8.

Jin, T., Hou, X., Li, P. and Zhou, F. (2015) A novel method of automatic plant species identification using sparse representation leaf tooth features. *PLoS ONE* 10:1-20.

Jones, D.R. (2003) Plant viruses transmitted by whiteflies. *European Journal of Plant Pathology* 109: 195-219.

Jones. P.G (2004) The ecology and epidemiology of whitefly transmitted viruses in Latin America. *Virus Research* 100:57-65.

Jove,l J., Reski, G., Rothenstein, D., Ringel, M., Frischmuth, T. and Jeske, H. (2004) *Sida micrantha* mosaic is associated with a complex infection of begomoviruses different from *Abutilon mosaic virus*. *Archives of Virology* 149: 829-841.

Kardani, S.G., Heydarnejad, J., Zakiaghl, M., Mehrvar, M., Kraberger, S.and Varsani, A. (2013) Diversity of *Beet curly top Iran Virus* isolated from different host in Iran. *Virus Genes* 46: 571-575.

Kashina, B.D., Mabagala, R.B. and Mpunami, A.A. (2002) Reservoir weed hosts of tomato yellow leaf curl *Begomovirus* from Tanzania. *Archives of Phytopathology* 35: 269-278.

Kehoe, M.A., Coutts, B.A., Buirchell, B. J., and Jones, R.A.C. (2014). Plant Virology and Next Generation Sequencing: Experiences with a Potyvirus. *PLoS ONE* 9:104580.

King, A.M.Q., Adams, M.J., Carstens, E.B. and Lefkowits, E.K. (2012) Virus taxonomy Ninth report of the International Committee on Taxonomy of Viruses. *San Diego: Elsevier Academic Press*.

Knief, C. (2014) Analysis of plant microbe interactions in the era of next generation sequencing technologies. *Frontiers in Plant Science* 5:216.

Kon, T. and Gilbertson, R.L. (2012) Two genetically related begomoviruses causing tomato leaf curl disease in Togo and Nigeria differ in virulence and host range but do not require a betasatellite for induction of disease symptoms. *Archives of Virology* 157 (1) 107-120.

Kozich, J.J., Westcott, S.L., Bexter, N.T., Highlander, S.K. and Schloss, P.D. (2013) Development of a dual index sequencing strategy and curation pipeline for analysing amplicon sequence data on the MiSeq Illumina sequencing platform. *Applied and Environmental Biology* 79: 5112-5120.

Kress, W.J., Wurdack, K.J., Zimmer, E.A., Weigt and L.E., Janzen, D.H. (2005) Use of DNA Barcodes to identify flowering plants. *Proceedings of the National Acadamy of Science* 102:8369-8374.

Kreuze, J. (2014) siRNA deep sequencing and assembly: piecing together viral infections, in Detection and Diagnostics of Plant Pathogens, eds Gullino, M.L., Bonants, P.J.M., editors. (Dordrecht: Springer) 21–38.

Kreuze, J.K., Perez, A., Untiveros, M., Quispe, D., Fuentes, S., Barker, I. and Simon, R. (2009) Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: A generic method for diagnosis, discovery and sequencing of viruses. *Virology* 338:1-7

Krishnamurthy, S.R. Wang, D. (2017) Origins and challenges of virus dark matter. *Virus Research* 239;136-142.

Krupovic, M., Ghabrial, S.A., Jiang, D. and Varsani, A. (2016) *Genomoviridae*: a new family of widespread single-stranded DNA viruses. *Archives of Virology* 161:2633–2643.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution* 33(7):1870-4

Lapidot, M. and Friedmann, M. (2002) Breeding the resistance for whitefly-transmitted geminiviruses. *Annals of Applied Biology* 140:109-127.

Lazano, G., Trenado, H.P., Fiallo-olivé, E., Chirinos, D., Geraud-Pouey, F., Briddon, R.W. and Navas-Castillo, J. (2016) Characterization of Nom-coding DNA Satellites associated with Sweepoviruses (Genus *Begomovirus*, *Geminiviridae*)-Definition of a distinct class of Begomovirus-Associated satellites. *Frontiers in Microbiology* 7:13.

Lefeuvre, P., Martin, D.P., Hoareau, M., Naze, F., Delatte, H., Thierry, M., Varsani, A., Becker, N., Reynaud, N. and Lett, J.M. (2007) Begomovirus 'melting pot' in the South-West Indian Ocean islands: molecular diversity and evolution through recombination. *Journal of General Virology* 88:3458-3468.

Lefeuvre, P., Lett, J. snd Martin, D.P. (2008) Recombination, decreased host specificity and increased mobility may have driven the emergence of *maize streak virus* as an agricultural pathogen. *Journal of General Virology* 89: 2063-2074.

Lefeuvre, P. and Moriones, E. (2015) Recombination as a motor of host switches and virus emergence: geminiviruses as case studies. *Current Opinion in Virology* 10:14-19

Legg, J.P. (1996) Host associated strains within Ugandan populations of whitefly *Bemisia tabaci* (Genn.), (Homo.,Aleyrodidae). *Journal of Applied Entomology* 120:523-527.

Legg, J.P., French, R., Dogan, D., Okao-Okuja, G. and Brown J.K. (2002) A distinct *Bemisia tabaci* (Gennadius) (Hemiptera: Sternorrhyncha: Aleyrodidae) genotype cluster is associated with the epidemic of server cassava mosaic virus disease in Uganda. *Molecular Ecology* 11:1219-1229.

Leke, W.N., Mignouna, D.B., Brown, J.K. and Kvarnheden, A. (2015) Begomovirus disease complex: emerging threat to vegetable production systems of West and Central Africa. *Agriculture and Food Security* 4: 14.

Leke, W.N., Khatabi, B., Mignouna, D.B., Brown, K.J. and Fondong, V.N. (2016) Complete genome of a new bipartite begomovirus infecting cotton in the Republic of Bennin in West Africa. *Archives of Virology* 161: 2329-2333.

Levy, S.E. and Myers, R.M. (2016. Advancements in Next generation sequencing. The *Annual Review of Genomics and Human Genetics* 17:95-115.

Lima, A.T.M., Sobrinho, R.R., Gonza´lez-Aguilera, J., Rocha, C.S., Silva, S.J.C., Xavier, C.A.D., Silva, F.N., Duffy, S. and Zerbini, F.M. (2013). Synonymous site variation due to recombination explains higher genetic variability in begomovirus populations infecting non-cultivated hosts. *Journal of General Virology* 94, 418–431.

Liu, L., van Tonder T., Pietersen, G., Davies, J.W. and Stanley, J. (1997). Molecular characterization of a subgroup I geminivirus from a legume in South Africa. *Journal of General Virology* 78: 2113-2117.

Liu, P., Xie, Y. and Zhou, X.P (2009) *Malvastrum coromandelianum* is an alternative host of *Tomato yellow leaf curl China virus*. *Plant Pathology* 58: 403.

Lozano, G., Trenad, H.P, Fiallo-Olive E., Chirinos, D., Geraud-Pouey F., Briddon, R.W and Navas-Castillo, J. (2016) Charecterization of non-coding DNA satellites associated with Sweepoviruses (Genus *Begomovirus*, *Geminiviridae*)- Definition of a distinct class of Begomovirus-Associated satellites. *Frontiers in Microbiology* 7: 13.

López, M.M., Llop, P., Olmos, A., Marco-Noales, E., Cambra, M. and Bertolini, E. (2009). Are molecular tools solving challenges posed by detection of plant pathogenic bacteria and viruses? *Current Issues on Molecular Biology* 11:13-46.

Magi, A., Benell,i M., Gozzini, A., Girolami ,F., Torricelli, F., Brand,i M.L., (2010) Bioinformatics for next generation sequencing data. *Genes* 1:294-307.

Malmstrom, C.M. andAlexander, H.M. (2016) Effects of crop viruses on wild plants. *Current opinion on Virology* 19:30-36.

Mankga, L.T., Yessoufou, K., Moteetee, A.M., Daru, B.H. and van der Bank, M. (2013) Efficacy of the core DNA Barcodes in identifying poorly conserved plant materials commonly used in South African traditional medicine. *Zookeys* 365:215-233.

Mansoor S., Briddon R.W., Zafar Y., Stanley J. (2003) Geminivirus disease complexes: an emerging threat. *Trends in plant science* 8:128-134.

Mansoor, S., Zafar, Y. and Briddon, R.W. (2006) Geminivirus disease complexes: the threat is spreading. *Trends in Plant Science* 11:209–212.

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

Mardis, E.R. (2013) Next generation sequencing platforms. *Annual Review for Analytical Chemistry* 6:287-303.

Massart S., Olmos A., Jijakli H. and Candresse, T. (2014) Current impact and future directions of high throughput sequencing in plant virus diagnostics. *Virus Research* 188:90-96

Marston, D.A., McElhinney, L.M., Ellis, R.J., Horton, D.L., Wise, E.L., Leech, S.L., David, D., de Lamballerie, X. and Fooks, A.R. (2013) Next generation sequencing of viral RNA genomes. *BMC Genomics* 14:444.

Marozova, O and Marra M.A., (2008) Applications of next generation sequencing techniques in functional genomics. *Genomics* 92:255-264.

Marzano, S.L and Domier, L.L (2016) Reprint of Novel mycovirus discovered from metatranscriptomics survey of soybean phyllosphere phylobiomes. *Virus Research* 219:11-21.

Martin, D.P. and Shepherd, D.N. (2009) The epidemiology, economic impact and control of maize streak disease. *Food Security* 1:305-315.

Maruthi, M.N., Colvin, J., Thwaites R.M., Banks, G.K., Gibson, G. and Seal S.E (2004) Reproductive incompatibility and cytochrome oxidase 1 gene sequence variability amongst host adapted and geographically separate *Bemisia tabaci* populations (Hemiptora: Aleyrodidae). *Systematic Entomology* 29: 560-568.

Maxam, A.M. and Gilbert, W. (1977) A new method for sequencing DNA. *Proceedings of the National Academy of Sciences* 74:560-564.

McKenzie C.L., Hodges G., Osborne L.S., Byrne F.J. & shatters Jr. R.G. (2009) Distribution of *Bemisia tabaci* (*Hemiptera*: *Aleyrodidae*) biotpes in Florida- investigating the Q type invasion. *Journal of Economic Entomology* 102: 670-676.

Melgarejo, T.A., Tatsuya, K., Rojas, M.R., Paz-Carrasco, L., Zerbini, F.M. and Gilbertson, R,L (2013) Characterization of a new world monopartite Begomovirus causing leaf curl disease of tomato in Ecuador and Peru reveals a new direction un geminivirus evolution. *Journal of Virology* 87: 5397-5413.

Metzker, M.L. (2010) Sequencing technologies-the next generation. *Nature Reviews* 11:31-46.

Miller, J.R., Koren, S., Sutton, G. (2010) Assembly algorithms for next-generation sequencing data. *Genomics* 95:315-327.

Mochizuki, T., Krupovic, M., Pehau-Arnaudet, G., Sako, Y., Forterre, P. and Prangishvili, D. (2013) Archeal virus with exceptional virion architecture and the largest single-stranded DNA genome. *Proceedings of the National Academy of Sciences* 109:13386-13391.

Morales, F.J and Anderson, P.K (2001) The emergence and dissemination of whitefly-transmitted geminiviruses in Latin America. *Archives of Virology* 146:415-441.

Morales, F.J. (2001) conventional breeding for resistance *Bemisia tabaci*-transmitted geminiviruses. *Crop Protection* 20:825-834.

Morales, F.J. and Jones, P.G. (2004) The ecology and epidemiology of whitefly transmitted viruses in Latin America. *Virus Research* 100:57-65.

Morales, F.J. (2006) Tropical whitefly IPM Project. *Advances in Virus research* 69:249-311

Moriones E. and Navas-Castillo J. (2000) Tomato yellow leaf curl virus, an emerging virus complex causing epidemics worldwide. *Virus Research* 71: 123–134.

Moodley, V., Gubba, A. and Mafongoya, P. (2018) A Survey of Whitefly-Transmitted Viruses on Tomato Crops in South Africa. *Plant Disease* (In press).

Mubin, M., Akhtar, S., Amin, I., Briddon, R.W. and Mansoor S. (2012) *Xanthium strumarium*: a weed host of components of begomovirus-satellite complexes affecting crops. *Virus Genes* 44: 112-119.

Mubin, M., Shahid, M.S., Tahir, M.N., Briddon, R.W.and Mansoor, S. (2010) Characterization of Begomovirus components form a weed that suggests that begomoviruses may associate with multiple DNA satellites. *Virus genes* 40:452-457.

Muthukumar, V., Melcher, U., Pierce, M., Wiley, G.B., Roe, B.A., Palmer, M.W., Thapa, V., Ali, A. and Ding, T. (2009) Non-cultivated plants of the Tallgrass Prairie Preserve of northeastern Oklamhoma frequently contain virus-like sequences in particulate fractions. *Virus Research* 141: 169-173.

Nassar, N.M.A. (2002) Cassava, *Manihot esculenta* Crantz, genetic resources: origin of the crop, its evolution and relationships with wild relatives. *Genetics and Molecular Research* 1 (4): 298-305

Navas-Castillo, J., Fiallo-Olive, E. and Sanchez-Campos, S. (2011) Emerging virus diseases transmitted by whiteflies. *Annual Review of Phytopathology* 49: 15.1-15.30.

Ng, T.F.F., Manire C., Borrowman K., Langer, T. & Ehrhart L. (2009) Discovery of a novel single-stranded DNA virus from a sea turtle fibropapilloma using viral metagenomics. *Journal of Virology* 83: 2500-2509.

Ng, J.C.K. and Falk, B.W. (2006) Virus-vector interactions mediating nonpersistent and semipersistent transmission of plant viruses. *Annual Review of Phytopathology* 44: 183-212.

Ng, T.F.F., Willner, D.L., Lim, Y.W., Schmieder, R., Chau, B., Nilsson, C., Anthony, S., Ruan, Y.J., Rohwer, F. and Breitbart, M. (2011). Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. *PLoS One* 6:20579.

Ng, T. F. F., Duffy, S., Polston, J. E., Bixby, E., Vallad, G. E., and Breitbart, M. (2011). Exploring the diversity of plant DNA viruses and their satellites using vector-enabled metagenomics on whiteflies. *PLoS ONE* 6:19050

Nweke, F., Spencer, I., Dunstan, S. and Lynam, J. (2001) The Cassava Transformation: Africa's best kept secret. Michigan State University Press, East Lansing.

Olufemi, J.A., Kumar, P.L., Mgbechi-Ezeri, J.U. and Naidu, R.A. (2010) Two new 'legumoviruses' (genus *Begomovirus*) naturally infecting soybean in Nigeria. *Archives of Virology* 155: 643-656.

Palanga, A., Filloux, D., Martin, D.P., Fernandez, E., Gargani ,D., Ferdinand, R., Zabre, J., Bouda, Z., Neya, J.B., Sawadogo, M., Traore, O., Peterscmitt, M. and Roumagnac, P. (2016)

Metagenomic-based screening and molecular characterization of Cowpea infecting virus in Burkina Faso. *PLoS ONE* 11:10.

Padidam, M., Sawyer, S. and Fauquet, C.M. (1999) Possible emergence of new geminiviruses by frequent recombination. *Virology* 265:218-225.

Palmer, K.E. and Rybicki, E.P. (1998) The molecular biology of mastreviruses. *Advances in Virus Research*. 20:183-234.

Paul, S., Ghosh, R., Roy, A., Ghosh, S.K. (2012) Analysis of the coat protein gene sequences of begomoviruses associated with different weed species in India. *Phytoparasitica* 40 (1): 95-100

Phan, T.G., Kapusinsizky, B., Wang, C., Rose, R.K., Lipton, H.L. and Delwart, E.L. (2011) The faecal viral flora of wild rodents. *PLoS Pathology* 7:9.

Pico, B., Diez, M.J. & Nuez, F. (1996) Viral diseases causing the greatest economic losses to the tomato crop .2. The *Tomato yellow leaf curl virus* - A review. *Scientia Horticulturae* 67:151-196.

Pieterson, G., Idris, A.M., Kruger, K. and Brown, J.K. (2008) Characterization of tomato curly stunt virus: A new tomato-infecting begomovirus from South Africa. *Plant Pathology* 57: 809-818.

Pieterson, G., Idris, A.M., Krüger, K. and Brown, J.K (2000) *Tomato curly stunt virus*, a new begomovirus of tomato within the *Tomato yellow leaf curl virus*-Is cluster in South Africa. *Plant Virus* 84:810.

Perring, T.M. (2001) The *Bemisia tabaci* species complex. *Crop Protection* 20:725-737.

Postol, J.E., McGovern, R.J. and Brown, L.G. (1999) Introduction of Tomato yellow leaf curl virus in Florida and implications for the spread of this and other geminiviruses of tomato. *Plant Disease* 83:984-988.

Prabhaker, N., Castle, S., Henneberry, T.J.and Toscano, N.C. (2005) Assessment of cross-resistance potential to neonicotinoid insecticides in *Bemisia tabaci* (*Hemiptera*: *Aleyrodidae*). *Bulletin of Entomological Research* 95: 535-543.

Prabha, K., Baranwal, V.K. and Jain R.K. (2013) Applications of next generation high throughput sequencing technologies in characterization, discovery and molecular interaction of plant viruses. *Indian Journal of Virology* 24:157-165.

Prajapat, R., Marwal, A. and Gaur R.K. (2013) Begomovirus associated with alternative host weeds: a critical appraisal. *Archives of Phytopathology* 47: 157-170.

Prosser, S.W., Goszczynski, D.E. and Meng B. (2007) Molecular analysis of double stranded RNAs reveals complex infection of grapevines with multiple viruses. *Virus Research* 124:151-159.

Quali, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni ,A., Swerdlow, H.P. and Gu, Y. (2012. A tale of three next generation sequencing platforms: Comparison of Ion Torrent, Pacific Bioscience and Illumina MiSeq sequencers. *BMC Genomics* 13:341.

Rasario K. and Breitbart, M. (2011) Exploring the viral world through metagenomics. *Current Opinion in Virology* 1: 289-297.

Rasairo, K.,Dayaram, A., Marinov, M., Ware, J., Kraber ,S., Stainton, D., Breitbart, M. and Varsani, A. (2012) Diverse ssDNA viruses discovered in dragonflies (Odonata:Epiprocta). *Journal of General Virology* 93:2668-2681.

Rasairo, K., Marr,C., Varsani, A., Kraberger, S., Stainton, D., Morions, E., Polston, J.E. and Breitbart, M. (2016) Begomovirus-associated satellite DNA diversity captured through vector-enabled metagenomics (VEM) surveys using whiteflies (*Aleyrodidae*). *Viruses* 8: 16.

Rattray, J. (2012) The implications of increasing global demand for corn. *UW-L Journal of Undergraduate Research XV*.

Ratnasingham, S and Herbet, P.D.N. (2007) BOLD: The Barcode of Life Data System. *Molecular Ecology Notes* 7: 355 –364.

Rejeshwari, R., Reddy, R.V.C., Maruthi, M.N., Colvin, J., Seal, S.E. and Muniyappa, V. (2005) Host range, vector relationships and sequence comparison of a begomovirus infecting hibiscus in India. *Annals of Applied Biology* 147:15-25.

Rey, M. E., Ndunguru, J., Berrie, L. C., Paximadis, M., Berry, S., Cossa, N., Nuaila, V. N., Mabasa, K. G., Abraham, N., Rybicki, E. P., Martin, D., Pietersen, G., … Esterhuizen, L. L. (2012). Diversity

of dicotyledenous-infecting geminiviruses and their associated DNA molecules in southern Africa, including the South-west Indian ocean islands. *Viruses*: 1753-91.

Ribeiro, S.G., Ambrozevicius, L.P., Ávila, A.C., Bezerra, I.C., Calegario, R.F., Fernandes, J.J., Lima, M.F., de Mello Rocha, H. and Zerbini, F.M. (2003) Distribution and genetic diversity of tomato-infecting begomoviruses in Brazil. *Archives of Virology* 148:281-295.

Rocha, C.S., Castillo-Urquiza, G.P., Lima, A.T.M., Silva, F.N., Xavier, C.A.D., Hora-Junior, B.T., Beserra-Junior, J.E.A.., Malta, A.W.O., Martin, D.P., Varsani, A., Alfenas-Zerbini, P., Mizubuti, E.S.G.and Zerbini, F.M. (2013) Brazilian begomovirus populations are highly recombinant, rapidly evolving and segregated based on geographical location. *Journal of Virology* 87:5784-5799.

Rojas, M.R., Hagen, C., Lucas, W.J. and Gilbertson, R.L. (2005) Exploiting chinks in the plant's armor: Evolution and emergence of geminiviruses. *Annual Reviews of Phytopathology* 43: 361-394.

Rojas, M.R. and Gilbertson, R.L. (2008) Emerging plant viruses: a diversity of mechanism and opportunities. In: Roossinck M.J. (eds) *Plant Virus Evolution*. Springer, Berlin, Heidelberg 27-51.

Roossinck, M. (2011) The big unknown: plant virus biodiversity. *Current opinion on virology* 1:63-67.

Roossinck, M.J., Martin D.P., Roumagnac P. (2015) Plant virus metagenomics: Advances in virus discovery. *Phytopathology* 105: 116-127.

Roossinck, M. J. (2010). Lifestyles of plant viruses. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *365*(1548), 1899-905.

Roossinck, M.J. (2012) Plant virus metagenomics: Biodiversity and ecology. *Annual Review of Genetics 46*:357-367.

Rossouw, L.T. (2016) Metabolite profiling of defence-related secondary metabolites in *Solanum lycopersicum*, in response to whitefly mediated begomovirus infection. MSc. Dissertation. University of Johannesburg, Johannesburg, South Africa.

Rose, R., Constantinide, B., Tapinos, A., Robertson, D.L. and Prosperi, M. (2016) Challenges in the analysis of viral metagenomes. *Virus Evolution* 2:

Saitou, N. and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4:406-425.

Sanger, F., Nicklen S., Coulson A.R. (1977) DNA sequencing with chain-terminatin inhibitors. *Proceedings of the National Academy of Sciences* 74:5463-5467.

Saunders, K., Bedford, I.D., Briddon, R.W., Markha, P.G., Wong, S.M. and Stanley, J. (2000) A unique virus complex causes Ageratum yellow vein disease. *Proceedings of the National Academy of Sciences* 97:6890-6895.

Seal, S.E., Jeger M.J. & Van den Bosch F. (2006a). Begomovirus evolution and disease management. *Plant Virus epidemiology* 67:297-316.

Seal, S.E., van den Bosch, F. and Jeger, M. (2006b) Factors influencing begomovirus evolution and their increasing global significance: Implications for sustainable control. *Critical Reviews in Plant Sciences* 25:23-46.

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

Shokralla, S., Spal,l J.L., (2012) Next generation sequencing technologies for environmental DNA research. *Molecular Ecology* 21:1794-1805.

Siebrasse, E.A., Reyes, A., Lima, E.S., Zhao, G., Mkakosya, R.S., Manary, M.J., Gordon, J.I. and Wang, D. (2012) Identification of MW Polyomavirus, a novel polyomavirus in human stool. *Journal of Virology* 86:10321-10326.

Simmonds, P., Adams, M.J., Benkő, M., Breitbart, M., Brister, J.R., Carstens, E.B., (26 authors) Zerbini, F. M. (2017). Virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology* 15 161–168.

Singht, A.K., Chattopadhyay, B. and Chakraborty, S. (2006) Biology and interactions of two distinct monopartite begomoviruses and betasatellite associated reddish leaf curl disease in India. *Virology Journal* 9:18.

Singht, A.K., Chattopadhyay, B. and Chakraborty, S. (2012) Biology and interactions of two distinct monopartite begomoviryses and betasatellites associated with radish leaf curl disease in India. *Virology Journal* 9:43

Staats, M., Arulandhu, A.J., Gravendeel, B., Holst-Jensen, A., Scholtens, I., Peelen, T., Prins, T.W. and Kok, E. (2016) Advances in DNA metabarcoding for food and wildlife forensic species identification. *Analylitcal and Bioanalytical Chemistry*. 408:4615-4630.

Stanley, J., Bisaro, D.M., Briddon, R.W., Brown, J.K., Fauquet, C.M., Harrison, B.D., Rybicki, E.P. and Stenger, D.C. (2005) Family *Geminiviridae*. In Virus Taxonomy: Eighth Report of the International Committee on Taxonomy of Viruses, pp. 301–326, Elsevier, London.

Sneath, P.H.A. and Sokal, R.R. (1973) Numerical taxonomy: principles and practice of numerical classification. *Freeman and Company San Francisco*: 573

Sseruwagi, P., Legg, J.P., Maruthi, M.N., Colvin, J., Rey, M.E.C. and Brown, J.K. (2005) Genetic diversity of *Bemisia tabaci* (Gennedius) (Hemiptera: Aleyrodidae) populations and pressure of the B biotype and a non-B biotype that can induce silverleaf symptoms in squash, in Uganda. *Annals of Applied Biology* 147:253-265.

Sseruwagi, P., Maruthi, M.N., Colvin, J., Rey, M.E.C. and Brown, J.K., and Legg, J.P. (2006) Colonization of non-cassava plant species by cassava whiteflies (*Bemisia tabaci*) in Uganda. *Entomologia Experimentalis et Applicata* 119:145-153.

Stobbe, A.H. and Roossinck, M.J. (2014) Plant virus metagenomics: what we know and why we need to know more. *Frontiers in Plant Science* 5:1-5.

Tahir, M., Amin, I., Haider, M.S., Mansoor, S. & Briddon, R.W. (2015) Ageratum enation virus- A begomovirus of weeds with the potential of infecting crops. *Viruses* 7:647-665.

Tahir, M., Haider, M.S. and Briddon, R.W. (2010) Complete nucleotide sequences of a distinct bipartite begomovirus, *Bitter gourd yellow vein virus*, infecting *Momordica charantia. Archives of Virology* 155: 1901-1905.

Tamura, K., Nei, M., and Kumar, S. (2004). Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proceedings of the National Academy of Sciences (USA*) 101:11030-11035.

Thatcher, J. (1976) Whitefly transmission and etiology of Tobacco leaf curl disease. M.Sc thesis, University of Natal, South Africa.

Thomas, T., Gilbert, J., Meyer, F. (2012) Metagenomics- a guide from sampling to data analysis. *Microbial Informatics and Experimentation* 2:1-12.

Tilman, D., Balzer, C., Hill, J., Befort, B.L. (2011) Global food demand and sustainable intensification of agriculture. *Proceedings of the National Academy of Sciences* 108:20260-20264.

Tokarz, R., Williams, S.H., Sameroff, S., Leon, M.S., Jain, K. and Lipkin, W.I. (2014) Virome analysis of *Amblyomma americanum*, *Dermacentor variabilis* and *Ixodes scapularis* ticks reveals a novel high divergent vertebrate and invertebrate viruses. *Journal of Virology* 88:11480-11492.

Tshiala, M.F. and Olwoch, J.M. (2010) Impact of climate variability on tomato production in Limpopo Province, South Africa. *African Journal of Agricultural Research* 5:2945-2951.

Turcatti, G., Romeeu, A., Fedurco, M. and Tairi, A (2008) A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acid Research* 36:13

Valentine, D.L., Thurber, R.V., Wegley, L., White, B.A. and Rohwer, F. (2008) Functional metagenomic profiling of nine biomes. *Nature* 452: 692-632.

Varma, A. and Malathi, V.G. (2003) Emerging geminivirus problems: A serious threat to crop production. *Annals of Applied Biology* 142:145-164.

Varsani, A., Shepherd, D.N., Monjane, A.L., Owor, B.E., Erdmann, J.B., Rybicki, E.P., Peterschmitt, M., Briddon, R.W., Markham, P.G., Oluwafemi, S., Windram, O.P., Varma, A. and

Malathi, V.G. (2003) Emerging geminivirus problems: A serious threat to crop production. *Annals of Applied Biology* 142:145-164.

Varsani, A., Shephered, D.N., Monjane, A.L., Owor, B.E., Erdman, J.B., Rybicki, E.P., Peterschmitt, M., Briddon, R.W., Markham, P.G., Oluwafemi, S., Windram, O.P., Lefeuvre, P., Lett, J. and Martin, D.P (2008) Recombination, decreased host specificity and increased mobility may have driven the emergence of maize streak virus as an agricultural pathogen. *Journal of General Virology* 89:2063-2074.

Varsani, A., Navas-Castillo, J., Moriones, E. Hernandez-Zapeda, C., Idris, A., Brown, J.K., Zerbini, F.M. and Martin, D.P. (2014) Establishment of three new genera in the family Geminiviradae: Becurtovirus, Eragrovirus and Turncurtovirus. *Archives of Virology* 159:2193-2203.

Varsani, A., Roumagnac, P., Fuchs, M., Navas-Castillo, J., Moriones, E., Idris A., Briddon R.W., Rivera-Bustamante R., Zerbini F.M., Martin D.P. (2017) *Capulavirus* and *Grablovirus*: two new genera in the family *Geminiviridae*. *Archives of Virology* 162:1819-1831.

Varsani, A. and Krupovic, M. (2017) Sequence-based taxonomic framework for the classification of uncultured single-stranded DNA viruses of the family *Genomoviridae*. *Virus Evolution* 3:37

Wang, G., Liu, Y., Zhu, D., Klau, G.W. and Feng, W. (2015) Bioinformatics methods and biological interpretation for next generation data. *Biomed Research International* 1-2.

Waston, J.D. and Crick, F.H.C. (1953) Molecular structure of nucleic acids. *Nature* 171: 737-738.

Wicke, S. and Quandt, D. (2009) Universal primers for the amplification of the plastid *trnK/matK* region in land plants. *Anales del Jardin Botanico de Madrid* 66:285-288.

Wren, J.D., Roossinck, M.J., Nelson, R.S., Scheets, K., Palmer, M.W., Melcher, U. (2006) Plant virus diversity and ecology. *PLoS Biology* 4: 0314-0315.

Wyant, P.S., Gotthardt, D., Scha¨fer, B., Krenz, B. and Jeske, H. (2011) The genomes of four novel begomoviruses and a new *Sida micrantha* mosaic virus strain from Bolivian weeds. *Archives of Virology* 156: 347-352.

Wylie, T.N., Wylie, K.M., Herter, B.N. and Storch, G.A. (2015) Enhanced virome sequencing using targeted sequence capture. *Genome Research 25*: 1910-1920.

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

Yan, D., Luo, J.Y., Han, Y.M., Peng, C., Dong, X.P., Chen, S.L., Sun, L.G. and Xiao, X.H. (2013) forensic DNA Barcoding and Bio-response studies of Animal horn products used in traditional medicine. *PLoS ONE* 8:

Zerbini, F.M., Briddon, R.W., Idris, A., Martin, D.P., Moriones, E., Navas-Castillo, J., Rivera-Bustamante, R., Roumagnac, P., Varsani, A., and ICTV Report Consortium (2017) ICTV Virus Taxonomy Profile: *Geminiviridae, Journal of General Virology* 98:131-133.

Zheng, L., Wayper, P.J., Gibbs, A.J., Fourment, M., Rodoni, B.C. and Gibbs, M.J. (2008) Accumulation variation at conserved sites in potyvirus genome is driven by species discovery and affects degenerative primer design. *PLoS ONE* 3.

Zheng, Y., Gao, S., Padmanabhan, C., Li R., Galvez, M., Gutierrez, D., Fuentes, S., Ling, K., Kreuze, J. and Fei, Z. (2017) Virus Detect: An automated pipeline for efficient virus discovery using deep sequencing of small RNAs. *Virology* 500:130-138.

Zhou, X., Xie Y., Peng, Y. and Zhang, Z. (2003) *Malvastrum yellow vein virus*, a new *Begomovirus* species associated with DNA satellite molecule. *Chines Science Bulletin* 48:2206-2210.