GENOMIC CHARACTERIZATION OF THE *CACAO SWOLLEN SHOOT VIRUS* COMPLEX AND OTHER *THEOBROMA CACAO*-INFECTING BADNAVIRUSES

by

Nomatter Chingandu

_____

A Dissertation Submitted to the Faculty of the

SCHOOL OF PLANT SCIENCES

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY
WITH A MAJOR IN PLANT PATHOLOGY

In the Graduate College

THE UNIVERSITY OF ARIZONA
2016

THE UNIVERSITY OF ARIZONA

GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Nomatter Chingandu, entitled "Genomic characterization of the *Cacao swollen shoot virus* complex and other T*heobroma cacao*-infecting badnaviruses" and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

_____ Date: 7.27.2016

Dr. Judith K. Brown

_____ Date: 7.27.2016

Dr. Zhongguo Xiong

_____ Date: 7.27.2016

Dr. Peter J. Cotty

_____ Date: 7.27.2016

Dr. Barry M. Pryor

_____ Date: 7.27.2016

Dr. Marc J. Orbach

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

_____ Date: 7.27.2016

Dissertation Director: Dr. Judith K. Brown

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of the requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Nomatter Chingandu

ACKNOWLEDGEMENTS

I would like to thank God, the giver of life, for without Him, this would not have been possible.

I am also grateful to my dissertation advisor, Dr. J.K. Brown, for the exceptional guidance and support I received under her supervision.

Lastly, to my family who were not able to be with me here, thank you very much for loving and encouraging me. My dear sister Heather Johnson, life would not have been the same without you. May God richly bless you all.

DEDICATION

This dissertation is dedicated to my lovely daughter, Nikita, who encouraged and supported me incessantly.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES-continued

ABSTRACT

The cacao swollen shoot disease of *Theobroma cacao* L. (cacao) is caused by *Cacao swollen shoot virus* (CSSV; genus, *Badnavirus*, family, *Caulimoviridae*). The virus is endemic to West Africa, where it poses a serious threat to cocoa production. Despite efforts to control CSSV spread by replacement of infected trees with tolerant cultivars and mealybug vector management, the disease is widespread in West Africa. In Trinidad, leaf mosaic and vein-banding symptoms have been observed in cacao plants in the field since the 1940s, and recently at the International Cocoa Genebank (ICGT), a custodian of cacao germplasm resources. The strains A and B of the suspect Cacao Trinidad virus (CTV) caused the symptoms, and were thought to be related to CSSV, however, viral causality was not demonstrated, until now.

To develop molecular detection methods for CSSV in infected plants, polymerase chain reaction (PCR) amplification of eight regions of the CSSV genome was implemented. The PCR results showed variable amplification frequencies of 19 - 42% at each region, for 124 isolates collected in Cote d'Ivoire and Ghana. Pairwise nucleotide (nt) analyses of the eight regions showed 66-99% shared identities, indicating that CSSV isolates exhibit extensive variability with respect to primer design. The results provided preliminary evidence for the existence of a CSSV complex consisting of four divergent species. The full length genome of 14 CSSV isolates from cacao determined using the Illumina HiSeq platform showed 70-99% shared nt identities. The pairwise nt identities placed CSSV sequences into a group of four distinct species, one of which represented a previously undescribed species. Moreover, the full-length genomes grouped phylogenetically with other badnaviruses and revealed two CSSV subclades with three types of genome arrangements; four, five or six open reading frames (ORFs). Predicted functional protein domains were conserved on each ORF.

Two distinct, full-length genome sequences were determined using the Illumina HiSeq platform, from DNA isolated from cacao leaves exhibiting distinct symptoms in Trinidad. The sequences were validated by PCR-amplification and sequencing of overlapping viral genome fragments. Pairwise nt analysis indicated that each genome shared 52-62% nt identities with CSSV and other badnaviruses, suggesting that the two are distinct species. Phylogenetic analysis indicated that the two sequences are not strains of the same virus, as supposed, but they represent two previously undescribed species in the genus, *Badnavirus*, and they have been named Cacao mild mosaic virus (CaMMV) and Cacao yellow-vein-banding virus (CYVBV). Despite sharing

the same host and causing similar symptoms in cacao, CSSV, CaMMV, and CYVBV are phylogenetically-distinct species.

The discovery of a CSSV species complex and the identification of three new cacao-infecting badnavirus species will support the development of molecular detection tools using the partial and complete genome sequences determined in this study. The ability to develop validated molecular tools for the detection of CSSV and related viruses, CaMMV and CYVBV, in cacao will aid quarantine efforts and safe movement of germplasm from the ICGT in Trinidad to cacao-growing countries, worldwide. Also, molecular diagnostics tools are expected to be useful in efforts underway to develop CSSV-resistant planting material for countries in West Africa, which are currently experiencing continued or new disease outbreaks.

CHAPTER 1
INTRODUCTION

Literature Review

Theobroma cacao and cocoa production

The plant *Theobroma cacao* L. (cacao) is an important neotropical crop that is a source of chocolate, and also some intermediate products, including cocoa powder, cocoa butter, cocoa liquor and cocoa cake. Its origins are in the Amazon basin, South America (Motamayor *et al.*, 2002), and it belongs to the family Malvaceae, subfamily Sterculioideae (Alverson *et al.*, 1999; Bhattacharjee & Kumar, 2013). There are at least 22 Theobroma species, but *T. cacao* is the most widely cultivated, followed by *T. bicolor* and *T. grandiflorum*. Three varieties of cacao are grown for chocolate: Criollo, Forastero and Trinitario (Motamayor *et al.*, 2002; Smulders *et al.*, 2008). The Criollo variety produces fine flavor chocolate and it originated from South America, from which it was later introduced to Central America. Although this variety produces the highest quality of chocolate, it performs poorly in terms of yield and it is also susceptible to diseases. Because of these reasons, foreign genes were introgressed into the Criollo genome, resulting in the hybrid, Forastero. The third variety, Trinitario, is a cross between Criollo and Forastero (Motamayor *et al.*, 2002). Trinitario is more productive and more disease resistant than Criollo, and it is more widely grown (about 9.9% of total cocoa production) than Criollo (0.1% of total cocoa production). Forastero, however, is widely grown and amounts to about 90% of the total cocoa bulk production (Argout *et al.*, 2011).

Stem cuttings, buds, seeds and grafting are the methods used for commercial planting of cacao. Flowers are borne on the cacao stem, rather than on the branches. The cacao pods formed from the flowers have between 40 and 60 seeds (beans). The pods are usually white, green, or red, when not ripe, but they turn green, yellow, red, or purple when they are ripe. The beans from the mature pods are processed and used to make chocolate. The mucilage from the pods is useful in the food industry for making jam and jellies, and for fermented drinks, including wine, and non-fermented glucose drinks. The seed/bean covering (testa) is used for animal feed or as organic fertilizers (Bhattacharjee & Kumar, 2013).

From its origins in the Amazon Basin, cacao was introduced to other continents in the humid tropics, including West Africa, around the 18th – 19th century (Johns & Gibberd, 1951; Motamayor *et al.*, 2002). Currently, cacao is West Africa's major cash crop and 80% of the world's cocoa supply for the 2015-2016 year was produced there (International Cocoa Organization, ICCO, 2016, http://www.icco.org). The remainder of cacao was produced in Indonesia (8%), South America (12%), and others.

Constrains to cacao production

The production of cocoa is hindered by several constrains, including pests and diseases (Clough *et al.*, 2009; Ploetz, 2006). It is estimated that up to 40 % of yield loss is attributed to pests and diseases, which are often endemic to certain geographic locations (ICCO, 2016). Several diseases of cacao have been reported, but those of economic importance in terms of causing significant yield loss are of fungal (and oomycetes) and viral origin (Bowers *et al.*, 2001).

Fungal and fungal-like diseases of cacao

The most important fungal and fungal-like diseases of cacao are the black pod, witches' broom, frosty pod and vascular streak dieback (Bowers *et al.*, 2001). The black pod disease is caused by several species of the oomycete *Phytophtora*, including *P. palmivora, P. megakarya, P. capsici* and *P. citrophthora.* These species cause extensive destruction in almost all cacao-growing regions, worldwide (Bowers *et al.*, 2001). The pathogen infects all parts of the plant, causing stem canker and seedling blight. On the pod, brown lesions are observed, and shriveling of cocoa beans may occur (Bowers *et al.*, 2001).

The witches' broom disease of cacao, caused by *Crinipellis perniciosa*, is prevalent in Latin America. The symptoms are the vegetative brooms from the infection of terminal and axillary buds, and also stem cankers form from the infection of leaves and petioles. The disease causes yield loss by affecting the development of beans (Bowers *et al.*, 2001; Frison *et al.*, 1999). *Moniliophthora roreri* infects pods causing monilia pod rot or frosty pod in Latin America (Bowers *et al.*, 2001). The symptoms appear as small swellings on the pod, which usually turn to water-soaked lesions. The lesions enlarge to form necrotic areas, leading to premature pod ripening and necrosis of beans (Frison *et al.*, 1999).

The vascular streak dieback is characterized by leaf chlorosis and green/yellow mottling. On the wood, brown streaking symptoms are apparent when split longitudinally. The causal pathogen is *Oncobasidium theobromae* (Bowers *et al.*, 2001; Frison *et al.*, 1999) and it is endemic to Asia. However, it is of less economic importance compared to the witches' broom and frosty pod diseases.

Viral diseases of cacao

To date, six viral diseases have been reported in cacao, and most of them are endemic to specific geographic locations (Ollennu, 2001). The viral pathogens include *Cacao necrosis virus* (CNV; genus *Nepovirus*, family *Secoviridae*), *Cocoa yellow mosaic virus* (CoYMV; genus *Tymovirus*, family *Tymoviridae*), *Cocoa yellow vein-banding virus* (CoYVBV), Cacao Trinidad virus (now known as two separate viruses: Cacao mild mosaic virus (CaMMV) and Cacao yellow vein-banding virus (CYVBV), both belonging to the genus, *Badnavirus*, family *Caulimoviridae*), and *Cacao swollen shoot virus* (CSSV; genus, *Badnavirus*, family, *Caulimoviridae*).

Symptoms caused by CNV are characterized by necrotic and translucent spots along the midrib and main veins of the leaves. The virus was reported from Ghana and Nigeria (Kenten, 1972; Owusu, 1971), where it was suspected to be transmitted by nematodes.

The CoYMV was originally described from Sierra Leone in 1958 where it caused mosaic symptoms on the leaves (Brunt *et al.*, 1965; Ding *et al.*, 1990), however the disease symptoms have not been observed again. It is not seed borne, and no vector has been found yet, but it is sap (mechanically) transmissible (Frison *et al.*, 1999). The CoYVBV was reported once from Malaysia but the disease is no longer being reported (Liu & Liew, 1975). The three viral diseases caused by CNV, CoYMV and CoYVBV have not been thoroughly studied, because they are not of significant economic importance, or the associated symptoms are no longer being reported.

The Trinidad viruses, CaMMV and CYVBV, were initially thought to be two strains of the same virus (Posnette, 1944), but they have been recently shown to be distinct viruses (Chapter 4; Chingandu et al., *submitted*). They were first described in 1944 in Trinidad (Posnette, 1944), where they caused symptoms similar to those of CSSV on cacao. A detailed description of the

two viruses is found in Chapter 4. The remainder of this dissertation will be focused on CSSV, which is the most economically important cacao-infecting virus due to the large yield loss incurred when trees are infected (Ollennu, 2001; Thresh, 1958).

History of the cacao swollen shoot disease

The cacao swollen shoot disease caused by CSSV was first described in Ghana 1936 (Steven, 1936). Following its discovery, subsequent experiments and successful graft transmission showed that the disease was of viral origin (Posnette, 1940). A few years later, transmission by mealybugs vector was demonstrated (Box, 1945). Since then, the disease was reported in most of West African cacao producing countries, including Nigeria in 1944 (Thresh & Tinsley, 1959), Cote d'Ivoire in 1946 (Mangenot *et al.*, 1946), Sierra Leone in 1963 (Attafuah *et al.*, 1963), and in Togo it was first reported in 1978 (Partiot *et al.*, 1978). The disease continues to spread within each of the affected countries, with new outbreaks occurring in areas where it was not previously reported, and also, new CSSV infection epicenters are being reported (Domfeh *et al.*, 2011). In Cote d'Ivoire, the disease appears to be spreading west-ward, from the eastern region where it was first discovered, to center-west regions, including Bouaflé, Sinfra, and Issia (Kouakou *et al.*, 2012). Similarly, in Ghana, new epicenters of infection have been described in the Western Region.

Symptoms exhibited by CSSV-infected plants

Symptoms caused by CSSV are evident on leaves, pods, shoots, and roots of the cacao plant (Fig. 1.1), and they may be transient or permanent. Infected plants may not show symptoms for several months, up to 20 months, and quarantine procedures often require assessment for at least 24 months for symptom induction (Frison *et al.*, 1999). The type of symptoms expressed depend on the infecting strain and the conditions in which the plant is grown (Posnette, 1947). Altering growth conditions for a plant infected by one virus strain may result in different symptom phenotypes under the different conditions (Adegbola, 1975). Infected plants go through gradual yield loss in 1 -3 years, followed by tree death in under five years, or as early as two years for virulent strains (Muller, 2008; Posnette, 1947). Symptom development also depends on the method of virus inoculation, for example, inoculation by mealybugs or graft transmission may result in variable frequencies for symptom expression (Posnette & Strickland, 1948).

Foliar symptoms are usually transient on the newest leaves (new flush), and they may be permanent on the mature leaves (Posnette & Strickland, 1948). Symptoms on the flush growth are exhibited as red vein-banding of primary veins, which, as the leaves mature, changes to mosaic, fern pattern, mottle or vein-clearing (Fig. 1.1). Vein clearing is seen in fine veins, and it is accompanied by reduction of chlorophyll content in adjacent mesophyll cells. In such leaves, the mesophyll cells may be undifferentiated and the palisade cells may be absent. Vein-clearing symptoms may also produce fern-like symptoms as the leaf matures, sometimes leading to chlorotic vein banding of larger veins. The chlorosis is due to reduced chloroplast size and quantity (Posnette, 1947). The different leaf symptoms often lead to early leaf senescence and defoliation.

In shoots and roots, there is pronounced swelling of chupon or sucker shoots, and on tap and lateral roots (Fig. 1.1) (Posnette, 1947). The swellings are present on the nodes, internodes or on the terminus of the shoot, and apical buds on swollen shoots typically die (Posnette, 1947). Some mild CSSV strains do not cause swelling of shoot and roots, and cause foliar symptoms instead, however, virulent strains are almost always associated with the swollen shoot symptom (Ollennu & Owusu, 2003). Studies conducted by Jacquot *et al.* (1999) on infected shoots showed that the swellings were due to an increase in size of the xylem parenchyma and phloem tissue. In their study, the xylem had three additional cell layers, and the abnormal thickening of the phloem was caused by an increase in size by two cell layers. They also found that the radial organization of phloem fibers had been disrupted. In addition, the cortex was found to increase by four cell layers, but the cambium was reduced from the normal six to a few layers. However, the pith had not been modified.

The unripe (green) pods of the Amelonado type from an infected tree show dark green mottling, which turns to dark red blotches during the later stages of ripening or infection. Also, the pods become smoother and more rounded, as opposed to the long ovoid shape of the healthy pods (Fig. 1.1). The pods from infected plants also tend to be smaller, and they produce beans that are paler in color and flatter, when compared to pods from non-infected trees (Posnette, 1947).

Taxonomy of CSSV

CSSV is a member of the family, *Caulimoviridae* and genus, *Badnavirus*. Plant viruses grouped in this family are distributed worldwide, but they are usually prevalent in tropical regions, where they affect the production of important tropical crops.

Plant viruses classified in the family, *Caulimoviridae* contain circular, double-stranded (ds) DNA viruses of 6.9 – 9.3 kb in length (International Committee on Taxonomy of Viruses (ICTV): 2015 release, http://www.ictvonline.org/virustaxonomy.asp). The dsDNA strands are not covalently circular but they have discontinuities (gaps) at specific sites (Medberry *et al.*, 1990b). The positive strand has one discontinuity, and the first nucleotide after the discontinuity is designated the nucleotide 1 coordinate, for genome numbering. The negative strand has between one and three gaps, and they differ depending on the genus or species. The gap on the plus strand is created during reverse transcription an used as the transcription start site for the minus strand DNA synthesis. The 5' end of the gap is a reverse complement of the plant cytosolic initiator methionine tRNA ($tRNA^{met}$), which serves as a primer for the virus reverse transcription by the viral-encoded reverse transcriptase (RT) and ribonuclease H (RNase H), collectively known as the replicase (Medberry *et al.*, 1990b). The viruses belonging to the family *Caulimoviridae* replicate through an RNA intermediate (reverse transcription), and are therefore called pararetroviruses (Geering, 2014).

The type species of this family is the *Cauliflower mosaic virus* (CMV), and it was the first to be described for dsDNA plant viruses. Virus species are grouped into eight genera primarily based on the genome organizations (King *et al.*, 2012, ICTV 2016). The eight genera are *Badnavirus*, *Caulimovirus*, *Cavemovirus*, *Petuvirus*, *Rosadnavirus*, *Solendovirus*, *Soymovirus*, and *Tungrovirus*. Of these, six genera have isometric-shaped particles while only *Badnavirus* and *Tungrovirus* have baciliform-shaped particles. The genus, *Badnavirus* is the largest of all, and it contains at least 37 species, of which CSSV is one (ICTV 2016). The most economically important species of the genus, *Badnavirus* include *Banana streak virus*, *Citrus yellow mosaic virus*, and CSSV (Geering, 2014).

Physical properties and molecular biology of CSSV

The particles of CSSV are bacilliform-shaped (Brunt, 1964) and they are not enveloped. The length of the particles is 33 – 346 nm, the modal length being 113, and the width, 28 nm (Hagen *et al.*, 1994)**.** The partially purified virus preparations are inactivated at 50 °C for 10 min, but not at 45 °C for 10 min (Brunt, 1964). However, the virus was not inactivated when infected plant material was immersed in hot water at 52°C for 10 min, at 50°C for 12 min or at 45°C for 30 min (Posnette, 1947).

The reported genome lengths of CSSV genomic sequences vary from 6.9 – 7.2 kb, and they have 4 - 6 open reading frames (ORFs); ORF1, ORF2, ORF3, ORF4, ORFX, and ORFY (Hagen *et al.*, 1993; Muller & Sackey, 2005) that encode proteins of approximately 16 kDa, 15 kDa, 212 kDa, 95 kDa, 13 kDa, and 14 kDa, respectively. Of the six ORFs, ORF4 and ORFX are only present in some genomes but not others, and to date, they are unique to CSSV. The functions of the proteins encoded by ORFs 1, 4, X and Y are unknown. The ORF2 encodes a nucleic-acid binding protein that binds DNA and RNA in a non-specific manner (Jacquot *et al.*, 1996). The largest ORF, ORF3, encodes a polyprotein that is processed to yield the mature viral movement (MP), capsid (CP), the aspartic protease (AP), the viral reverse transcriptase (RT), and the ribonuclease H (RNase H) proteins (Hagen *et al.*, 1993). Different strains of CSSV have been described, based on symptom phenotype (Posnette, 1947), serological analyses (Kenten & Legg, 1971; Sagemann *et al.*, 1985) and molecular methods (Kouakou *et al.*, 2012). Despite this, there is no genomic level information available to link differential symptoms to a particular serotype, of which thus far two have been reported (mild and severe), host range, or putative differences in vector competency.

Transmission of CSSV

Mealybugs

The natural transmission of CSSV is by at least 14 species of mealybugs (family Pseudococcidae), and the most efficient species are *Ferrisia virgata* Ckll., *Planococcus citri* (Rossi), *Pseudococcus njalensis* Laing, and *Pseudococcus exitiabilis* (Box, 1945). One species can transmit more than one CSSV isolate, with similar or variable rate of transmission. All stages of *Pseudococcus njalensis* Laing and *Ferrisia virgata* Ckll. can transmit the virus to healthy trees

(Posnette & Strickland, 1948). The virus is retained during insect molting but it does not replicate in the mealybug vector.

Mealybugs feed on all parts of the plant, including the leaves, flowers, pods, cherelles, or young pods, shoots and cacao beans from which testa has been removed (Posnette & Strickland, 1948). However, the virus acquired from symptomatic young leaves is more readily transmitted than when acquired from other plant parts (Posnette & Strickland, 1948). The species that vector CSSV feed from the phloem (Entwistle & Longworth, 1963) and they have easy access to the virus because CSSV is a phloem-limited virus, and it has been observed in the phloem companion cells (Jacquot *et al.*, 1999a). The minimum time which mealybugs must feed on infected plant material to acquire the virus is three hours, and transmission occurs within the three hours of feeding, after which the virus is lost (Posnette & Strickland, 1948). Thus the mode of transmission is considered to be non-persistent and stylet-borne.

The developmental stage of the mealybug vector determines the frequency of CSSV transmission. Although all stages of the mealybug can transmit CSSV, the young nymphs (crawlers) are the most efficient as they can move from tree to tree, especially when there are intertwining branches (Posnette & Strickland, 1948). In contrast, the adults are sedentary and are less likely to spread the virus. There are two modes that mealybugs use to spread virus; radial and jump spread. Radial spread involves movement from one tree canopy to an adjoining one (Cornwell, 1958). The jump spread occurs when mealybugs move long distances to new hosts, away from the initial outbreak (Cornwell, 1960). This type of movement is facilitated by wind (wind dispersal). With time, the radius of each infection point increases. Mathematical modelling has predicted the consequence of jump spread to be lesser than that of radial spread, especially for the fields that have non-cacao plant species as barriers (Jeger & Thresh, 1993). Other factors affecting the rate of CSSV spread by mealybugs include the virus isolate (mild or virulent), the number of mealybugs feeding and the type of tissue the vector feeds on (Posnette & Strickland, 1948). CSSV strains causing severe symptoms (virulent) are more readily transmissible by mealybugs than mild strains. Although a single mealybug is capable of transmitting CSSV, transmission will occur faster if more mealybugs feed/inoculate the host simultaneously.

Other natural means of local CSSV spread between plants include transmission by vascular contact between interlocking branches, or by root grafting. Long distance CSSV spread can occur by the human-mediated movement of infected cacao seedlings, buds or stems used for grafting, within the same farm or to new locations locally or at a distance.

Experimental virus transmission

Mechanical inoculation of CSSV has been achieved, albeit with difficulty, because of the presence of plant inhibitors (Adomako & Owusu, 1974; Brunt & Kenten, 1963; Wessel-Riemens, 1965). Oxidized phenolic compounds and tannins inactivate CSSV in vitro, and therefore make mechanical inoculation challenging (Brunt & Kenten, 1963). However, the addition of protein (casein, egg and blood albumin, and hide powder) to the virus extraction can improve the infectivity of CSSV (Brunt & Kenten, 1963). The protein binds the tannin that would otherwise interact with CSSV particles. More recent and efficient virus inoculation methods include agro-inoculation (Jacquot *et al.*, 1999a) and particle bombardment (Hagen *et al.*, 1994).

CSSV was detected in pollen from infected cacao trees, but it is not transmitted to healthy trees through cross-pollination (Ameyaw *et al.*, 2013). The beans germinated from cross pollinated pods do not have detectable levels of CSSV, neither are the seedlings resulting from the germination of the beans, therefore CSSV is not seed-transmissible.

Alternative (wild) hosts for CSSV

There are several non-domesticated (wild) plant species that can serve as hosts for CSSV. There is evidence that, before cacao was introduced in West Africa in the early 18th century, CSSV was present in the wild hosts (Tinsley, 1971a). The virus likely went through a host shift from the endemic species to cacao, facilitated by mealybug feeding. The alternative hosts include *Adansonia digitata* L., *Ceiba pentandra* L., *Cola chlamydantha* K.Schum., *Cola gigantean* A. Chev., and *Sterculia tragacantha* Lindl. (Posnette *et al.*, 1950; Tinsley, 1971b; Todd, 1951), and they are thought to have a significant role in the spread of CSSV. The wild hosts belong to the Bombaceae, Sterculiaeceae and Tiliaceae families (Posnette *et al.*, 1950).

Transmission assays of CSSV from cacao to wild hosts using mealybugs for inoculation resulted in variable symptoms. Some plant species were symptomless, while others showed mild and severe symptom phenotypes, similar to those presented in cacao. For *Adansonia digitata* L., severe stunting was observed, while the symptoms in *Bombax buonopoxense* included vein-clearing on the leaves. *Ceiba pentandra* leaves showed vein-banding and stem necrosis. Most wild host species showed symptoms only on newly infected leaves, and they typically became symptomless as they matured. Additionally, the transfer of CSSV to cacao from the alternative host plants one year post infection proved difficult (Posnette *et al.*, 1950). The above mentioned observations led to the hypothesis that virus titer was lower in wild hosts than in cacao.

For optimum growth and yield production, *T. cacao* requires shade, which is provided by various plant species that are often planted near the cacao plants. In Ghana, at least 90 species, representing 30 families, are grown in and around cacao fields, and the most common species include *Albizia zygia, Amphimas pterocarpoides, Antiaris toxicaria, Cola nitida, Ficus exasperata, Milicia excelsa, Morinda lucida, Newbouldia laevis, Persea americana, Ricinodendron heudelotii, Terminalia ivorensis,* and *T. superba* (Richard & Ræbild, 2016). The role of some of the shade trees in the transmission of CSSV is yet to be investigated. However, if not properly pruned, they may facilitate mealybug movement from infected trees to healthy trees. Also, because most mealybugs are polyphagous, the shade trees may sustain mealybug feeding before they move to cacao. *Pseudococcus njalensis,* for example, feeds on over a hundred plant species (Posnette, 1950).

Management of CSSD

A number of methods have been implemented for the control of the swollen shoot disease, however, none have proved to be effective. The methods include elimination of inoculum source (cacao and wild hosts), mealybug control, mild strain cross protection, use of "resistant" or tolerant cacao cultivars, and some cultural practices such as avoiding infected areas.

Elimination of infected trees

To the present, at least 200 million infected trees have been cut in Ghana since the eradication campaign that started in the 1940's (Dzahini-Obiatey *et al.*, 2006; Ollennu, 2001). However, the

program was not successful because of the resistance by farmers to participate, citing yield loss, and also the lack of funding for compensation to the farmers. During the campaign, cacao trees in the 30 m radius of infected were to be removed and new plantations were to be at least 10 m from the old plots. However, not all farmers have adhered to these guidelines and as a result, re-infection rates were high for newly planted plots (Ameyaw *et al.*, 2014; Dzahini-Obiatey *et al.*, 2006). Further, when infected plants are not removed when symptoms of CSSV infection occur, they become virus sources that contribute to nearby spread to other trees. Additionally, asymptomatic trees may go unnoticed, and so also serve as virus inoculum for mealybug transmission, or as a source of bud or grafted material used by farmers to invigorate other trees.

Mealybug vector management

Control of mealybugs is very difficult and currently, no method is effective, because the mealybugs have an impermeable waxy layer on their body that protects them from most insecticides. This makes the insecticide ineffective. The few chemical pesticides that effectively eliminated the mealybugs were either too costly for regular use, or they were toxic to be handled by the farmers (Hanna & Heatherington, 1957). Other control methods target the attendant ants, which play a role in the movement of mealybug nymphs to feeding sites, and also in the protection of the mealybugs by building nests around them. The ants benefit by feeding on the honey dew secreted by the mealybugs. The species *Crematogaster straiatula* (Emery) is the most important in its association with mealybugs (Bigger, 1972b). Some insecticides such as dieldrin were effective at killing the attendant ants, but they disrupted the balance of insect populations, and this caused increased populations of other destructive pests like the pod borers and leaf miners (Entwistle *et al.*, 1959). Chemical spraying also reduced the marketability of the crop, so the use of the insecticides was stopped or reduced. Also, biological control using fungal species predatory insects or parasitoids was tested but it was not successful. The parasites either did not become successfully established, or they were not adequate to control the mealybugs on their own (Ackonor, 1995; Bigger, 1972b).

Mild strain-mediated cross-protection

Mixed CSSV isolate infections can occur in a single host (Posnette & Todd, 1955). The individual strains cause variable symptoms depending on the host and the strain itself. However,

the result of mixed infections is variable. Extant CSSV isolates have been classified as mild or virulent strains. The virulent strains generally cause the death of trees in under two years and have persistent, prominent swellings and variable foliar symptoms (Box, 1945; Posnette, 1947). Mild strains may induce only transient symptoms that appear in the first flush but not in the next. Mild strain cross protection involves the inoculation of a mild strain on a host to protect it from the effects of virulent strains. The result is that the host shows mild or no symptoms at all (Posnette & Todd, 1955). Mealybugs were not able to transmit the virulent strain from such asymptomatic plants to healthy plants, suggesting that the replication of the virulent strain was suppressed.

Interestingly, the mild strain-protected plants tended to produce more yield than non-infected or healthy plant controls (Ameyaw *et al.*, 2016; Posnette & Todd, 1955). The yield and growth rates on cross protected plants were much improved when compared to the non-protected trees (Ameyaw *et al.*, 2016). Trees inoculated with mild strain only had an accumulative 8.9% yield reduction while those inoculated with a combination of mild and virulent strains caused 28.7% yield loss in a 4-year period (Ollennu & Owusu, 2003). This showed that the mild strain protected the trees from the virulent strain, which would have otherwise killed the trees in 2 – 3 years. However, the protection from the mild strain is not durable, and it started breaking after 15 years, suggesting that it might be a short term rather than long term measure (Ameyaw *et al.*, 2016).

Resistant cultivars

Currently, no cacao variety is immune to CSSV. Tolerant hybrids are not easily available to farmers therefore the susceptible genotypes are still being planted (Ameyaw 2013). Experiments that were conducted between 1969 and 1978 showed some form of CSSV tolerance in the Upper Amazon hybrids (Kenten & Legg, 1970), leading to their recommendation to farmers for replacement of the susceptible cultivars (Ameyaw *et al.*, 2014). However, with time, it became evident that they had limited tolerance (Dzahini-Obiatey *et al.*, 2010), and that the tolerance weakened over time (Padi *et al.*, 2013).

Avoidance

Farmers are encouraged to isolate their new cacao plants from other plants with a distance of at least 10 m around the field to discourage mealybug movement. However, the new plantings seemed to get re-infected after some time, at a rate of up to 43% of the number of trees planted (Ollennu & Owusu, 2002). Non-host crops like citrus, cola, and oil palm, which are a source of income for the farmers, have been suggested for use as barrier crops around new fields (Ollennu & Owusu, 2002).

Detection of CSSV

While visual symptoms are used frequently to diagnose infected cacao plants, it is not uncommon for a CSSV-infected plant to be asymptomatic. A more reliable detection of CSSV involves serological, molecular, microscopy methods and grafting to a susceptible cultivar.

Serological methods

The Enzyme linked immuno-sorbent assay (ELISA) is the commonly used serological assay for CSSV detection (Dongo & Orisajo, 2007; Kenten & Legg, 1971; Sagemann *et al.*, 1983). The serum is raised against the coat protein of several mild and virulent strains. The dot blot hybridization has also been used with some degree of success (Sackey & Hull, 1994).

Molecular methods

Polymerase chain reaction (PCR) has been the most effective tool in detection of CSSV. It has been combined with serology (Immunocapture PCR), and also used separately (Hoffmann *et al.*, 1997)

Grafting

Buds from suspected infected plant material is grafted to West African Amelonado genotype seedlings or seedling rootstock and assessed for symptom expression (Frison *et al.*, 1999). This method takes the longest time because symptoms may not be expressed in several months. Also, different virus species in different families may cause similar symptoms, or a new strain might have symptoms not previously described.

Challenges to CSSV management

Due to the high molecular variability among the CSSV isolates, it is challenging to detect all strains using molecular methods. CSSV is known to have an uneven titer distribution in the plant tissue (Sagemann *et al.*, 1985) and variable results may be obtained, depending on the part of the leaf sampled. For example, it is more likely to detect the virus from a symptomatic part of a leaf than on an asymptomatic part of the same leaf (Sagemann *et al.*, 1985). The virus is more readily detected from cotyledons, which have at least 74% higher virus titer compared to detection on leaves, which have about 9% of the total viral titer (Dzahini-Obiatey & Fox, 2010). On the cacao leaf, the highest concentration of the virus was shown to be in the lamina and lowest in the petiole, so the outcome will largely depend on the part selected (Sagemann *et al.*, 1983). Finally, there are not enough genomic sequences to consider for development of comprehensive diagnostic tools.

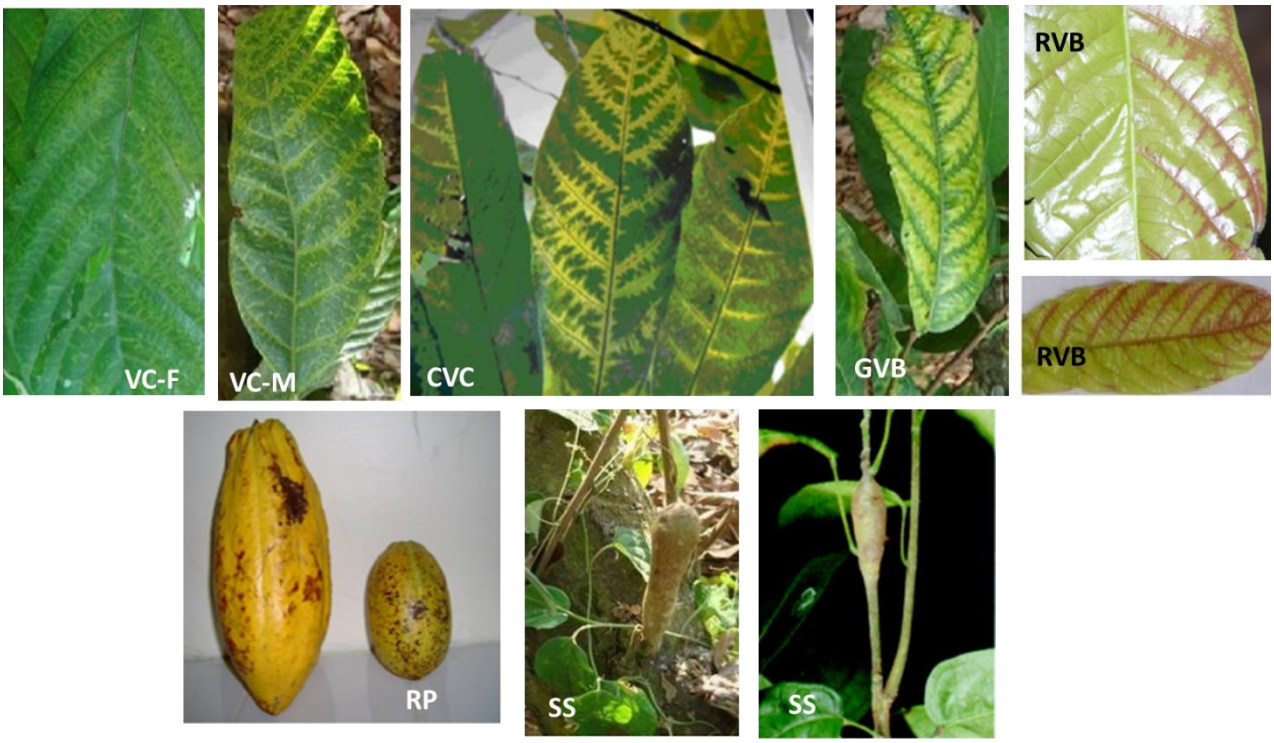

Fig. 1.1. Various symptoms exhibited by CSSV-infected plants. Foliar symptoms include vein clearing on fine veins (VC-F), vein-clearing on major veins (VC-M), chlorotic vein-clearing (CVC) or chlorotic vein banding, green vein banding (GVB) and red vein banding (RVB). Infected pods are rounder and smaller (RP) than non-infected pods. The swollen shoot symptom (SS) can be observed on shoots from mature trees or from young seedlings (Domfeh *et al.*, 2011; Dzahini-Obiatey *et al.*, 2010; Jacquot *et al.*, 1999a).

CHAPTER 2

UNEXPECTED GENOMIC VARIABILITY OF *CACAO SWOLLEN SHOOT VIRUS* AT MULTIPLE LOCI PROVIDES PRELIMINARY EVIDENCE FOR MULTIPLE VIRAL SPECIES

Abstract

The swollen shoot disease of cacao *Theobroma cacao* (L.), caused by *Cacao swollen shoot virus* (CSSV) [*Badnavirus*, *Caulimoviridae*] poses a substantial threat to cocoa production in West Africa. Currently, no serological or molecular diagnostic test has been capable of detecting all isolates in plants symptomatic for swollen shoot disease, suggesting the possible emergence of new or previously uncharacterized variants. To test this hypothesis, the genomic variability of CSSV-like isolates associated with leaf or shoot samples were collected from symptomatic cacao trees, and amplified using five sequence-specific and/or degenerate polymerase chain reaction (PCR)-primer pairs designed based on CSSV sequences available in the GenBank database. In addition, the utility of three previously published CSSV and 'general' Badnavirus primers were included for comparison. In total, the eight primer pairs had potential to PCR-amplify one or more fragments from 75% of the full-length genome of previously reported CSSV isolates. Total DNA was isolated from 124 symptomatic leaf or shoot samples collected in Cote d'Ivoire, and subjected to enrichment of circular DNA using rolling circle amplification, followed by PCR amplification, cloning, and DNA sequencing. Frequency of PCR primer amplification was highly variable among the eight primer pairs tested, yielding from zero to eight amplicons per sample. However, no single primer pair was capable of detecting CSSV in all symptomatic plant samples. Phylogenetic analyses of PCR amplicon sequences obtained using the 8 different primers resolved two groups of CSSV-like isolates like those reported previously, plus an additional group that was basal to the other two. Based on the International Committee on Taxonomy of Viruses-accepted species demarcation, at 80%, the pairwise nucleotide (nt) identities of the RT-RNase H locus revealed four groups of CSSV-like isolates that comprise at least four putative species. The genomic variability among CSSV-like isolates is therefore more extensive than previously reported, underscoring the need to understand the complete range of genomic variability, to guide diagnostics tool development to support epidemiological studies and the development of disease resistance.

Introduction

Theobroma cacao (L.) or cacao (Malvaceae) is cultivated for the cocoa bean, which is used in the confectionary industry and to make cosmetics and other products used worldwide. Sustainable cocoa bean production is essential for the economic well-being of farmers in West Africa (Clough *et al.*, 2009). The world's production of beans used for bulk cocoa is centered in the West African countries of Cameroon, Cote d'Ivoire, Ghana, and Nigeria, contributing to over 70% of the global supply (International Cocoa Organization; ICCO). The remainder of the cacao crop is produced in the Americas and Caribbean Basin, and in Indonesia. Since the cultivation of cacao, following its introduction into West Africa from the New World over one hundred years ago, the production of the cocoa crop there has faced challenges, including reduced production caused by insect pests and diseases caused by fungal (Guest, 2007) and viral pathogens (Ollennu, 2001; Ploetz, 2006). Among the most damaging pathogens of the cacao tree in West Africa is *Cacao swollen shoot virus* (CSSV), which causes the cacao swollen shoot disease (CSSD) (Thresh, 1958). The productivity of trees affected by CSSV is characteristically reduced within 1-3 years after the initial infection, and death of trees can occur within five years, or more, post-infection (Posnette, 1947).

Although cacao is native to the neotropics, CSSV occurs only in West Africa in the countries of Cote d'Ivoire (Mangenot *et al.*, 1946; Muller & Sackey, 2005), Nigeria (Dongo & Orisajo, 2007; Thresh & Tinsley, 1959), Sierra Leone (Attafuah *et al.*, 1963), and Togo (Oro *et al.*, 2012b; Partiot *et al.*, 1978). The *T. cacao* tree has its origin in the Amazon Basin of South America, and the plant was introduced into West Africa during the late 1880s when it became widely cultivated for the bean throughout the region (Motamayor *et al.*, 2002). During 1936, symptoms of a virus-like disease, later named CSSV, were first observed in Ghana by Steven (Steven, 1936). Because cacao is an introduced species, and almost immediately was found to be susceptible to CSSV, it has been hypothesized that the virus is indigenous with its endemic plant host species, and that it underwent a host shift to infect this exotic, introduced species. Results of grafting and mealybug vector transmission experiments, indicated that a number of endemic tree species were naturally infected with CSSV, including *Adansonia digitata, Ceiba pentandra, Cola chlamydantha, Cola gigantean,* and *Sterculia tragacantha,* representing species in the

Bombacaceae, Malvaceae, Sterculiaceae, and Tiliaceae families (Posnette *et al.*, 1950; Tinsley, 1971c; Todd, 1951).

When cacao is infected with CSSV, variable symptoms are expressed, depending on cultivar, time of year, and virus strain (Adegbola, 1975; Posnette, 1947). Characteristic disease symptoms include shoot and root swelling associated with proliferation of phloem tissue (Jacquot *et al.*, 1999b), accompanied by green and then red vein banding on the newly developing leaves and foliar chlorosis and/or mosaic patterns (Cilas *et al.*, 2005; Posnette, 1947). Symptoms usually disappear altogether as the leaf matures (Dzahini-Obiatey & Fox, 2010). Severe CSSV variants are known that cause tree dieback (Dongo & Orisajo, 2007) whereas others strains, considered to be mild, are reported to be asymptomatic or to cause transient symptoms.

The CSSV is a member of the genus, *Badnavirus* (family, *Caulimoviridae*) (King *et al.*, 2012; Lockhart, 1990). The virus has non-enveloped bacilliform-shaped particles of 128 x 28 nm (Brunt, 1964), each containing a circular double-stranded DNA molecule of approximately 7.0-7.3 kb in size. The genome contains one discontinuity in the plus strand (Lot *et al.*, 1991; Uhde *et al.*, 1993), which serves as the priming site for initiation of reverse transcription of the DNA minus strand synthesis (Geering, 2014; King *et al.*, 2012). The CSSV genome has between four and six open reading frames (ORFs) namely, ORF1, ORF2, ORF3, ORF4, ORFX, and ORFY (Hagen *et al.*, 1993; Lot *et al.*, 1991; Muller & Sackey, 2005). The functions of ORFs 1, 4, X, and Y are unknown, but ORF2 encodes a non-specific nucleic-acid binding protein (Jacquot *et al.*, 1996), and ORF3 encodes a polyprotein which is processed by protease catalytic activity to yield the mature viral movement (MP), capsid (CP), aspartate protease (AP), reverse transcriptase (RT), and ribonuclease H proteins (RNase H) (Hagen *et al.*, 1993).

Transmission of CSSV is facilitated by at least 14 mealybug species in a semi-persistent manner (Box, 1945; Posnette, 1950) The virus is ingested and acquired by the insect in less than 4 hrs of feeding, and transmitted within 3 hrs post-acquisition-access feeding, and persists in the vector for up to 3 hrs (Posnette & Strickland, 1948). However, it is not transmitted by pollen or through seed (Ameyaw *et al.*, 2013), although it has been shown to be detectable for a short time post-

germination of seeds, collected from virus-infected trees (Quainoo *et al.*, 2008a). At least a portion of virus spread between cacao trees has been exacerbated by the human-mediated movement of infected cacao cuttings within and between countries in West Africa.

Outbreaks of CSSV and disease incidence have been correlated to the proximity of cacao plantings to forests, and ecosystem factors including elevation, precipitation, and temperature (Lartey, 2013). Thus, disease management has relied primarily on removal of infected trees (Ameyaw & Ollennu, 2006; Dzahini-Obiatey *et al.*, 2006; Jeger & Thresh, 1993), mealybug control (Bigger, 1972a; Hanna & Heatherington, 1957), and relocating production to previously unplanted areas to escape infection, which resulted in abandonment of trees that then provided inoculum to nearby production areas (Clough *et al.*, 2009). Other disease control measures have included chemical control of the mealybug to reduce the population size and decrease transmission frequency. Cross-protection by inoculating plants with mild CSSV strains to protect them against more severe strains was tested as a control measure (Ameyaw *et al.*, 2016; Hughes & Ollennu, 1994). Mild strains offered protection to plants and they performed better compared to non-protected plants in growth, yield and survival rates for over a decade (Ameyaw *et al.*, 2016).

For long-term management, breeding programs have been established to develop disease resistant cultivars, however, genetic resistance to CSSV has broken down in certain cacao genotypes (Padi *et al.*, 2013). Despite these measures, new outbreaks involving a severe decline phenotype were reported in Ghana (Ollennu & Owusu, 1989), and then in Cote d'Ivoire (Kebe *et al.*, 2006) and Togo (Cilas *et al.*, 2005). Previously unrecognized virus variants have been detected in Cote d'Ivoire (Kouakou *et al.*, 2012), however, their status as resistance breaking strains or their relationship to the new outbreaks and disease re-emergence remains to be determined.

Initial disease diagnosis of swollen shoot disease relied on experimental transmission using the mealybug vector or by grafting to 'Amelanado' indicator plants, a hyper-susceptible cacao genotype (Posnette, 1940). A serological (ELISA) test was developed using antisera raised against purified CSSV particles, and it has been widely used for CSSV detection for

epidemiological studies and in breeding programs (Dongo & Orisajo, 2007; Kenten & Legg, 1971; Sagemann *et al.*, 1983). Recently, approximately 25-30% of field isolates from Ghana were found to be undetectable using the ELISA test (personal communication, Ameyaw), suggesting the possibility of previously undiscovered variants of CSSV. Based on the first CSSV genome sequence, which was determined in 1993 for an isolate found in Togo (Hagen *et al.*, 1993), polymerase chain reaction (PCR) amplification tests were implemented (Muller *et al.*, 2001; Quainoo *et al.*, 2008a). Since then, the genome sequence has been determined for six additional isolates collected in Ghana, Cote d'Ivoire, and Togo (Kouakou *et al.*, 2012; Muller & Sackey, 2005). Initially the isolates were thought to be geographically associated, but as additional isolates have been studied, it appears that most or all may be widespread throughout the region (Kouakou *et al.*, 2012).

Seven complete CSSV genome sequences of 7006 to 7297 bp in size are available in the NCBI-GenBank database. They share 71-98% nucleotide (nt) identity. Based on the ICTV cut-off for badnavirus species demarcation at the RT-RNase H region (King *et al.*, 2012), the CSSV variants known thus far represent three badnaviral species. Based on these seven genome sequences, PCR primers have been designed from conserved regions, including the ORF1 (Quainoo *et al.*, 2008b), ORF3 (Kouakou *et al.*, 2012) and universal Badnavirus genus primers targeting the RT-RNase H on ORF3 (Yang *et al.*, 2003). Using the PCR primers based on the 5'-end of ORF3, the virus was detectable in symptomatic leaf samples collected in Cote d'Ivore and Togo (Kouakou *et al.*, 2012; Oro *et al.*, 2012a). Similarly, primers designed to amplify ORF1 were found to detect certain but not all symptomatic isolates from cacao (Ameyaw *et al.*, 2013). Thus, the availability of CSSV primers that facilitate detection of all CSSV-like isolates has become essential to inform epidemiological studies and management, including tree removal programs for the short term and resistance breeding efforts in the long term for West Africa.

The aim of this study was to design and validate PCR primers based on the seven genome sequences available in the GenBank database, for selected cacao and non-cacao samples representing east, central, and western Cote d'Ivoire. To examine the variation among isolates, sequences obtained for each PCR-amplified locus were aligned and analyzed by phylogenetic analysis. In addition, the 570 bp fragment of the reverse transcriptase-RNase H (RT-RNase H)

coding region is the partial genome fragment that has been accepted by the International Committee on Taxonomy of Viruses (ICTV) for species demarcation (King *et al.*, 2012). Primers that were designed to amplify this region were included for comparison with additional genome regions targeted for PCR-amplification in this study. Here, the results have provided the first insights into the genomic variability of the CSSV complex at multiple loci representing more than 75% of the genome sequence for 124 field isolates from cultivated cacao plants and nearby endemic, wild (uncultivated) suspect hosts of CSSV. Of the 124 field isolates, PCR amplicons were obtained for only 69 of 124 isolates, despite the presence of foliar symptoms. Results suggest that previously unrecognized viral variants or even different badnavirus species may be associated with CSSV-like symptoms in cacao, including the recently reported rapid decline phenotype.

Materials and Methods

Plant samples. Symptomatic and non-symptomatic cacao leaf and shoot material (n=124) were collected from fields located in three major cacao growing locations of eastern, western and central Cote d'Ivoire during 2012 (Fig. 2.1). Leaves were also collected from cacao plants maintained in the greenhouse at Centre National de Recherché Agronomique (CNRA), Divo, Côte d'Ivoire. Most of the leaf samples exhibited characteristic CSSV mosaic and green or red vein banding symptoms, and symptomatic shoots showed characteristic swellings. Leaves were also collected from wild host plant species growing in or around cacao fields that were suspected to serve as a reservoir of CSSV. Plant samples were preserved in glycerol and transported to the University of Arizona, Tucson AZ and stored at 4 °C.

Total DNA isolation. The plant material was washed to remove the 100% glycerol storage solution, and blotted dry. Total DNA was isolated according to the Cetyl trimethylammonium bromide (CTAB) DNA isolation procedure (Doyle & Doyle, 1990). For each sample, 100 mg of leaf tissue was ground in liquid nitrogen using a sterile plastic pestle in a 1.6 mL Eppendorf tube, followed by transfer to a 2-mL tube containing 1.2 mL of CTAB buffer with 2 % β-mercaptoethanol (Sigma-Aldrich), and four 3.2 mm stainless steel beads (Next Advance). The samples were pulverized for 5 min by placing tubes in the Mini Beadbeater™ (Biospec

Products), the standard CTAB isolation steps followed. The DNA pellet was dissolved in a final volume of 100 µL of low TE buffer (10 mM Tris-HCL (pH 7.5) containing 0.1 mM EDTA (pH 8.0), and stored at -20 °C**.**

Primer design. Five primer pairs were designed to amplify CSSV sequences from the samples based on the seven CSSV reference sequences (Accession numbers AJ534983, AJ608931, AJ609019, AJ609020, AJ781003, JN606110, and L14546) available in the NCBI GenBank database. The sequences were aligned using the MUSCLE algorithm (Edgar, 2004) implemented in CLC Sequence viewer 7.5, available at http://www.clcbio.com/products/clc-sequence-viewer. Four of the primers were designed from the ORF3 coding region (nt coordinates 1272-6776 on the GenBank Accession number L14546 sequence), and they were designated P1, P2, P3 and RT. The fifth primer (P4) was designed to amplify mostly the non-coding intergenic region at coordinates 6777 - 440 on Accession number L14546 sequence, with some overlapping sequences on ORFs 1 and 3. The position of the each primer and approximate size of each PCR amplicon are provided in Table 2.1 and Fig. 2.2. All of the primers, except the RT, had degenerate bases incorporated, where necessary, based on the CSSV reference sequences alignment. For comparison, the previously published primers that amplify the 5' end of ORF3 (Kouakou *et al.*, 2012) and ORF1 (Quainoo *et al.*, 2008b), and Badnavirus 'universal', degenerate PCR primers (Yang *et al.*, 2003) designed to amplify all the then available badnavirus sequences, were included in the study (Table 2.1; Fig. 2.2). The three primers are designated as ORF3A, ORF1, and Badnavirus, respectively. To determine the quality of total DNA used for PCR, representative samples for which no primer or only one or two primer pairs successfully amplified CSSV sequences, the CoxExon1F/2R (Demesure *et al.*, 1995) primer pair was used in PCR amplification to detect a 2 kbp non-coding fragment of plant mitochondrial DNA. The primer sequences were: CoxExon1F-5'-CAGTGGGTTGGTCTGGTATG-3' and CoxExon2R-5'-TCATATGGGCTACTGAGGAG-3'.

Rolling circle amplification. Circular viral DNA was preferentially enriched using rolling circle amplification (RCA) that employs phi29 DNA polymerase (Dean *et al.*, 2001), available in the Templiphi RCA kit, (GE Healthcare Bio-Sciences Corp, NJ, USA), according to the manufacturer's instructions with modifications, as previously described (Rector *et al.*, 2004;

Stevens *et al.*, 2010). The RCA reactions contained 2µL purified DNA in 10 µL of sample buffer. The preparation was denatured at 95°C for 3 min, and cooled on ice for 3 min. The RCA enzyme mixture, which contained 10 µL of the reaction buffer, 0.4 µL of the Templiphi enzyme mix, and 450 µM of dNTP mix (Sigma-Aldrich, MO, USA), was added to the denatured DNA. The reaction mixture was incubated at 30°C for 18 hrs, followed by enzyme deactivation at 65°C for 15 min. After cooling to 4°C, the product was used as template for PCR amplification of CSSV genomic fragments.

***Polymerase chain reaction*.** The reaction volume for PCR was 25 µL and contained 10.5 µL nuclease-free water, 12.5 µL of RedTAQ mix (Sigma-Aldrich, St. Louis, MO, USA), 0.2 µM final concentration of primers, and 1 µL of RCA template. The PCR reaction parameters were: 2 min initial denaturation step followed by 35 cycles each with 94°C for 30s, annealing temperature for 30s, and extension at 72°C for 1 min/kb, and a final extension at 72°C for 10 min. Amplification using the Badnavirus primers was carried out, as previously described (Yang *et al.*, 2003). The PCR products were fractionated by electrophoresis on a 0.8% agarose gel (90 min, 100 V) stained with GelRed (10 µL/mL) stain (Biotium, Aurora, CO, USA), in 1X Tris-acetate EDTA (TAE) buffer, pH 8.0.

Cloning and DNA sequencing of amplicons. The PCR products with the expected size were ligated using the pGEM T-Easy vector cloning kit (Promega) and transformed into *Escherichia coli* DH5α bacterial cells, according to the manufacturer's instructions. Insert sizes were confirmed by diluting selected white colonies in 20µL nuclease free water with 2µL as the template for PCR. The PCR reaction consisted of 1X PCR buffer and 0.5 U Platinum Taq polymerase (Invitrogen), 0.2 mM dNTP mix (Sigma-Aldrich, St. Louis, MO, USA), 0.2 µM each of M13 forward and reverse primers, and nuclease- free water to 50µL. The resultant products were separated by agarose gel electrophoresis, as previously outlined. The clones with the expected insert size were selected, two for each sample, and subjected to bi-directional capillary (Sanger) DNA sequencing at the University of Arizona Genetics Core (UAGC) sequencing facility (Tucson, AZ).

DNA sequence assembly and alignment. The DNA sequences were assembled using the DNASTAR SeqMan Pro Software Lasergene version 11 package (DNASTAR Inc., Madison, WI) and annotated using the BLAST2GO software (Conesa & Gotz, 2008). The CSSV sequences were aligned using MUSCLE software (Edgar, 2004) implemented in CLC Sequence viewer version 7.5. To reduce the number of sequences for analyses, identical sequences haplotypes were removed from the alignment using FaBox v1.41 software (Villesen, 2007), leaving one representative sequence of each to include in the phylogenetic analyses.

Pairwise nucleotide comparisons and phylogenetic analysis. To determine the pairwise nt identities, the aligned PCR fragments were analyzed using the Sequence demarcation tool (SDTv1.2) (Muhire *et al.*, 2014). The RT-RNase H region (580 bp) is delimited by primers specific to members of the badnavirus genus for which sequences were available at the time (Yang *et al.*, 2003), and the region represents the partial genome locus accepted for species demarcation of badnaviruses by the ICTV (http://www.ictvonline.org) (King *et al.*, 2012). The matrices obtained using SDT for this region were used for grouping sequences that shared at least 80% nt identity into species. For phylogenetic analyses, the Maximum Likelihood (ML) algorithm implemented in MEGA6 (Tamura *et al.*, 2013) was used to reconstruct a tree for each sequence alignment at 1000 bootstrap iterations. The best nt substitution model was predicted by the MEGA6 model prediction feature for each of the eight sequence alignments. The phylogenetic trees were viewed in MEGA6 using the default settings (Tamura *et al.*, 2013).

Results

Frequency of CSSV detection

A total of 124 plant samples collected from Cote d'Ivoire were amplified using eight primers, and CSSV was detected in 69 samples. However, the primers showed variable amplification depending on the sample used and, as a result, they had different amplification frequencies. The samples were amplified by a different combination of primers, from one to eight, and some primers failed to amplify CSSV in samples from which the other primers had amplified. Table 2.2 summarizes the frequency of each primer pair at detection of CSSV from all the 124 samples

and from the 69 confirmed to have CSSV sequences. The CSSV sequences were amplified from 24 to 52 per primer pair, representing approximately 19 to 42% of 124 tested samples, or 35 to 75% of the confirmed 69 CSSV positive samples, respectively.

The P4 primer pair had the highest amplification frequency at 42% (52/124), and it targeted the non-coding intergenic region on the CSSV genome, yielding the expected size of 1,123 bp. The RT primer pair had an expected size of 421 bp and achieved the second highest amplification frequency at 37% (46/124). The 421 bp region corresponds to the conserved reverse transcriptase coding region on the CSSV ORF3.

The previously published ORF3A primer pair, and the P3 primer pair, both amplified CSSV at 33% (41/124) frequency. The ORF3A set targeted part of the movement protein gene on the ORF3, with an expected size of 532 bp, while the P3 primer pair also targeted the ORF3, specifically, the pepsin-like aspartate protease. The Badnavirus primers originally designed to be universal primers based on the sequences in the database at that time amplified CSSV at a frequency of 28% (35/124). The 577 bp amplicon, which corresponds to a region encoding portions of the RT and RNase H (RT-RNase H) protein, is the region used for taxonomic classification of the genus, *Badnavirus* (King *et al.*, 2012).

Relatively low amplification frequencies were observed for three primer pairs, P2, P1, and ORF1 at 27% (34/124), 23% (29/124), and 19% (24/124), respectively. The P1 and P2 primer pairs targeted the region at the 5' and 3' of the ORF3A primers, and yielded expected sizes of 774 bp and 804 bp, respectively. The region amplified by the P1 primer pair overlapped with the region coding for the MP, while the P2 amplified the region with no specific functional domain, as predicted by the NCBI Conserved Domain Database (CDD) (Marchler-Bauer *et al.*, 2015). The ORF1 primers targeted a 375 region in ORF1 that contains a badnavirus-specific domain of unknown function, DUF1319.

BLASTn search

The BLASTn search of the sequences available in the GenBank database with the putative CSSV sequences obtained for the 69 samples yielded matches with e-values of zero, and sequences that

had at least an 88% nt similarity score to the seven CSSV reference genomes. However, five sequences that were amplified using the Badnavirus, ORF3A and P3 primers, consistently showed low percentage similarities, at 69 – 76%, with high e-values (8e-61 to 1e-102) and sequence coverage of 30 to 70%. These results were observed for the three primer pairs, and the results were consistent after repeating DNA sequencing on duplicate DNA extracts.

Among the CSSV sequences amplified by the eight primers, a few were recovered from non-cacao samples, representing six plant species in total. The plant species and the primers were for *Carica papaya* (Badnavirus, and P4, and RT primers), *Ceiba pentandra* (all primers but P1), *Commelina erecta* (Badnavirus primers), *Dioscorea cayenensis* (P4 primers), *Spigelia anthelmia* (P4 primers), *Tapinanthus bangwensis* (P4 primers), and *Xanthosoma maffafa* (ORF3A and P4 primers).

All the primers used for PCR amplification of CSSV, with the exception of ORF1, also amplified non-specific products, mostly plant host genomic DNA on different loci, in addition to CSSV sequences. The BLASTn search results showed that most of the sequences had the highest similarity to GenBank sequences labeled as *T. cacao* uncharacterized protein, whereas, only less than 10% of the sequences shared high similarity with sequences encoding specific cacao proteins, among which were the DNA/RNA polymerase superfamily, receptor kinase, retrotransposon, transport protein, reverse transcriptase and gag-pro-like proteins. Results of PCR amplification indicated that the general Badnavirus primers amplified the largest number of products containing host plant genomic sequences.

In addition, the general Badnavirus primers amplified six other badnavirus sequences, representing two species from six samples. Four of the six samples were also found to contain CSSV using all primers except the P1 pair, indicating mixed infections. Two isolates of *Dioscorea bacilliform virus* (DBV) and one isolate of *Banana streak Uganda E virus* (BSUEV) were amplified by the general Badnavirus primers. The BLASTn similarity search results indicated a similarity score of 73 – 82% for DBV and 73% for BSUEV over a range of e-values from 1e-53 to 6e-51 for the six sequences. The coverage for the six sequences ranged from 48 to

74% of the 577 bp fragment used for the search. The BSUEV was amplified was *Tapinanthus bangwensis*, and DBV was amplified from *Xanthosoma maffafa* and *Dioscorea cayenensis*. To determine if the variable or absence of amplification by the eight primers was related to the quality of total DNA, eight representative samples with no amplification, or those that were amplified by only one or two CSSV primers, were amplified by PCR using the CoxExon1F/2R primers. The eight selected samples yielded expected PCR products of 2 kbp in size, and they confirmed that the total DNA was amplifiable by PCR (Fig. 2.3).

Pairwise nucleotide comparisons

The shared pairwise nt identities were determined among the CSSV sequences amplified by each of the eight primers using the SDT. The analyses indicated that the sequences from each primer formed between two and five SDT groups. The nt sequences amplified by the P1, P4, and ORF1 primers were the least divergent, sharing 74 – 99% nt identity each among the sequences within each set (not shown). The sequences from the three sets formed two SDT groups each, based on analysis of 63, 98, and 53 clone sequences for the P1, P4, and ORF1 sets, respectively. The results suggest that the regions delimited by the P1, P4, and ORF1 primers are near equally divergent. However, the PCR primers that were designed to amplify each of the target regions amplified their respective viral targets at different frequencies (Table 2.2).

The nt sequences amplified by the Badnavirus and the ORF3A primers were equally divergent and shared nt identities at 68 – 99% among 87 and 85 clone sequences, respectively. The sequences from both regions formed four SDT groups each, based on the 80% species threshold. A matrix generated by SDT showing the nt identities (compressed to show all sequences) is shown for the Badnavirus- and ORF3A-amplified sequences in Fig. 2.7 and 2.8. The RT-RNase H region corresponding to the region delimited by the Badnavirus primer pair is used for the genus, *Badnavirus* taxonomy, therefore the four SDT groups represent four putative CSSV species.

The viral sequences amplified by the P2 and the RT primers formed two separate groups whose members shared 71–99% and 73-99% nt identities among the 68 and 77 cloned sequences, respectively. The most divergent region was that amplified by P3 primers, with 66-99% nt

identities shared among the 82 sequences analyzed. The sequences amplified by the P3 primers formed five groups, based on the ICTV 80% species threshold.

Phylogenetic analysis

To determine the phylogenetic relationships among the CSSV sequences amplified by the eight primers, phylogenetic trees were constructed from each of the data sets. The eight trees showed at least two major clades, into which the seven reference sequences from GenBank were grouped, and each clade was statistically supported at >70%. The five phylogenetic trees reconstructed from the sequences amplified by the ORF1, P1, P2, P4 and RT primers were similar in that they resolved two major clades, Clades I and II (Fig. 2.4). For each tree, Clade I consisted of the two subclades which appeared to contain sequences based on the geographic locations from which the GenBank reference isolates where collected. One of the subclades contained the majority of the sequences determined in this study, and the CSSV GenBank reference sequences from Ghana (Accession numbers, AJ608931, AJ609019, and AJ609020) (Fig. 2.4). The second subclade contained two isolates from Togo (Accession numbers, AJ534983 and L14546) and one or no sequences determined from this study (Fig. 2.4). The smaller of the two clades, Clade II, consisted of sequences from isolates collected from Cote d'Ivoire and the GenBank reference sequence from the same country (Accession JN606110). Less than one third of the total sequences often clustered in Clade II, which is interesting because all the samples were collected from Cote d'Ivoire. One divergent sequence, GenBank reference sequence (Accession AJ781003), originally isolated from Togo, was frequently found in Clade II, forming a sister subclade to the Cote d'Ivoire sequences.

The three phylogenetic trees reconstructed based on the sequences amplified by the Badnavirus, ORF3A, and P3 primers were similar to the five described for the ORF1, P1, P2, P4, and RT trees, but they were different in that, a third clade (Clade III) was observed on each. The phylogenetic trees for the sequences amplified using the Badnavirus and ORF3A primers are shown in Fig. 2.5 and 2.6, respectively. For each of the three trees, Clade III contained 11 sequences, or less, from five samples. The same sequences had showed low similarity and high e-values for BLASTn searches when compared to previously published CSSV reference sequences. Clade III for the P3 tree (not shown) and the Badnavirus tree (Fig. 2.5) did not

contain partial or complete GenBank reference sequences, but only the sequences determined in this study. However, the ORF3A Clade III contained four partial genomic GenBank reference sequences that were previously sequenced from isolates amplified using the ORF3A primers (Kouakou *et al.*, 2012), and they were classified as Group E and F isolates (Fig. 2.6). None of the complete genome GenBank reference sequences were found in Clade III. The sequences that grouped in Clade III were amplified from five samples, and the Badnavirus, ORF3A and P3 primers amplified CSSV from one, four and three samples, respectively. One of the samples, R295 was *X. maffafa* while the other four were cacao plants. Compared to the number of isolates in Clades I and II, the Clade III-type sequences were rare. For a point of reference, the isolates that grouped in clades I and II have been labeled in another study as Group A, B, and D isolates (Kouakou *et al.*, 2012) (Fig. 2.6). Also, certain of the CSSV sequences were observed to shift between Clades I, II or III, depending on the PCR primer pair used for amplification, suggesting that the clades based on sequences from loci are not congruent. The reference sequence from Togo (Accession AJ781003) shifted clades frequently from Clades I to II, or being basal to the two, but it was never part of Clade III (Fig. 2.4, 2.5 and 2.6). Similarly, sequences from different clones amplified from the one sample were found to group separately, or to shift to another clade when a different region was analyzed.

Discussion

Five PCR primer pairs were designed and tested for the ability to amplify CSSV sequences in cacao and non-cacao samples. Three primer pairs used previously to amplify CSSV-like isolates (Kouakou *et al.*, 2012; Oro *et al.*, 2012a; Yang *et al.*, 2003) were included as internal controls. Collectively the frequency of amplification ranged from 19- 42% (Table 2.2). Although six of the PCR primer pairs were degenerate, no single degenerate or sequence-specific primer pair detected CSSV in all of the samples, and 55 of the 124 samples were not positive by PCR amplification using any of the primer pairs, despite having characteristic CSSV-like symptoms. Among the 55 samples, eight were selected for PCR amplification of the non-coding plant mitochondrial DNA to determine the quality of the total DNA. The mitochondrial oxidase gene (Demesure *et al.*, 1995) was amplified from the eight samples, with only one sample showing

low concentration of PCR product compared to other samples (Fig. 2.3). Although not all 55 samples were tested because the DNA had been used for Illumina sequencing (Chapter 3), the results suggest that the genome sequence of CSSV-like isolates is highly divergent, as has been shown for other badnavirus species (Geering *et al.*, 2000; Harper *et al.*, 2005; Hoffmann *et al.*, 1997). Alternatively, the symptomatic plant samples had low level of CSSV titers, or they may not have been infected by CSSV.

The five primer pairs were designed based on previously sequenced genomes and failure to detect CSSV isolates suggests that virus populations may have changed since the first time the currently published CSSV isolates were sampled. The five primers detected more of the sequences related to Ghana reference sequences than from Togo or Cote d'Ivoire, suggesting that either more of Ghana-like isolates were present in the samples, or that the primers are limited in the number of isolates they can amplify from isolates present in other countries. Testing plant material collected from other cacao-growing countries in West Africa will confirm these speculations. The eight primers, however, were useful for determination of the genetic diversity of the corresponding eight regions. The analysis indicated that CSSV isolates have extensive genetic variability at the eight regions, and they shared 66 – 99% nt identity.

The P4 primers amplified the most CSSV sequences from the 124 tested samples, representing approximately 42% of the total, but they failed to amplify sequences from some isolates, which were amplified by other primers. In contrast, the Badnavirus, ORF1 and P2 primers showed the lowest PCR frequencies of amplification, at 28, 27, and 33%, respectively. Among the 69 samples with identified CSSV sequences, the eight primers failed, at least once, to detect the virus. The failure rates ranged from 25% for the P4 primers to 75% for the ORF1 primers. These results indicate that, although six of the primers had been designed with some degree of nt degeneracy, the primers remained specific for a certain group of isolates. Variable amplification of CSSV isolates has also been reported elsewhere (Kouakou *et al.*, 2012; Oro *et al.*, 2012a). Because the eight primers were designed based on the CSSV GenBank references, the failure to amplify all isolates suggests that there is more genetic variability among isolates than has been described.

The CSSV genome showed a wide range of nt diversity, based on pairwise nt identities of the regions amplified by the eight primers. The shared pairwise nt identities from SDT analysis for the eight regions, ranged from 66 % to 99 %. The region amplified by the P3 primer was the most divergent region, at 66-99 %, followed by the ORF3A- and Badnavirus-amplified regions, at 68-99 %. All the other primers were not as divergent and the shared pairwise identities were 71 – 99 %, combined. The sequences amplified by the P3 primers encode the aspartic protease, the enzyme that processes the polyprotein into mature proteins with individual functions. Proteases provide crucial functions during the virus life cycle and so would not be expected to be free to vary substantially. Studies on the structure of proteases from other organisms showed that the carboxyl terminus of the enzymes were of different orientations, however, the internal structure remained conserved (Dunn, 2010). Similarly, some regions of high structural variability were found on the surfaces of the protein, but were found not to affect the rate of cleavage or substrate specificity of the enzymes. The divergent nt sequences determined form P3 amplification in this study might reflect such variability.

The least divergent sequences were from regions amplified by the P1, P4, and ORF4 primers. The functions for the proteins encoded by ORF1 and P1 are not known. The ORF1 contains an unknown functional domain while no such domain is present on the P1 region. Because the three regions are conserved, they are most likely to perform a crucial function in the life cycle of the virus. The region amplified by the P4 primers contains promoters and important badnavirus regulatory sequences like the conserved $tRNA^{met}$ primer binding site that serves as a reverse transcription start site, and also the TATA boxes and poly A signal, both of which are important for the production of terminally redundant RNA transcripts (Medberry *et al.*, 1990b; Zhang *et al.*, 2015). This region is therefore not expected to undergo mutations that cause significant nt or amino acid changes.

High variability at the nt level was also observed in region amplified by the ORF3A and the Badnavirus primers, which amplified sequences that code for a partial movement protein and RT-RNase H, respectively. The Badnavirus primers were originally designed to amplify species from the badnavirus genus, and a high amplification frequency was observed (Yang *et al.*, 2003). The Badnavirus primers were published when only one complete CSSV isolate sequence was

available in the database, and this may be one reason why CSSV isolates from 89 of 124 samples were not detected using the primers used in this study. The CSSV-like sequence region amplified by the ORF3A primers, which encode the viral MP, have been shown to be highly divergent, compared to the badnavirus taxonomically informative RT-RNase region (King *et al.*, 2012), and has been reported to resolve six or more divergent subgroups of CSSV-like isolates from Cote d'Ivoire and Ghana (Kouakou *et al.*, 2012).

Four SDT groups, representing four distinct CSSV species, based on the ICTV-approved 80% nt identity species demarcation threshold for the RT-RNase H region (delimited by the Badnavirus primers) were observed (Fig. 2.7). These results suggest that CSSV exists as a complex of species, such as that described for *Banana streak virus* (Harper *et al.*, 2005). The Badnavirus primers detected CSSV isolates from only 35 of the 124 samples, and compared to other primers, showed extremely low PCR amplification frequency. It is possible that the undetected sequences may represent additional CSSV species.

A comparison between the SDT groups based on the taxonomic RT-RNase H and on the other seven regions revealed that, while the RT-RNase H consisted of four groups, the seven regions comprised between two and five groups. The number of sequences, including GenBank references, in each of the four RT-RNase H group also varied when compared to the corresponding groups based on the sequences from other regions. While the results from this study showed that all of the targeted regions are useful for evaluation of genomic diversity and CSSV detection, the variable amplification by primers indicated that using any of CSSV region, other than the RT-RNase H, for taxonomy and classification is likely to yield incongruent taxonomic groups. However, the Badnavirus primers do not detect all CSSV isolates, so an alternative to the taxonomic RT-RNase H region would be to use the complete genomic sequences. The advantage of using the complete genomic sequences is that they are more informative because both the coding and non-coding regions are taken into consideration.

The six degenerate CSSV primers amplified both plant genomic DNA, and CSSV sequences, even though at times the plant amplicons were of the expected size. The presence of products

from false positive amplification results also emphasizes the need to sequence PCR products to verify the presence of CSSV as part of the diagnostics routine.

Until CSSV primers with higher amplification frequencies are validated, such as those targeting specific CSSV species/groups, or broad range primers, extensive CSSV detection will rely on the use of multiple primers to eliminate the false negative results. Based on this study, the 'best' candidates for CSSV detection are the P4, RT, ORF3A (or P3), and a combination of two or more of these is likely to significantly improve the detection frequencies. The remainder of the four primers with low detection rates will also be useful because, in some cases, they detected CSSV where the 'best' candidates did not. Undoubtedly, adopting the use of separate PCR reactions with multiple primers is time consuming and financially constraining. Instead, optimization of multiplex PCR using any combinations of the primers designed herein will allow more efficient diagnosis than uniplex PCR. Alternatively, designing SDT group (species)-specific primers for each of the four putative species within the CSSV complex will allow for more specific questions to be answered.

Phylogenetic analyses resolved two major clades for the amplicons produced by the ORF1, P1, P2, P4, and RT primers (Fig. 2.4). In contrast, some sequences amplified by the ORF3A, P3, and Badnavirus primers grouped separately, and formed a third clade basal to the rest (Fig 2.5 and 2.6). At least two-thirds of the sequences in the eight phylogenetic trees grouped with Ghana and Togo reference sequences in Clade I, suggesting that the origin of most of the isolates was not Cote d'Ivoire. Clade I formed two subclades based on the geographical origin of the CSSV GenBank reference sequences, where those from Ghana grouped separately from those from Togo. The results may imply that, although the isolates from both countries form one major clade, they are divergent such that primers targeting all the sequences in the two subclades may fail to amplify the related CSSV sequences. This degree of divergence suggests that primers specific for subclades may be more effective than those targeting major clades. CSSV sequences were also amplified from six wild host species, three of which have been previously described (Posnette *et al.,* 1950). The CSSV sequence from the six wild host plants and those from cacao were grouped in the same clade, suggesting that they are phylogenetically related. This supports findings from previous studies that the wild hosts play a role in the transmission of CSSV, and

that CSSV was present in West Africa in wild hosts before cacao was introduced there in the early 1800's (Posnette *et al.*, 1950; Todd, 1951).

The sequences contained in Clade III were amplified by the Badnavirus and ORF3A (Fig. 2.5 and 2.6), and the P3 primers, and they contained highly variable CSSV sequences that shared <80% nt identity with the previously published sequences. The virus isolates from which these sequences were derived may have been present in cacao or wild hosts in geographical areas where surveys had not been conducted. Alternatively, they may have been a result of recombination between CSSV species, or between CSSV and unknown badnavirus species. No recombinant CSSV species have been described to date, but in this study there was evidence of the presence of recombinant isolates. The GenBank reference, Accession AJ781003, originally isolated from Togo, shifted from Clade I to II, and to being basal to the two clades. Additionally, it was found in different SDT groups, based on the primer used for amplification. A similar trend was observed for some isolates from this study, and it might be indicative of recombinant genomes or the presence of mixed/co-infecting isolates. Recombination analysis will elucidate the status of these putative recombinant sequences.

Overall, the results of this study indicate that what is considered to be a single virus, CSSV, exists instead as a highly divergent group of variants that will probably be found to represent a CSSV complex comprising at least four species. If this hypothesis is borne out, it is unlikely, based on the data reported herein that it will be feasible to design and use a single diagnostic primer pair to facilitate universal detection of all members of the putative species group. Thus, it is necessary to determine the full-length genome sequence for representative isolates of CSSV occurring in the cacao-producing countries in West Africa where the disease is known to occur, such that the full extent of CSSV variability can be determined.

Table 2.1. Primer pairs used for PCR-amplification of eight *Cacao swollen shoot virus* genomic regions. The primer coordinates are based on the CSSV reference sequence, Genbank Accession number NC_001574.1.

Primers forward (F), reverse (R)	Primer pairs 5'– 3' orientation	Primer coordinates	Primer Tm °C	Expected size (bp)
RT_F	AACGACAACACTGAAAAGGA	5325 - 5344	56	421
RT_R	GTGCCCAAAAATTCAATCTC	5727 - 5746	56	
ORF3A_F	GTYRTACCRRAYAYYATGATGAC	1848 - 1870	55	532
ORF3A_R	GTTYCCRTTRSYRGAYTCYTCCCATAC	2355 – 2380	55	
ORF1_F	AACCTTGAGTACCTTGACCT	498 - 517	56	375
ORF1_R	TCATTGACCAACCCACTGGTCAAG	849 - 872	56	
P1_F	RGCAGCWGAARWGGCWAAGR	1244 - 1263	56	774
P1-R	TTDGGWGTRTTKGAYARYCK	1999 – 2018	56	
P2_F	ACDGGHTGGGRMRRYGATRA	2461 – 2480	56	804
P2_R	TRTCYTTKATRTTGTKGCADGT	3244 – 3265	56	
P3_F	ATNMHRGTCCARCAGCAGCC	4089 – 4108	56	1042
P3_R	TTDATGGGCTTRTCTTCWAT	5112 – 5131	56	
P4_F	TGGCAACDGAACATGCCATCTC	6585 – 6606	56	1123
P4_R	TGGTTGTTGGTCACTTTACT	528 - 547	56	
BadnavirusFP	ATGCCITTYGGIAARAAYGCICC	5513 - 5536	50	577
BadnavirusRP	CCAYTTRCAIACISCICCCCAICC	6066 - 6090	50	

Table 2.2. Frequency of polymerase chain reaction amplification of a fragment of the *Cacao swollen shoot virus* (CSSV) genome using the eight primer pairs. Panel I shows the number of samples amplified (and the percentage) from the 69 samples that had been confirmed to contain CSSV sequences. Panel II shows the number of samples amplified (and the percentage) from the 124 total samples tested.

	RT	ORF3A	ORF1	P1	P2	P3	P4	Badnavirus
I	46/69 (67%)	41/69 (59%)	24/69 (35%)	29/69 42%	34/69 49%	41/69 59%	52/69 75%	35/69 51%
II	46/124 (37%)	41/124 (33%)	24/124 (19%)	29/124 (23%)	34/124 (27%)	41/124 (33%)	52/124 (42%)	35/124 (28%)

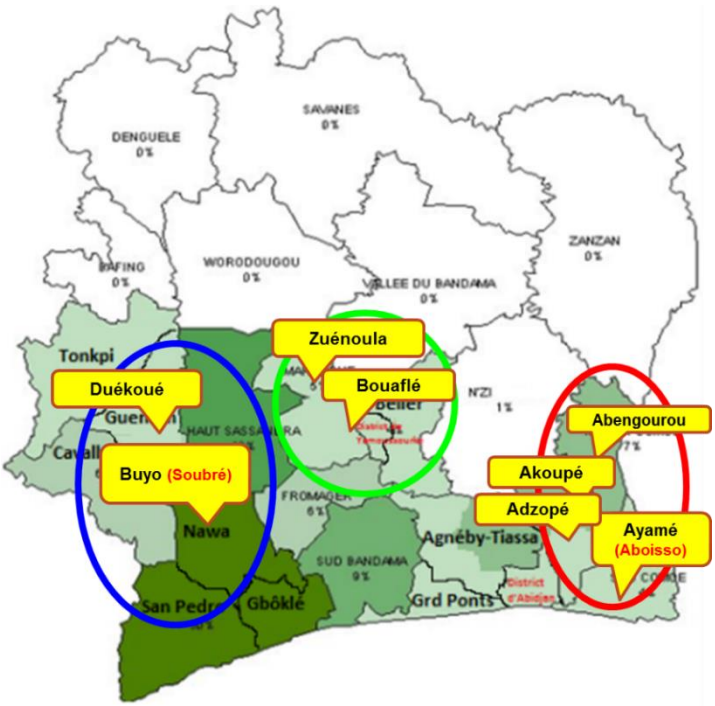

Fig. 2.1. Map of Cote d'Ivoire showing the cacao growing regions and specific sites from where the plant samples used in this study were collected. The three regions designated as western, central and eastern are indicated in blue, green and red, respectively.

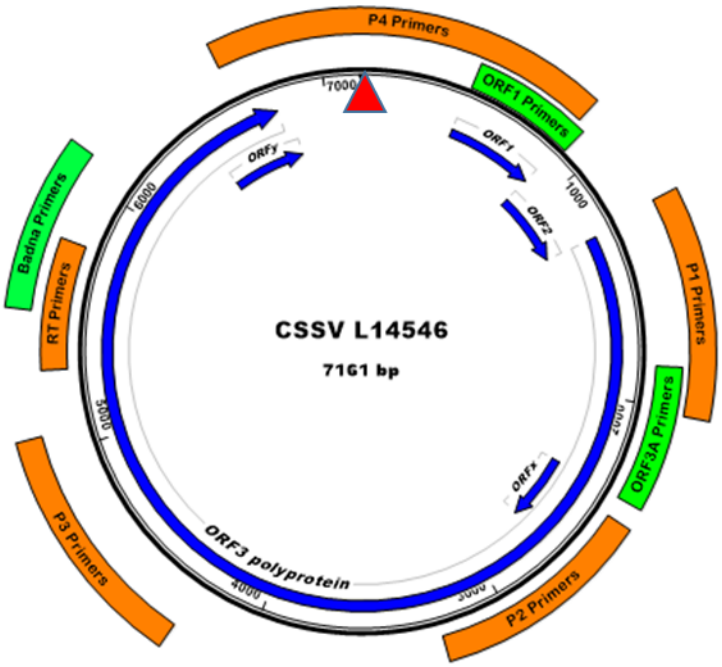

Fig. 2.2. The genome map of *Cacao swollen shoot virus* (CSSV) showing the location of each primer pair used for polymerase chain reaction amplification in relation to the GenBank CSSV reference sequence, Accession number NC_001574. Previously published primers, Badna (Yang *et al.*, 2003), ORF1 (Quainoo *et al.*, 2008b) and ORF3A (Kouakou *et al.*, 2012) amplify regions designated as green blocks, and those designed in this study amplify regions designated as orange blocks. The first nucleotide is indicated by the red arrow.

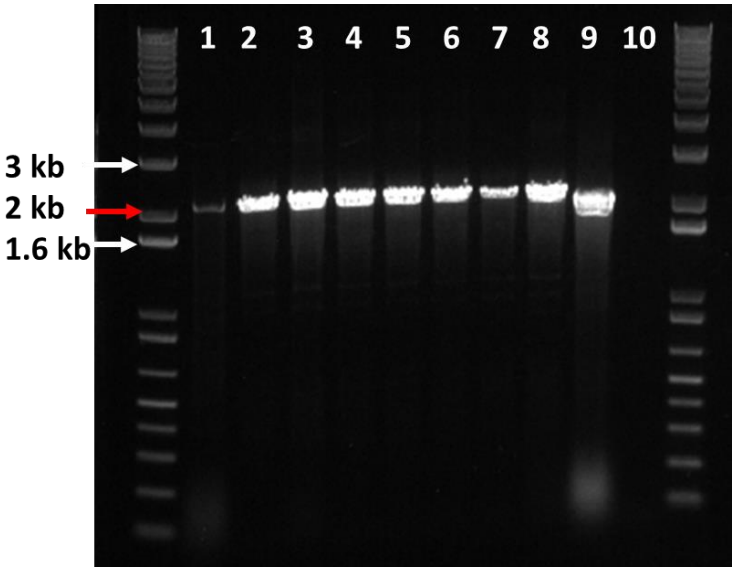

Fig. 2.3. Polymerase chain reaction (PCR) results from amplification of total DNA from cacao and non-cacao plant samples to determine the quality of the DNA. Total DNA from selected samples for which one, two or no *Cacao swollen shoot virus* (CSSV) primers amplified CSSV sequences was used as the template. The PCR was carried out using the CoxExon1F/2R (Demesure *et al.*, 1995) primers dessigned to amplify a part of the plant mitochondrial oxidse gene. Lanes 1 – 3: samples in which no primer amplified CSSV, lanes 4 – 7: samples in which only one primer amplified CSSV, lane 8: sample in which two primers amplified CSSV, lane 9: positive control cotton plant DNA, lane 10: negative no template control. A 1 kb DNA ladder was included on the first and last lanes, and the expected size (2 kbp) is indicated by a red arrow.

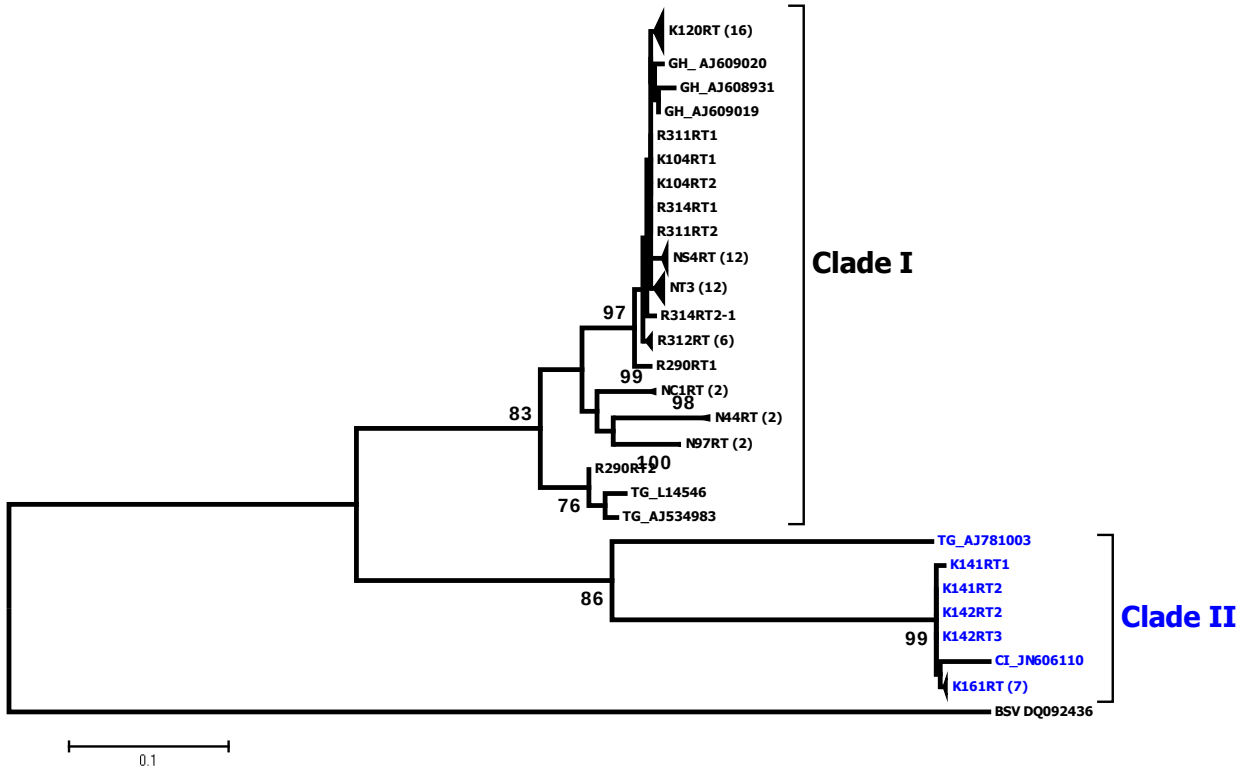

Fig. 2.4. Phylogenetic tree of the sequences coding for the reverse transcriptase (RT) region of *Cacao swollen shoot virus* (CSSV), corresponding to the region delimited by the RT primers (Table 1), using the Maximum Likelihood method (1000 bootstrap replications) implemented in MEGA6 (Tamura *et al.*, 2013). Bootstrap values greater than 70% are shown at the statistically supported clades. The CSSV reference sequences downloaded from the NCBI- GenBank database are indicated with Accession numbers. The numbers in parentheses represent the number of sequences collapsed at each taxon name. The *Banana streak virus* RT-RNase H region (GenBank Accession number DQ092436) was used as the outgroup sequence. The letter codes CI (Cote d'Ivoire), GH (Ghana), or TG (Togo) indicate the country from which the cacao samples were collected.

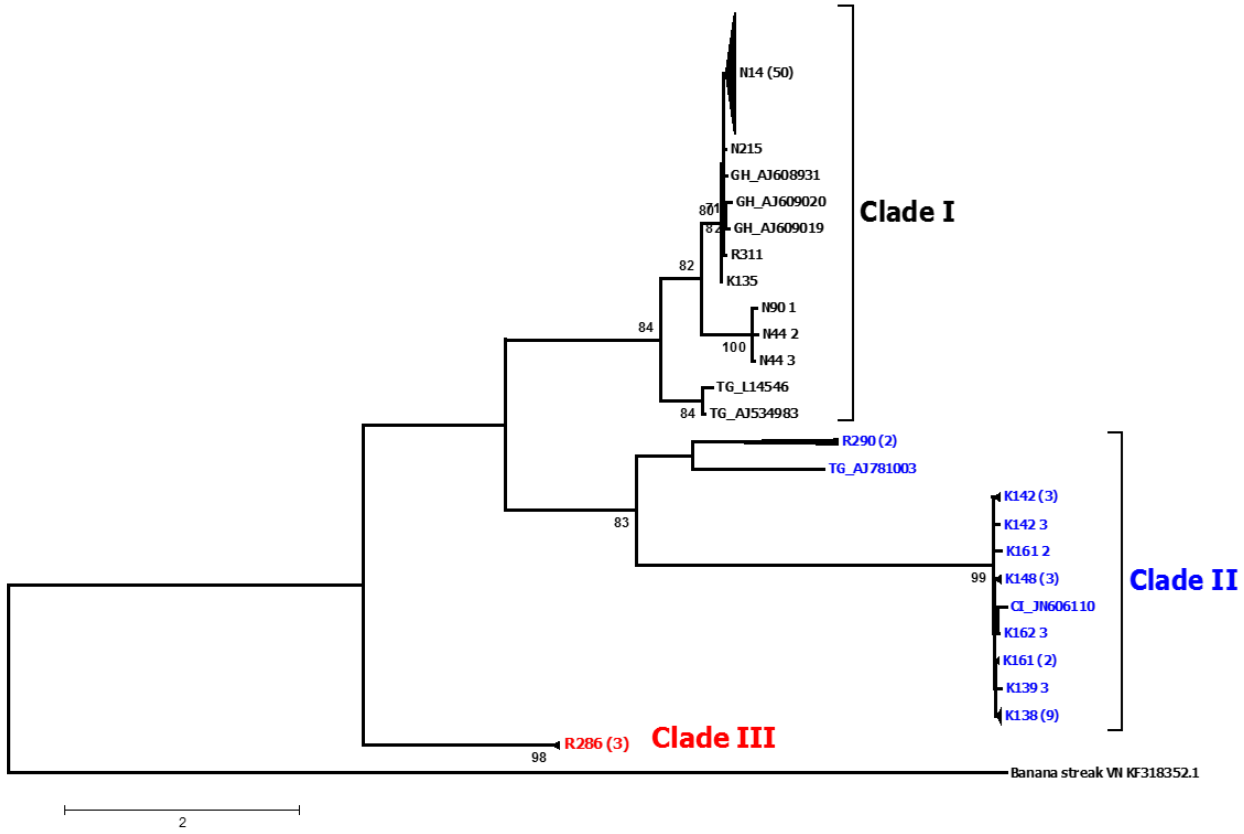

Fig. 2.5. Phylogenetic tree of the RT-RNase H sequence of *Cacao swollen shoot virus* (CSSV), corresponding to the region delimited by the Badnavirus primers (Yang *et al.*, 2003), the using the Maximum Likelihood method (1000 bootstrap replications) implemented in MEGA6 (Tamura *et al.*, 2013). Bootstrap values greater than 70% are shown at the statistically supported clades. The CSSV reference sequences downloaded from the NCBI- GenBank database are indicated with Accession numbers. The numbers in parentheses represent the number of sequences collapsed at each taxon name. The *Banana streak virus* RT-RNase H region (GenBank Accession number KF318352.1) was used as the outgroup sequence. The letter codes CI (Cote d'Ivoire), GH (Ghana), or TG (Togo) indicate the country from which the cacao samples were collected.

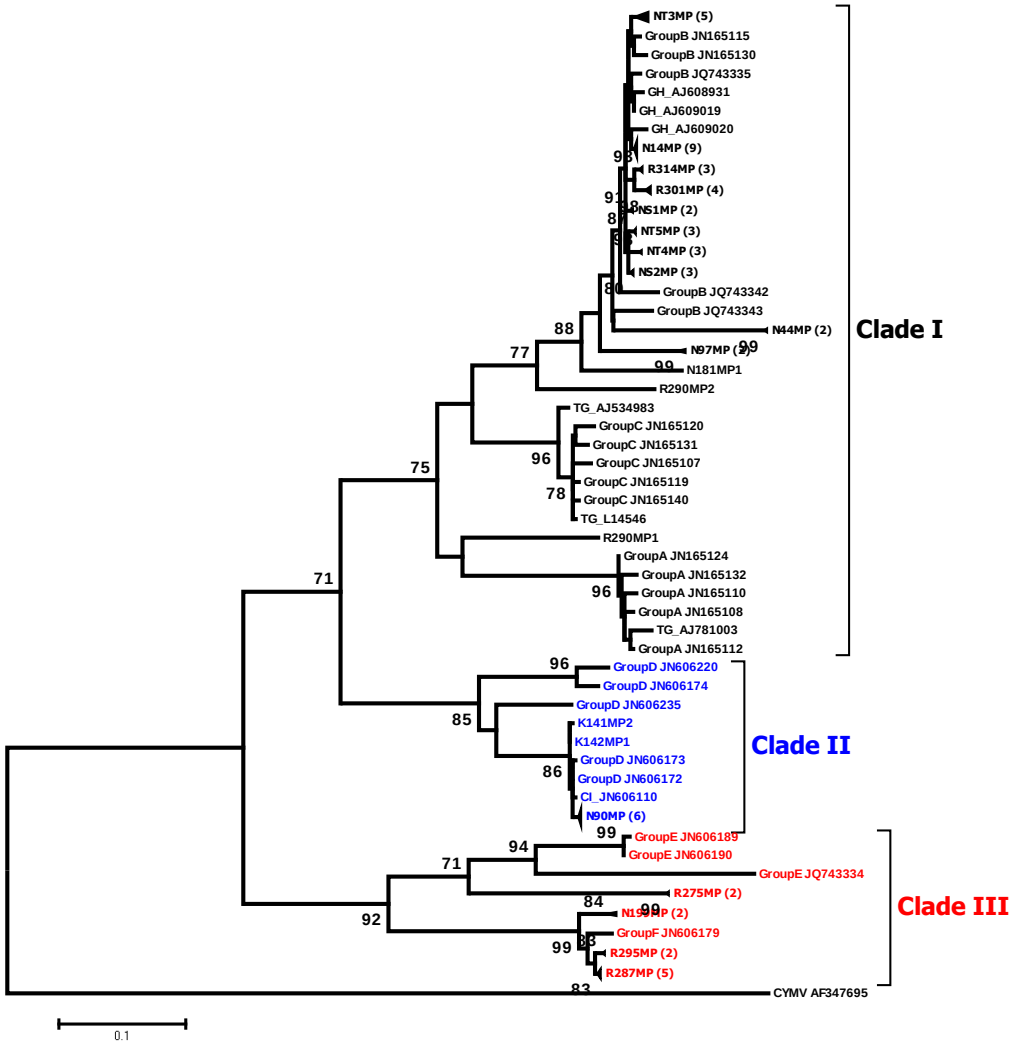

Fig. 2.6. Phylogenetic tree of the 5' end of the open reading frame 3 of *Cacao swollen shoot virus* (CSSV), corresponding to the region delimited by the ORF3A primers, encoding the viral movement protein (Kouakou *et al.*, 2012). The Maximum Likelihood method (1000 bootstrap replications) implemented in MEGA6 (Tamura *et al.*, 2013) was used. Bootstrap values greater than 70% are shown at the statistically supported clades. The CSSV reference sequences downloaded from the NCBI- GenBank database are indicated with Accession numbers. The Group A – F labels were adapted from (Kouakou *et al.*, 2012) and were used to compare to the current grouping system. The numbers in parentheses represent the number of sequences collapsed at each taxon name. The analogous region of the *Citrus yellow mosaic virus* (Genbank Accession number AF347695) genome was used as the outgroup sequence. The letter codes CI (Cote d'Ivoire), GH (Ghana), or TG (Togo) indicate the country from which the cacao samples were collected.

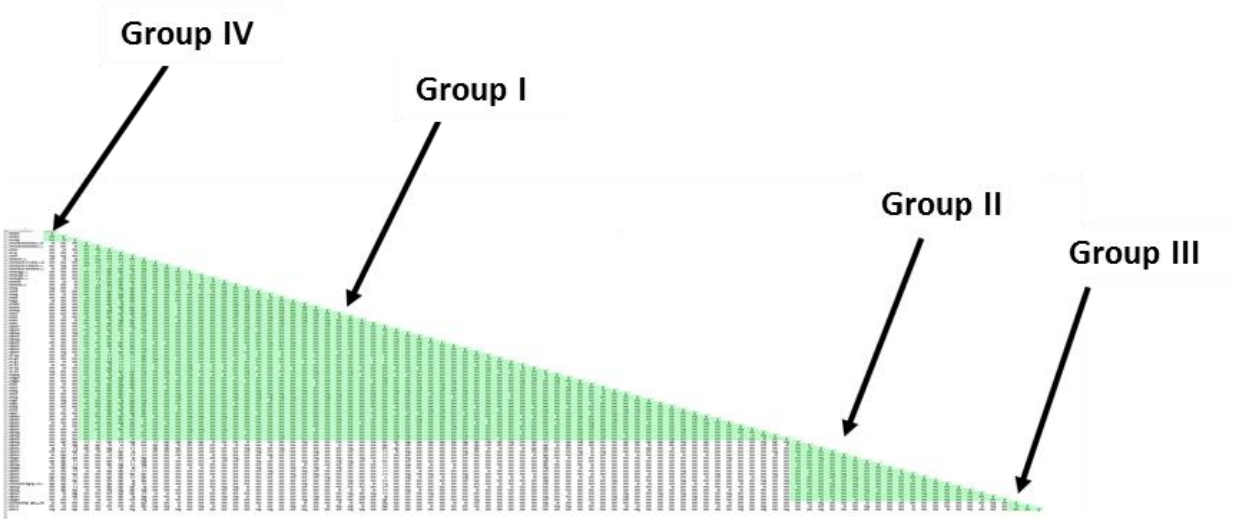

Fig. 2.7. Percentage pairwise nucleotide (nt) identity for the RT-RNase H region of *Cacao swollen shoot virus* (CSSV) sequences, delimited by the Badnavirus primers (Yang *et al.*, 2003), in relation to the CSSV GenBank reference sequences using MUSCLE alignment implemented in the Sequence demarcation tool (SDTv1.2) software (Muhire *et al.*, 2014). The green shaded shaded box indicates isolates that share greater than or equal to 80% nt identity, the threshold used for species demarcation, approved by the International Committee on Taxonomy of Viruses. Group I includes CSSV reference sequences assigned the GenBank Accession number AJ609019, AJ608931, AJ609020, L14546, and AJ534983. Groups II and III contain sequences assigned the Accession number AJ781003 and JN606110. Group IV contains a group of unique CSSV variants detected for the first time in this study.

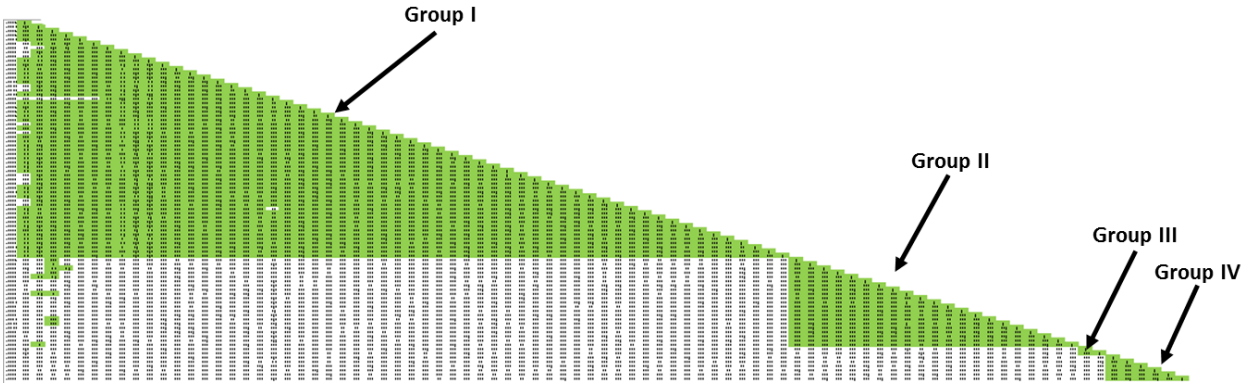

Fig. 2.8. Percentage pairwise nucleotide (nt) identity for 5' end of the open reading frame 3 of *Cacao swollen shoot virus* (CSSV), corresponding to the region delimited by the ORF3A primers (Kouakou *et al.*, 2012) and encoding the viral movement protein, in relation to the CSSV GenBank reference sequences. The MUSCLE alignment implemented in the Sequence demarcation tool (SDTv1.2) software (Muhire *et al.*, 2014) was used. The green shaded box indicates isolates that share greater than or equal to 80% nt identity. Group I consists of CSSV reference sequences assigned the GenBank Accession number AJ609019, AJ608931, AJ609020, L14546, AJ534983, and AJ781003. Group II sequences have as their closest relatives the previously described variants Accession number JN606110. Groups III and IV are unique CSSV variants detected for the first time in this study.

CHAPTER 3

THE *CACAO SWOLLEN SHOOT VIRUS* COMPLEX IN WEST AFRICA COMPRISES FOUR DIVERGENT SPECIES THAT VARY BY GENOME ARRANGEMENT AND CONSERVED PROTEIN DOMAINS

Abstract

The production of cacao (*Theobroma cacao*) in West Africa, which supplies over 70% of the world's bulk cocoa, is threatened by *Cacao swollen shoot virus* (CSSV) (*Caulimoviridae, Badnavirus*) infection. The CSSV is endemic to West Africa where it has caused swollen shoot disease of cacao there since 1936. Recent outbreaks beginning in Ghana and Cote d'Ivoire during 2000-2003, are characterized by atypical rapid tree decline symptoms together with shoot swelling and reduced pod set in cacao trees. Serological and molecular methods that previously detected CSSV have proven ineffective, suggesting the emergence of new or previously uncharacterized CSSV-like variants. The genome variability among extant CSSV isolates was investigated for symptomatic cacao samples from Cote d'Ivoire and Ghana using the Illumina HiSeq platform, with validation by Sanger DNA sequencing. Pairwise nucleotide analysis of apparently full length 87 badnavirus genome sequences (including fourteen determined herein), and of the taxonomically informative RT-RNase H coding region, overall distinguished four groups based on the 80% species demarcation approved by the International Committee on Taxonomy of Viruses. Phylogenetic analysis using Maximum Likelihood (70% bootstrap) of the RT-RNase H region resolved several unsupported clades, whereas, the full-length badnavirus genome sequences resolved three well-supported clades, indicating that the full-length viral genome sequence is a robust predictor of evolutionary relationships. The CSSV subclade was further distinguished by having isolates with a genome arrangement consisting of four, five, or six open reading frames (ORFs). Four full-length genome sequences of Cote d'Ivoire and Ghana isolates contained four ORFs, and grouped together in a previously undescribed clade, basal to the two clades harboring previously recognized CSSV isolates, having either five or six ORFs, respectively. The discovery of a third CSSV-like clade of viruses suggests a possible association between the new CSSV-genome type and the rapid decline symptoms observed for the first time in Ghana during 2000 and is spreading in West Africa.

Introduction

Theobroma cacao L. (cacao) produces cocoa beans, an important commodity worldwide in the confectionary industry. Production is mainly centered in West Africa (Cote d'Ivoire, Ghana, Nigeria and Cameroon) where over 70% of the world's supply is produced. Cocoa production is threatened by several pests and diseases (Ploetz, 2006), with one of the greatest threats to West Africa being the *Cacao swollen shoot virus* (CSSV) [*Caulimoviridae* family; *Badnavirus* genus] that causes the swollen shoot disease (Posnette, 1947; Thresh, 1958). The CSSV virus is transmitted in a semi-persistent manner by approximately 20 mealybug species (Dzahini-Obiatey & Fox, 2010), however, it is not transmitted through pollen or seed.

Since swollen shoot disease was first described in Ghana during 1936 (Steven, 1936), it has been reported in the nearby cocoa-producing West African countries, including Cote d'Ivoire (Mangenot *et al.*, 1946; Muller & Sackey, 2005), Nigeria (Dongo & Orisajo, 2007; Thresh & Tinsley, 1959), Sierra Leone (Attafuah *et al.*, 1963), and Togo (Oro *et al.*, 2012b; Partiot *et al.*, 1978), where crop losses can reach up to 30% or more when the trees die (Ollennu & Owusu, 2003). Recently, outbreaks of CSSV have been reported in Ghana and Cote d'Ivoire in which trees characterized by a combination of symptoms characteristically associated with swollen shoot disease, and/or a previously unprecedented, rapid tree decline that results in death in 1-3 years after symptoms of the infection are observed (Kouakou *et al.*, 2012). In Cote d'Ivoire, phylogenetic analysis of sequences corresponding to the 5' end of open reading frame (ORF) 3 from isolates collected from areas with new outbreaks and other regions identified three new CSSV "groups" in cacao (Kouakou *et al.*, 2012), in addition to the three that had been described in Togo (Oro *et al.*, 2012b).

Cacao plants infected with CSSV exhibit different symptoms, depending on the cacao genotype, infecting strain, age of maturity, and environment (Adegbola, 1975; Posnette, 1947). The most obvious and common symptoms of CSSV infection are swelling of shoots and roots, and red-vein banding, mosaic, fern pattern, mottle and vein-clearing on leaves (Cilas *et al.*, 2005; Jacquot *et al.*, 1999a; Posnette, 1947). Virulent strains can cause have persistent, prominent swellings and variable foliar symptoms, especially on the flush, or newest leaves (Box, 1945; Posnette, 1947). Mild strains may induce transient symptoms that usually appear only in the first flush but

not in the next (Box, 1945). Swollen shoot affected plants also produce beans of smaller than expected size, and the pods are often misshapen. In pods, virus infection can result in small and rounded, instead of large oval-shaped pods. Cocoa beans from infected pods are paler and flatter when compared to those from non-infected trees (Posnette, 1947). Diseased cacao trees show reduced productivity within 1-3 years after infection, and death occurs within five years after virus infection occurs (Posnette, 1947) leading to economic losses. However, virulent strains can cause the death of trees in less than two years. Management of the disease by rogueing infected trees from farms has been generally unsuccessful because of the high cost and that many farmers are unwilling to participate (Dzahini-Obiatey *et al.*, 2006).

The CSSV is classified in the family, *Caulimoviridae* and genus, *Badnavirus*, and its genome is a circular, double stranded DNA molecule (King *et al.*, 2012; Lot *et al.*, 1991) that is 6,920 – 7,297 kb in length (current study). The virus particles are bacilliform-shaped and they are non-enveloped. Typical of other badnaviruses, CSSV replicates through an RNA intermediate (pararetrovirus) and it encodes four to six open reading frames; ORF1, ORF2, ORF3, ORF4, ORFX, and ORFY (Hagen *et al.*, 1993; Muller & Sackey, 2005). The function of the predicted CSSV protein encoded by the ORF1 (16 kDa) is not known, and the ORF2 encodes a predicted protein of approximately 15 kDa that has been associated with DNA and RNA binding activity (Jacquot *et al.*, 1996). The largest predicted protein, ORF3, encodes a polyprotein of approximately 212 kDa, and contains domains located at the 5'end attributed to within-plant viral movement (MP). The remainder of ORF3 encodes the capsid, the aspartic protease, the viral reverse transcriptase (RT), and the ribonuclease H (RNase H) proteins. The ORF4, ORFX, and ORFY, at approximately 95, 13, and 14 kDa in size, respectively, overlap with ORF3 but the function of the respective predicted proteins is not known (Jacquot *et al.*, 1999a; Muller & Sackey, 2005).

Serological and molecular detection methods developed for detection of CSSV in symptomatic cacao trees have been shown to produce inconsistent results (Kouakou *et al.*, 2012; Ollennu & Owusu, 2003; Oro *et al.*, 2012b; Sagemann *et al.*, 1985). Serological methods using the Enzyme linked-immunosorbent assay (ELISA) were developed to detect CSSV isolates causing symptoms related to infection by mild or severe strains. Using antisera raised against the severe strain (1A) detected some severe and mild CSSV strains, but also failed to detect the virus in

some plants showing typical symptoms caused by the severe strains (Ollennu & Owusu, 2003; Sagemann *et al.*, 1985). These findings suggested that the undetected virus isolates causing the different symptoms were not serologically related to the 1A strain, although they were causing similar symptoms. For example, in some instances CSSV was detectable using one set of polymerase chain reaction (PCR) primers designed to amplify the 5' end of the ORF3, but not by another, or the inverse (Kouakou *et al.*, 2012). The inconsistency in virus detection suggests that CSSV genomic sequences are highly divergent.

Recent studies involving PCR-amplification of several isolates of CSSV at the regions coding for the CP, MP, RNase H, and RT proteins, and also the non-coding intergenic region showed that, at least once, all PCR primers failed to amplify CSSV in apparently symptomatic plants, or in cases where other primers had successfully amplified it. These findings suggest that no single primer pair is capable of detecting all isolates of the virus (Chapter 2). The latter discovery has provided evidence to support the hypothesis that CSSV (nt) genomic sequences are highly variable, and they also suggested that more than one species of CSSV may exist. As early as the 1940s, at least ten distinct CSSV isolates were shown to occur based on the production of differential symptoms in different virus-infected cacao genotypes (Posnette & Todd, 1955; Posnette, 1947). Also, some of the variants caused only shoot swelling, while others caused symptoms only on leaves or pods. The putative 'virulence' of the isolates was also reported to be variable, with some over others causing greater yield reduction and/or more rapid or slow death of trees, after foliar symptoms became apparent. Since then results of PCR-amplification, cloning, and phylogenetic analysis of the viral MP coding region has distinguished six "groups" of CSSV variants in Cote d'Ivoire, Ghana, and Togo (Kouakou *et al.*, 2012; Muller & Sackey, 2005; Oro *et al.*, 2012)

The lack of complete understanding of the genomic variability that resides within the putative CSSV complex has hindered the development of effective molecular-based diagnostics, especially, given the limited number of complete genomic sequences, e.g. seven, that are available in the NCBI GenBank database. To develop molecular diagnostic tools that enable the detection of all CSSV variants a significant number of complete nt genomic sequences is needed. The objective of this study was to determine the genomic sequences and the molecular variability of CSSV isolates associated with cacao and non-cacao plants in Cote d'Ivoire and Ghana.

Materials and Methods

Plant samples. Sixty-five of 124 symptomatic and non-symptomatic cacao leaf and shoot samples (Chapter 2) collected from fields located in three major cacao growing locations (eastern, western and central), and from greenhouses at Centre National de Recherché Agronomique (CNRA) in Cote d'Ivoire in 2012 and are labeled here starting with CI, followed by sample number. Leaf material was also collected from symptomatic non-cacao plants that were found in and around the symptomatic cacao plants. Also, 35 cacao leaf samples were collected in 2015 from three cacao-growing regions in Ghana, as well as from the "museum", the collection of different virus isolates established at the Cocoa Research Institute of Ghana (CRIG). The Ghana samples are labeled starting with GH, followed by sample number. All samples were preserved in 100% glycerol before they were transported to the School of Plant Sciences University of Arizona, Tucson AZ, where they were stored at 4 °C.

DNA isolation. The cacao leaves were washed to remove the 100% glycerol storage solution, and blotted dry. Total DNA was isolated and purified using two methods for a total of 100 samples. The first method was used to isolate DNA from 100 samples, and it was according to the Cetyl trimethylammonium bromide (CTAB) (Doyle & Doyle, 1990). One hundred mg of leaf tissue was finely powered in liquid nitrogen using a sterile plastic pestle in a 1.6 mL Eppendorf tube. The frozen powder was transferred to a 2-mL tube with four 3.2 mm stainless steel beads (Next Advance) and 1.2 mL CTAB buffer containing 2 % β-mercaptoethanol (Sigma-Aldrich). Samples were pulverized for 5 min by placing tubes in the Mini Beadbeater™ (Biospec Products). The DNA pellet was dissolved in a final volume of 100 µL low TE buffer (10 mM Tris-HCL (pH 7.5) containing 0.1 mM EDTA (pH 8.0)), and stored at -20 °C. The second DNA isolation method involved partial purification of virus particles from leaves using polyethylene glycol (PEG) 6000 (Sigma) for 18 of the 100 samples used for the first method. Leaf tissue (300mg) was ground in liquid nitrogen as described for the first method, and frozen tissue pulverized for 30 sec, as described above, but in 3 mL Tris-glucose buffer (0.5 M Tris, pH 6.8 with 0.5 M glucose). After pulverization, chloroform: butanol (1:1 ratio) was added to a final vol:wt of 0.25 per wt starting tissue. The suspension was mixed by inverting each tube 5 times. The mixture was centrifuged at 10,000 rpm for 5 mins at 4°C, and the supernatant was

transferred to a sterile Eppendorf tube. Sodium chloride and PEG 6000 were added to a final concentration of 0.2 M and 8.0 %, respectively, and the solution was mixed by gentle rocking at 4°C for 2 hrs. The suspension was collected by centrifugation at 10,000 rpm for 10 min at 4°C, resuspended in 100 µl 10mM Tris buffer, pH 7.5, and treated with DNase (Ambion DNA-free kit), according to the manufacturer's instructions. The partially purified virus particles were incubated at 65°C for 15 min. Chloroform: butanol (1:1) was added to a final vol:wt of 0.25 per wt starting tissue. The preparation was mixed by inversion, and centrifuged at 10,000 rpm for 5 min at 4°C to collect the supernatant. Two-thirds vol isopropanol and one-fifth vol 3 M ammonium acetate were added to the supernatant, followed by incubation at -20°C for 30 min. The precipitated DNA was collected by centrifugation at 10,000 rpm for 10 min, and washed with 1 mL cold 70% ethanol. The pellet was collected and dissolved in 30 µl low TE, pH 7.5.

Rolling circle amplification. The DNA isolated using the CTAB method (CTAB-DNA) or the PEG method (PEG-DNA) were amplified using Templiphi rolling circle amplification (RCA) kit (GE Healthcare Bio-Sciences), according to the manufacturer's instructions with the modifications of Rector *et al.*, (2004) and Stevens *et al.* (2010). Two µL DNA were combined with 10 µL of the kit sample buffer and the preparation was denatured at 95°C for 3 min, and cooled on ice for 3 min. The RCA reaction mixture containing 10 µL of the kit reaction buffer, 0.4 µL Templiphi enzyme mix, and 450 µM of dNTP mix (Sigma-Aldrich) was added to the denatured DNA. The reaction was incubated at 30°C for 18 hours, held at 65°C for 15 min to denature the enzyme, cooled to 4°C, and stored at -20°C. Hereafter, the RCA products are referred to as CTAB-RCA and PEG-RCA, respectively.

Illumina sequencing and assembly of paired-end reads. Paired-end libraries were prepared using a TruSeq PE cluster kit with a mean insert size of 350 bp, and individually indexed (tagged) for template produced from CTAB-DNA, CTAB-RCA, PEG-DNA, and PEG-RCA viral isolation methods. Three Illumina HiSeq 2500 DNA sequencing runs were carried out for each tagged template at the University of Arizona Genomics Core (UAGC). The resulting DNA reads were de-multiplexed, and the quality of reads was assessed using FASTQC software (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). The adapter sequences were removed and reads were trimmed using TRIMMOMATIC v0.32 software (Bolger *et al.*, 2014).

Sequences were assembled *de novo* using DNASTAR SeqMan NGen v.12, with or without the option of filtering background host DNA prior to *de novo* assembly. The filtering step was carried out using the two genome sequences for *T. cacao* available in the GenBank Accession numbers CM001879.1 to CM001888.1, FR7222157.1, and KE132922.1 (Argout *et al.*, 2011; Motamayor *et al.*, 2013), the cacao chloroplast genome, GenBank Accession NC014676.2, and the mitochondrial genome of *Gossypium hirsutum* L., GenBank Accession JX065074.1, a related species in the Malvaceae. Sequences were downloaded from the NCBI GenBank database. Only contigs greater than 100 nt in length were considered in the subsequent analyses.

Annotation of badnavirus-like genomes. The assembled contigs were annotated using the BLAST2GO software that accommodates a large batch of sequences simultaneously (Conesa & Gotz, 2008), to confirm badnavirus identity of sequences. Each sequence identified as badnavirus-like was subjected to a BLASTn search using the software and sequence database available at the NCBI-Genbank website (Altschul *et al.*, 1990). The nucleotide (nt) sequence of each apparently full-length CSSV genome was ordered by assigning the first nt of the predicted tRNAmet primer binding site as coordinate number 1, in accordance with badnavirus convention (Dixon & Hohn, 1984; Hull & Covey, 1983).

Pairwise analysis of CSSV coding regions and complete genome sequences. The genome sequence for 122 badnavirus isolates, representing 37 viral species (King *et al.*, 2012) were downloaded from the NCBI GenBank database. To reduce the number of sequences, the redundant sequences were identified using a haplotype search implemented in the FaBox tool v1.41 (Villesen, 2007) and removed from the dataset, leaving 87 haplotypes, which were used as reference genome and partial genome sequences for the pairwise nt and phylogenetic analyses. The RT-RNase H region (580 bp), at <80% nt sequence identity, represents the currently accepted sequence for species demarcation by the International Committee on Taxonomy of Viruses (ICTV) (http://www.ictvonline.org; (King *et al.*, 2012) and is the region delimited by the BadnavirusFP/RP primers (Yang *et al.*, 2003). To calculate the pairwise nt identities, either the conserved RT-RNase H region, or the complete CSSV genome sequences, respectively, were aligned with their 87 counterpart badnavirus haplotype sequences using MUSCLE, implemented in the CLC Sequence Viewer 7.5 (http://www.clcbio.com/products/clc-sequence-viewer). The

aligned sequences were subjected to pairwise nt analysis implemented in the Standard Demarcation Tool (SDTv1.2) software (Muhire *et al.*, 2014), which creates a matrix with pairwise identities and enables the user to apply various experimental % nt cutoff values and to identify optimal species cutoff values, based on frequency of sequence distributions. Likewise, a conventionally-excepted species cutoff may be applied to parse species into groups (Muhire *et al.*, 2014), based on previously established species demarcations (King *et al.*, 2012).

Phylogenetic analysis of CSSV coding regions and complete genome sequences. The aligned RT-RNase H and complete genome sequences were subjected to phylogenetic analysis using the Maximum Likelihood (ML) algorithm implemented in MEGA6 (Tamura *et al.*, 2013). The ML phylogenetic trees were reconstructed using the best model determined by the software (General Time Reversible model and Gamma distribution with Invariable sites), for 1000 bootstrap replicates (Tamura *et al.*, 2013).

Validation of CSSV genome sequences by Sanger sequencing. To validate the presence and sequence accuracy of the *de novo*-assembled Illumina sequences in field-collected samples, primers were designed and used to PCR-amplify, clone and sequence six unique, full-length CSSV-like genome sequences, three from Cote d'Ivoire and three from Ghana. The six samples were selected as representatives for the SDT RNase H Groups 1 (all four sequences) and 2 (two sequences, one sharing 89 – 94% nt identity and the second sharing 94-100% nt identity with the Group 2 sequences). There were no full-length sequences determined for Groups 3 and 4 in this study. The respective RCA products were subjected to PCR-amplification to obtain a putative full-length genome.

The PCR-amplification was carried out using sequence-specific abutting primers (Table 3.1). The PCR primers were designed manually and those with optimum parameters based on the analysis using the Netprimer software (http://www.premierbiosoft.com) were selected. The PCR reactions were carried out using the Invitrogen CloneAmp^{TM} HiFi PCR Premix (Clonetech Laboratories), according to the manufacturer's instructions. The reaction mixture contained 1X CloneAmp^{TM} HiFi PCR Premix, 0.2 µM each of the reverse and forward primer, respectively, 2 µL of the RCA product as template, and nuclease-free water to a final volume of 50 µL. The

PCR conditions were: initial denaturation at 98°C for 2 min, followed by denaturation at 98°C for 20 sec for 40 cycles, annealing at 55°C for 15 sec, extension at 72°C for 8 min, with a final extension at 72°C for 10 min. The PCR amplicons were separated by agarose gel electrophoresis (0.8%) for 90 min at 100 V, and stained with GelGreen (10µL/mL) stain (Biotium) in 1X TAE buffer (40mM Tris, 20mM acetic acid, and 1mM EDTA), pH 8.0. Bands approximately 7 kbp in size were cut from the gel, and purified using the GE Healthcare Bio-Sciences kit. The DNA concentration was determined using the Nanodrop 2000 UV-Vis spectrophotometer at 340 nanometers (nm) (Thermo Scientific).

The purified amplicons were cloned into the pGEM5 plasmid vector (Promega) previously digested using the restriction endonuclease *Not* I (New England Biolabs). Ligation and transformation were carried out using the In-Fusion HD Cloning kit (Clontech), according to the manufacturer's instructions, using the bacterial strain *E. coli* DH5α. The insert size for each clone was confirmed by purifying the plasmid DNA from transformed *E. coli* colonies, based on blue-white colony selection, using the GeneJET Plasmid Miniprep Kit (ThermoFisher Scientific), according to the manufacturer's instructions. The cloned viral inserts were released from the plasmid vector by digestion with the *Not* I endonuclease, for which a single cloning site is available on the plasmid vector. The cloned PCR products were separated by agarose gel electrophoresis, as described above. Clones having an insert of >7 kbp in size were subjected to bi-directional capillary (Sanger) DNA sequencing, using primer walking, at Eton Bioscience (San Diego, CA). The cloned DNA sequence fragments were assembled using DNASTAR SeqMan Pro v.12 and the ORFs were identified using the NCBI ORF Finder tool. The viral ORFs, non-coding regions, and the conserved protein domains were compared to the Illumina-derived sequence determined for the corresponding cacao field isolate, using the NCBI ORF Finder, CDD and SDT, respectively, as described above.

Badnaviral genome coding and non-coding regions, and conserved protein domains. The predicted ORFs at >100 amino acids (aa) encoded on the apparently full-length CSSV genomes were identified using the standard genetic code and the ORF Finder tool, available at the NCBI website (http://www.ncbi.nlm.nih.gov/gorf/gorf.html). The Conserved Domain Database (CDD) tool (Marchler-Bauer *et al.*, 2015), available at the NCBI website

http://www.ncbi.nlm.nih.gov/cdd/, was used to locate the conserved protein domains identifiable in the predicted coding regions of the CSSV-like genomes. The SDT algorithm (Muhire *et al.*, 2014) was used to calculate the pairwise nt distances between each CSSV ORF and its homologous ORF when present in the full length badnaviral genome sequences available in GenBank.

Results

Illumina sequencing and assembly of paired-end reads

Twenty-two of the initial 100 field samples yielded contigs with at least one CSSV sequence from the CTAB-DNA, CTAB-RCA, PEG-DNA, and PEG-RCA combined tagged templates sequenced using the Illumina HiSeq platform. The number of sequence reads for each isolate ranged from 410,662 – 31,417,994. The reads were assembled into 350 contigs containing CSSV sequences. Of these, 16 were full-length CSSV-like genomes whose sizes ranged from 6,920 to 7,172 bp, and 14 of the 16 genomic sequences were unique. The remaining 334 contigs were found to be partial CSSV genomic sequences ranging in size from 136 bp to 6,900 bp. All the complete genomes, except for two, were also accompanied with partial genomic sequences derived from the same sample. Of the 350 virus contigs, 11%, 41%, 11% and 37%, were derived from CTAB-DNA, CTAB-RCA, PEG-DNA, and PEG-RCA preparations, respectively.

The number of partial CSSV contig sequences ranged from 1 - 52 per sample for three combined sequencing runs. The 350 contigs were of very variable depth of coverage from a minimum of 2 to a maximum of 6600 sequences. In general, the number of virus contigs and the depth of sequence coverage were independent of the total sequences reads. However, most contigs with full-length viral sequences (11) were determined for sequence reads greater than 3.5 million, and most of the contigs with partial CSSV sequences were assembled in samples with less than 3.5 million sequence reads. The 350 contigs were organized and distributed among six groups as follows: >500bp (25%); 501 – 2,000 bp (53%); 2,001 – 3,500 bp (9%); 3,501 – 5,000 bp (5%); 5,001 – 6,500 bp (3%) and >6,501 bp (5%). The maximum sequence coverage for each group was 550, 2750, 2750, 304 and 6600 sequences, respectively. Nine full-length viral genome contigs were determined from CTAB-RCA enrichment of samples, the remainder were from CTAB-purified DNA (three), and only one from the PEG virion-purified DNA sample. The *de*

novo assembly method determined three full-length contigs, while the *de novo* method with the option to remove host genomic sequences determined eight full-length virus contigs.

Comparison of full-length badnaviral genomes from Sanger and Illumina platforms

The assembled full-length CSSV-like genome Sanger-determined sequences were identical in length (within 3 nt) with the Illumina-produced genome sequences (Table 3.2). The exception was the isolate CI286 from Cote d'Ivoire, whose sequence was determined by PCR amplification using sequence-specific abutting primers that were based on the partial sequence determined by the Illumina platform. By pairwise distance analysis, the sequences determined by Sanger sequencing were 99.0-99.9% identical to the respective sister sequence (data not shown), indicating a small number of single nt polymorphisms probably due to infidelity of the enzymes used for RCA enrichment of circular DNA, PCR amplification, and/or error inherent in either of the sequencing platforms. Based on phylogenetic analysis of the aligned sequences, the CSSV-like genomes determined by Sanger sequencing of the cloned PCR amplicons, were grouped on the ML tree with the respective Illumina-derived sequence either within the well-resolved CSSV subclades (data not shown). The results of the phylogenetic, pairwise distance (SDT), and the conserved protein domain analyses (see below), were carried out using only the Illumina HiSeq-derived sequences, with the exception of the CI286 sequence, for which Sanger result was used because the Illumina platform produced only a partial genome for that isolate.

Pairwise nucleotide identities

The shared pairwise nt identity of the RT-RNase H region for CSSV-like sequences determined herein, and the analogous coding region from 87 badnaviral haplotype sequences in GenBank ranged from 58% to 70%, compared to 58-65% among the 87 full-length genome sequences.

CSSV RT-RNase H region comparison. Pairwise distance analysis of the RT-RNase H region for the 14 CSSV-like sequences determined herein and the seven available in GenBank indicated they shared 70% to 100% nt identity (Table 3.4). Based on the ICTV species threshold, at >80% for the RT-RNase H region, four CSSV groups were determined using the 14 sequences and the seven GenBank reference sequences. The first group contained four isolates, CI275, CI286, GH64 and GH67 (Group 1) and no GenBank references. Group 2 comprised 10 sequences determined herein, and five reference sequences represented by three isolates from Ghana

(GenBank Accessions, AJ608931, AJ609019, and AJ609020), and two isolates from Togo (GenBank Accessions, AJ534983 and L14546). Group 3 contained the GenBank reference sequence from Cote d'Ivoire, Accession JN606110 and Group 4 contained one reference from Togo, Accession number AJ781003, but none of the sequences from this study were grouped within these two.

CSSV complete genome sequence comparisons. The pairwise nt analyses of complete CSSV genomes (Table 3.5) indicated slightly somewhat different results, compared to pairwise analysis of the RT-RNase H fragment (Table 3.4). The complete CSSV-like genome sequences shared 68% to 100% nt identity, compared to shared 70% to 100% nt identity in the RT-RNase H region. Although the SDT analysis of the partial and complete genome sequences resolved four groups based on the species demarcation threshold of >80% threshold, the groupings of the isolates was incongruent. When the complete genome was considered, Group 1 isolates split into two, CI286 into a separate group where it was the only member, and three other sequences (CI275, GH64 and GH67) into a separate group (Table 3.5). Also, the RT-RNase H Group 4 shifted to Group 2 (Table 3.5). However, only the nt identities of the Group 4 member compared with two other Togo sequences were greater than 80%; the identities with all the other sequences were below the established 80% threshold, suggesting that it could possibly represent a recombinant badnavirus. Group 3, however, remained unchanged when the RT-RNase H and the complete genome analyses were compared. The results of pairwise analysis indicated that CSSV is a complex of at least four species, and that Group 1 contained the four sequences from Cote d'Ivoire and Ghana (CI275, CI286, GH64, and GH67), Group 2 isolates consisted of 10 isolates whose sequenced were determined in this study and GenBank reference sequences for Togo and Ghana, and Groups 3 and 4 contained one reference sequence each from Cote d'Ivoire and Togo, respectively.

Phylogenetic analyses

The RT-RNase region. Phylogenetic analysis of the RT-RNase H region of the 87 badnavirus sequences and 14 CSSV-like sequences resolved three major clades, Clades I, II and III, but none of the clades were statistically supported at >70% bootstrap (Fig. 3.2). However, the CSSV subclade in Clade I was statistically supported at 100% bootstrap, and indicated that all the 14

genomic sequences from this study group with the previously published CSSV sequences, suggesting that they are more closely related to CSSV than to any other badnavirus species. Three smaller clades Subclades A, B and C, were apparent in the CSSV subclade and they appeared to be grouped according to ORF arrangement. Subclade A contained sequences with five or six predicted ORFs, while Subclades B and C contained sequences with four and five predicted ORFs, respectively.

Full-length viral genome. Phylogenetic analysis of the apparently full-length CSSV-like and 87 badnavirus genomes resolved three well-supported major clades, Clades I, II and III, at >70% bootstrap each (Fig. 3.3). Also, two divergent genome sequences were positioned basal to the three badnavirus clades (Fig. 3.3). All of the 14 full-length sequences determined here and the seven CSSV GenBank reference sequences formed one well-supported subclade in Clade I, at 100% bootstrap, designated herein as CSSV. Unlike the RT-RNase H CSSV subclade which contained three subclades, two subclades were observed in the tree based on the full-length genomic sequences. The two subclades, Subclade A and B, were similar to the subclades formed in the RT-RNase H tree in that, the sequences in each subclade were grouped according to the ORF arrangement. Five or six, and four ORFs, were predicted for the Subclades A and B, respectively.

ORF arrangement

The CSSV-like isolates in the RT-RNase H SDT Group 1 were characterized by having four predicted ORFs, whereas, Group 2 species had either five or six predicted ORFs. The Groups 3 and 4 each had five ORFs, but grouped separately from Group 2 (Table 3.4). In comparison, the SDT analysis of the full-length genomes formed four groups in which, Groups 1 and 2 each had four predicted ORFs, and Groups 3 and 4 had five, and either five or six, predicted ORFs, respectively (Table 3.5).

The number of predicted ORFs for the RT-RNase H tree CSSV Subclades A, B and C were five or six, four, and five, respectively (Fig. 3.2). The grouping based on ORF arrangement was consistent with the pairwise nt identity analysis of the same region. Subclade A contained the same sequences as the SDT Group 2, Subclade B contained sequences found in Group 2, and

Subclade C had sequences from Groups 3 and 4. The phylogenetic tree based on the complete genomic sequences revealed Subclades A and B, containing sequences with five or six, and four, ORFs, respectively (Fig. 3.3). The results from pairwise distance analyses (Table 3.5) were similar to the phylogenetic analysis results in that, Subclade A contained sequences from Groups 3 and 4 with five or six ORFs, while Subclade B contained sequences with four ORFs from Groups 1 and 2. In both the pairwise identity and phylogenetic analyses, for both the RT-RNase H and the full-length genomic sequences, the isolates having four predicted ORFs were separated from those having five or six ORFs.

Genome characterization of CSSV-like isolates

The results of a BLASTn search for the 350 partial or full-length CSSV-like genome sequences against the GenBank database, indicated that all the sequences had the highest percentage similarity score, greatest percentage coverage, and robust e-value score of zero, with the other sequences in the database corresponding to CSSV isolates.

Among the Genbank CSSV sequences used as reference sequences, matches were found to all of them except to the GenBank Accession JN606110, for an isolate from Cote d'Ivoire. This was of interest because 65% of the cacao samples were collected in Cote d'Ivoire. The BLASTn search hits had matches from one to five of the six CSSV GenBank references in various combinations, for fragments of at least 191 bp in length. The distributions for the number of BLASTn hits were 14%, 36%, 23%, 18%, and 9% for 1, 2, 3, 4, and 5 GenBank Accessions combinations, respectively. The number of BLASTn hits for each sample was independent of the number of contigs produced by the sequence assemblies. None of the contigs delivered a match to other badnavirus sequences in the GenBank database with an e-value of zero.

Identification of the predicted ORFs found in the 14 unique full-length CSSV genomes using the NCBI ORF Finder tool, indicated that each full-length genome sequence encoded at least four predicted ORFs (>50 aa) on the viral plus strand (Fig. 3.1), and that three different genome arrangements were represented (Fig. 3.1). One type of genome arrangement (1) contained four predicted ORFs: ORF1, 2, and 3, and ORFY) (Fig. 3.1a). A second type (2) contained five

predicted ORFs: ORF1, 2, and 3, and ORFs Y and X (Fig. 3.1b), and a third type encoded six ORFS: 1, 2, 3, and 4 and ORFs Y and X (Fig. 3.1c).

Four genomes, CI275, CI286, GH64, and GH67, representing field collections from Cote d'Ivoire and Ghana were of the four ORF type (1). About half of the genomes, CI44, CI134, CI135, CI215, CIS2, CIS3, and CIT5, contained five ORFs (type 2) and represented only exemplars from Cote d'Ivoire. The three genomes, CI301, CI311, and GH75, representing collections from both Cote d"ivoire and Ghana encoded six ORFs (type 3).

The nt and predicted amino acid (aa) size, the molecular mass, and the BLAST percentage nt and aa similarity score for the predicted CSSV ORFs were in the same ranges as the previously published CSSV genomes, and to other badnaviral species assigned to the genus, *Badnavirus* (Table 3.2).

The predicted ORFs, except ORF3 and ORFX, were similar in size at the aa level e.g. within 2 to 3 aa, with a few exceptions. Overall, ORF1 was 143 aa in size, except GH64, which was 162 aa, ORF2 was 145 aa but GH64 and CI275 had 142 aa and 143 aa, respectively. The ORFY size was 130 - 131 aa, except for CI311, which encoded 135 aa. In contrast, the size of ORFX, when present, was widely variable at 57 - 91 aa, and ORF3 encoded 1,811- 1,879 aa. In the three genomes that encoded an ORF4, the coding region was 94, 95 or 99 aa in size. All of the genome sequences had a predicted $tRNA^{met}$ binding site located at the nt coordinates 1- 18, which is complementary to a sequence found previously in plants (Ghosh *et al.*, 1982). In CSSV-like genomes, this intergenic region is known to be the location at which reverse transcription of the viral genome is initiated (Geering, 2014; Medberry *et al.*, 1990a).

Conserved protein domains

Conserved protein domains were predicted in ORFs 1 and 3 for all full-length genomes, but none were predicted in ORFs 2, 4, X, and Y, with one exception (Table 3.3, Fig. 3.3). A badnavirus-specific, uncharacterized protein superfamily, known as domain of unknown function (DUF1319) was predicted in ORF1. Four domains were predicted for the ORF3 in all genomes: the zinc knuckle finger (Zn), the pepsin-like aspartate protease (Pep), the reverse transcriptase

(RT) and the ribonuclease H (RNase H) domains. The exception was the GenBank reference JN606110 for which had no Pep domain was found, and the domain of unknown function, DUF3187, was predicted in ORFY.

A comparison of CSSV to CaMMV and CYVBV, the only other cacao-infecting badnaviruses, revealed that CaMMV shared similar domains in the ORFs 1, 2, and Y, and CYVBV shared similar domains only in the ORFs 1 and 2 (Fig. 3.3). However, additional domains were located between the zinc knuckle and the pepsin-like aspartate protease domains of the ORF3 of CaMMV and CYVBV. The CaMMV ORF3 had a BAR (GTPase regulator associated with focal adhesion 2 (Bin/Amphiphysin/Rvs)), while the CYVBV ORF3 had a trim (trimeric-dUTPase) domain in the same coordinates. Also, an HTH (helix-turn-helix) domain was predicted in the CYVBV ORFY, but was absent from the CaMMV and CSSV ORFY (Fig. 3.4).

Discussion

CSSV genomic sequences from symptomatic cacao samples collected from Cote d'Ivoire and Ghana were determined using the Illumina sequencing technology. Assembly of the sequences resulted in 14 complete and 336 partial CSSV genomes, representing 22 samples of the 100 analyzed. Four complete genomes represented new CSSV species that had not been previously described. The genomic sequences of the new species have been validated using PCR and Sanger DNA sequencing and 99.0 – 99.9% nt identity was shared between the sequences from the two methods. Here, we describe the identification and the characterization of the newly sequenced CSSV genomes, and show that CSSV is a complex of species.

The CTAB-RCA enrichment method used to prepare the samples subjected to Illumina-HiSeq sequencing produced sequence data that resulted in the assembly of contigs containing CSSV sequences from the largest number of samples, and more full-length genomic sequences were determined than when CTAB-DNA, PEG-DNA or PEG-RCA were used. The RCA technique is specifically designed for the amplification and enrichment of circular DNA (Dean *et al.*, 2001). When the circular CSSV DNA is amplified, the products likely dilute any single stranded DNA

molecules in the sample, increasing the chances of the circular genomes being sequenced. In contrast, CTAB-DNA, which is usually found in low titers, has less chances of being sequenced.

The PEG virus purification method isolates virus particles (virions) from the host genomic DNA, and sequencing of the purified preparation should provide viral sequences with minimal host background sequences. However, the PEG-DNA and PEG-RCA preparations did not yield a significant number of virus contigs as expected, but rather, the numbers were extremely low compared to the CTAB-RCA preparation. This may be because the virus purification method implemented was not effective at isolating the virus particles from the plant samples. The optimization of the purification method to isolate more virus particles, including using larger quantities of starting plant material will likely improve assembly of full-length CSSV sequences (Lot *et al.*, 1991).

The BLASTn search for the CSSV contigs assembled in this study shared no homology with other known plant viruses, except for isolates named CSSV and certain other badnavirus species. Using the 87 badnavirus sequences, representing 37 species, in a pairwise distance analysis implemented in SDT software (Muhire *et al.*, 2014), none of the contigs were found to represent other known badnavirus species. The pairwise nt comparison (SDT) of the 87 badnavirus genome sequences, with 21 CSSV-like sequences (14 herein and seven GenBank references), indicated a shared nt identity of 58-70% for the putative, taxonomically-informative RT-RNase H region, and 58-65% for the complete badnavirus genomes. However, the 21 CSSV-like sequences shared 70-100% for the RT-RNase H region and 68-100% for the complete genomes. This range of shared nt identities, calculated using either a partial or complete genome sequence, indicated that the cacao-infecting isolates from West Africa are more closely related to CSSV than to other known badnavirus species. Further, 14 isolates reported here for the first time, are therefore considered divergent variants of the currently recognized CSSV. Four of the 14 isolates are highly divergent and share 70-78% nt identity with the seven CSSV GenBank reference sequences, and the remaining ten isolates determined in this study shared 80-100% nt identities with the reference sequences for the RT-RNase H and the full-length genomes, combined (Tables 3.4 and 3.5).

The pairwise identity analyses of the 21 CSSV-like sequences revealed four CSSV, Groups 1-4 within a larger CSSV-like group of species, of which Groups 2, 3 and 4 (Tables 3.4 and 3.5) have been previously reported (Hagen *et al.*, 1993; Muller & Sackey, 2005). In addition, the Group 1 genome type, which are represented by four CSSV-like variants (sequences) identified here for the first time, have not been previously described from cacao or from any host. Further, the sequences in Group I do not have a significant match to any available CSSV GenBank reference sequences. These results strongly suggest that CSSV, as the viruses are currently named, does not represent a single virus species, but that instead, it is a complex of cacao-infecting viruses. Virus species complexes are not uncommon among certain plant virus families. Among badnaviruses, complexes are recognized for *Banana streak virus* (BSV)-like variants (Geering *et al.*, 2000; Harper *et al.*, 2005; Lockhart & Olszewski, 1993), and for Sugarcane mosaic virus (genus *Potyvirus*, family *Potyviridae*)(Yang & Mirkov, 1997). And, in the family, *Geminiviridae*, a group of five distinct begomoviral species has been recognized, based on the ICTV-accepted 91% species cut-off for the genus, *Begomovirus*, that cause leaf curl disease of cotton in Pakistan (Briddon, 2003; Brown *et al.*, 2015).

Phylogenetic analyses of the RT-RNase H region and the complete badnavirus genome sequences showed that the 14 CSSV-like sequences determined in this study group phylogenetically with other badnavirus sequences in Clade I (Fig. 3.2 and 3.3). The RT-RNase H phylogenetic analyses resolve three unsupported badnaviral clades while the full length genome tree had three statistically-supported badnaviral clades, and therefore more informative than the former. However, the two trees showed a statistically supported CSSV clade, each at 100% bootstrap.

The RT-RNase H tree resolved three CSSV subclades, A, B and C, while the full-length genome tree resolved two CSSV subclades, A and B, within a larger group that does not include other badnaviruses. However, other badnavirus species cluster in the same major badnavirus clade, Clade I, with the CSSV species group (Fig. 3.3; Chingandu et al., 2016, *submitted*). The analyses based on the partial and complete genomes showed that the 14 sequences reported in this study are badnavirus sequences and that they represent isolates of CSSV, or species closely

related to CSSV. This finding is supported by the result of the pairwise distance analysis that resolved four groups among the CSSV full length genome sequences (Tables 3.4 and 3.5). The CSSV clade from the complete genomic tree groups phylogenetically in Clade I with eight other badnaviruses that infect a wide range of plant species, but not cacao. Only CaMMV, one other recently described cacao-infecting virus in Trinidad (Chingandu et al., 2016, submitted), is found in Clade I. The CYVBV, the second of the recently described cacao-infecting badnavirus also from Trinidad (Chingandu et al., 2016), is clustered in Clade II with seven other non-cacao-infecting badnavirus species (Fig. 3.3). Based on phylogenetic analysis of the complete genomes, the CaMMV and the CYVBV are not members of the CSSV species group. The results suggest that the three viruses probably originated from different, unrelated extant lineages, and further, that they may have evolved independently, first in different host species and from different geographical regions, and then underwent a host-shift to cultivated cacao.

The evolutionary history of the newly sequenced CSSV species has not been investigated, but a preliminary analysis based on the two phylogenetic trees showed that most of the isolates from Cote d'Ivoire formed a clade with Ghana reference sequences, suggesting the origin to be the latter. Indeed, CSSV was first described in Ghana (Steven, 1936), and it is believed to have spread to other neighboring areas, including Cote d'Ivoire, from where it was described about ten years later. In contrast, the four isolates (sequences) that did not group with previously published CSSV-like sequences, appear to represent isolates that have either gone unnoticed in cacao since CSSV was first described in 1936 (Steven, 1936), or have recently emerged in response to selection pressures, albeit unknown.

In this light, perhaps the new CSSV-like species discovered in cacao may have only recently shifted to cacao from indigenous wild hosts by the mealybug vector. Prior to the introduction of cacao into West Africa during the late 1880's, CSSV was known to infect a number of hosts endemic there (Posnette *et al.*, 1950; Tinsley, 1971b; Todd, 1951). The alternative wild hosts include *Adansonia digitata* L., *Ceiba pentandra* L., *Cola chlamydantha* K.Schum., *Cola gigantean* A. Chev., and *Sterculia tragacantha* Lindl (Posnette *et al.*, 1950; Tinsley, 1971b; Todd, 1951). These species are forest trees, and are among the 90 or so species among 30 plant families that are used by farmers as shade or cover tree/crops, including in cacao plantations

(Richard & Ræbild, 2016). Even so, many other suspect species have not been investigated as possible CSSV hosts.

In the full-length genome tree and SDT analysis, the Togo reference sequence (Accession AJ781003) grouped with the Subclade A that contained 3 of 3 Ghana and 3 of 3Togo reference sequences (Fig. 3.3), and shared at >80% nt identity with members of the same isolates based on pairwise sequence analysis (SDT) (Table 3.5). By comparison, by pairwise distance analysis, the RT-RNase H region sequence of the Togo isolate was an outlier to all other isolates included in the analysis, and the only example of this kind of divergent representative among CSSV isolates (Table 3.4). One possible explanation is that the isolate represents a recombinant between a Group 4 type isolate and an as yet unknown parent. Until now, recombination between CSSV-like species has not been reported.

Among the initial BLASTn searches for viral contig identification, from one to five CSSV reference sequences were identified in each cacao sample. An exception to this was observed for three of 22 samples positive for CSSV infection, all having only one match each to three different CSSV GenBank reference sequences. Although unsubstantiated by molecular data, naturally-occurring CSSV-mixed infections have been reported in cacao in Ghana (Posnette & Todd, 1955), based on differential symptom phenotypes. Also, the mealybug vectors are known to transmit several strains of CSSV, and also other non-cacao infecting badnaviruses (Kirkpatrick, 1950; Selvarajan *et al.*, 2016).

Mixed virus infections are also very common in other plant species, and they are known to occur within or between genera or families. Virus co-infection can result in variable outcomes, including synergism, competition or attenuation of symptoms (cross-protection) by another (Syller, 2012). Several studies have been carried out to evaluate the practicality of cross protection using CSSV mild strains to protect the plants against virulent strains (Ameyaw *et al.*, 2016; Ollennu & Owusu, 1989, 2003; Posnette & Todd, 1955). In these studies, the mild strains were shown to offer substantial protection against more virulent strains on cacao. For a period of over ten years, growth, yield and survival rates were significantly improved for the protected, compared to the non-protected cacao plants (Ameyaw *et al.*, 2016).

In the statistically supported (70% bootstrap value) CSSV-like clade in the full-length genome tree, the genomes within Subclade A were found to encode four ORFs, whereas the Subclade B isolates had five of six ORFs (Fig 3.3). Further, among the three cacao-infecting viruses, CSSV, CaMMV and CYVBV have different genome arrangements, with the CSSV sequences representing three different ORF arrangements of four, five or six ORFs, but CaMMV and CYVBV with only four ORFs, each. All badnavirus genomes contain between 3 and seven ORFs whereas, the CSSV genomes encode four to seven ORFs. The CSSV-like viral genomes were found to group with eight non-cacao-infecting badnaviruses in Badnavirus Clade I (Chingandu et al. 2016, *submitted)*, encoded from three to seven ORFs (Fig. 3.3). This indicates that the non-cacao infecting badnaviruses in this clade do not group according to a particular genome arrangement, except for the CSSV-like species placed in the CSSV Subclades A and B (Fig. 3.3), which do group by number of ORFs.

There was no robust relationship at the level of country, between most of the CSSV-like isolates, or with the number of ORFs encoded by the genomes. However, the viral genomes having five ORFs were from cacao samples collected eastern and central Cote d'Ivoire, whereas, those genomes with four and six ORFs were from western Cote d'Ivoire, and western and central Ghana samples. The possible geographical link among the latter three locales could possibly be associated with unintentional virus spread through the exchange of virus-infected planting materials (Kouakou *et al.*, 2012), and/or by inter-regional long-distance dispersal of viruliferous mealybug (Cornwell, 1960).

Differences in the number of ORFs observed among the CSSV-like group is corroborated by previous studies that have shown CSSV isolates are divergent and have different genomic arrangements. While ORFs 1, 4, X, and Y have as yet unknown functions, the presence or absence of the other three ORFS are reflective of phylogenetic relationships, however, no clear phylogeographical relationships are apparent. The functions of the different ORFs in the overall infection or viral life cycle, including symptom development and virulence, are mostly unstudied.

Although the number of predicted ORFs encoded by each CSSV genome varies between CSSV-clades, the domain search confirmed that several functional domains are conserved throughout the group. In this analysis, functional domains were predicted in the ORF3 that encodes a polyprotein. The action of the protease enzyme processes the polyprotein into mature proteins that are functionally distinct (Hagen *et al.*, 1993). The identification of these functional domains are in agreement with those predicted on previously published CSSV genome sequences, which are: 1) Zn, a functional motif associated the coat protein gene, 2) Pep, a protease that cleaves the polyprotein into several mature proteins, 3) RT, reverse transcription of RNA transcript to DNA for encapsidation or replication, and 4) RNase H, which is involved in degradation of RNA in a DNA-RNA hybrid synthesized during reverse transcription. These domains are also consistent with those identified in the ORF3 polyprotein of other well-studied badnaviruses (Hohn *et al.*, 1997).

By comparison, the CaMMV and CYVBV species contained additional (identifiable) functional domains on the ORF3, at the region between the zinc finger and aspartate protease domains. The CaMMV ORF3 had a predicted GTPase regulator previously associated with focal adhesion 2 - Bin/Amphiphysin/*Rvs* domain *(*BAR_GRAF2) is involved in protein-protein interaction, membrane binding and curvature sensing (Marchler-Bauer *et al.*, 2015), and, by comparison, two other badnaviruses, BSV and *Hibiscus bacilliform virus* (HBV), contain partial BAR_GRAF2 domains. The CYVBV encodes a predicted Trim domain that functions as dUTPase to catalyze the hydrolysis of dUTP-Mg complexes into dUMP and pyrophosphate (Marchler-Bauer *et al.*, 2015; Tormo-Mas *et al.*, 2013). This domain is also present in the genome of *Piper yellow mottle virus* (Hany et al., 2014), *Dioscorea bacilliform virus,* and *Taro bacilliform virus*, which are phylogenetically unrelated to CYVBV.

In several betaretroviral genomes the dUTPase gene is located adjacent to the viral nucleocapsid (NC) polypeptide gene. Here, it is thought that their close proximity facilitates the formation of a trans frame fusion protein that effectively, joins the dUTPase to NC protein, which has been shown to be resistant to proteolysis by retroviral protease. This protection is afforded owing to co-folding of the domains that participate in RNA/DNA folding, reverse transcription, and DNA repair (Nemeth-Pongracz *et al*. 2006), a co-evolutionary achievement hypothesized to allow a

shortened C-terminus to circumvent its (extant) shortened length and reach an active site. In viruses this is seen to provide an economic solution to encoding a shorter but still effective dUTPase fused to the NC protein, in order to accommodate size-limited genomes. Results from additional studies have shown that the fusion protein might efficiently decrease dUTP pools in the vicinity of viral reverse transcription, thereby conserving time and energy otherwise devoted to hydrolyzing cellular dUTP to dUMP and pyrophosphate, a reaction that minimizes hyper-misincorporation or uracil into DNA during replication and results in cell death (Vertessy and Toth, 2009). In contrast, HIV DNA uracilation was shown to confer a benefit during the early phase of the viral life cycle by inhibiting autointegration (Yan *et al*., 2011). Yet, many badnaviruses appear to carry out these essential functions using separate NC and dUTPase polypeptides without deleterious effects.

In addition, the CYVBV genome has the HTH domain on the ORFY that is implicated in DNA binding, and is consistent with transcriptional regulators (Marchler-Bauer *et al.*, 2015). No other viruses were found to have a similar domain. The functions of the three unique predicted domains in CaMMV and CYVCV are unknown, but they may confer important functions the virus life cycle, although they probably do not confer virulence to the two species. The CSSV generally causes more severe symptoms than CaMMV or CYVBV (Posnette, 1944), and does not have said domains. Even so, phylogenetic groups resolved here, suggest that the CSSV-like badnaviruses that group together do not necessarily encode identical functional domains, in particular, the isolates among them that have unique as well as shared predicted ORFs and/or domains. And, the domain for cell-to-cell movement, located at the 5' end of ORF3 (Hagen *et al.*, 1993), was not predicted in any of the CSSV genomes using the CDD tool.

Lastly, within ORF4, Y, and X of the CSSV-like genomes, no functionally conserved domains could be identified in the database. This either suggests that the unique ORFS encoded on these viral genomes confer no essential functions, or more likely, that the domains are potentially novel and so could not be predicted using the existing CDD search databases. One exception to the latter ORF-type patterns was in the genome of GenBank Accession JN606110, from Cote d'Ivoire (Kouakou *et al.*, 2012), which lacked the ORF3 Pep domain, and encoded a DUF3187 predicted within ORFY (Table 2). The absence of the Pep domain was unexpected because the

polyprotein encoded by ORF3 requires processing to confer functionality of its proteins (Hagen *et al.*, 1993). The predicted absence might imply a previously undocumented mutation, resulting in an unrecognized motif, with potentially conserved or similar function(s). The Cote d'ivoire isolate DUF3187 was annotated as an outer membrane hypothetical protein in Proteobacteria (Marchler-Bauer *et al.*, 2015), and though its function is not known, such a protein would not be expected to be homologous to the Pep domain.

Overall, the genomic variability of the CSSV putative species group has not been well studied in cacao-growing areas of West Africa, underscoring the need for stepped-up research. Important clues lie in this and previous studies that have examined the variability at the level of the movement protein for a large number of isolates from Cote d'Ivoire, Ghana and Togo. Based on these studies the MP region is highly divergent and so is potentially informative at the sub-species level (Kouakou *et al.*, 2012; Muller & Sackey, 2005; Oro *et al.*, 2012b). Even so, the evolutionary histories and taxonomy of the entire badnavirus genus remain to be better resolved as an increased number of complete-genome sequences become available.

Table 3.1. Six primer pairs used for PCR-amplification of full-length *Cacao swollen shoot virus* genomes. The virus-specific sequences contained within each primer are underlined, and the first 8-9 bases are those of the plasmid vector. A *Not* I restriction site was added (italics) to facilitate cloning. The letter codes GH (Ghana) or CI (Cote d'Ivoire) indicate the country from which samples were collected.

Primer name	**Primer sequence 5' – 3'**	**Nucleotide coordinates**
GH67_F	ATCACTAGTGCGGCCGCAATGGCGGATGAACTATGT	2419-2438
GH67_R	GACCTGCAGGCGGCCGCGAATCCACGTTAACTT	2399-2418
GH64_F	GACCTGCAGGCGGCCGCATCACTGGACAGCAATGGT	4672-4690
GH64_R	ATCACTAGTGCGGCCGCTGGTTTTTCCTAACCAATAGG	4651-4671
GH75_F	ATCACTAGTGCGGCCGCGGATTATCCATCCAGGAG	980-998
GH75_R	GACCTGCAGGCGGCCGCTTGGAGTTTTGTGATAAG	959-979
CI44_F	ATCACTAGTGCGGCCGCGCCTATCAAGGTTAAAGCTA	4344-4363
CI44_R	GACCTGCAGGCGGCCGCCTTCCTACTTCAGG	4325-4343
CI275_F	ATCACTAGTGCGGCCGCGTTAACACCCCGAACC	622-640
CI275_R	GACCTGCAGGCGGCCGCTTGGCCTTTTCTTCT	602-621
CI286_F	GAATGTAGCATTTGCCTATC	1722-1741
CI286_R	CTTTTGTTAAATCAATCTCC	1702-1721

Table 3.2. Analyses of predicted open reading frames identified on the plus-strand of *Cacao swollen shoot virus* genomes using the NCBI BLAST algorithm (Altschul *et al.*, 1990) and NCBI ORF Finder tool (http://www.ncbi.nlm.nih.gov/gorf/gorf.html). BLASTn and BLASTp results are based on the first 100 hits. The letter codes CI (Cote d'Ivoire), GH (Ghana), or TG (Togo) indicate the country from which the cacao samples were collected.

Sample		ORF1	ORF2	ORF3	ORFX	ORFY	ORF4	Total size (bp)
CI311	Amino acid size	143	145	1847	91	135	95	7118
	Molecular mass (kDa)	16.78	15.79	212.18	10.87	14.54	11.1	
	Nucleotide coordinates	296-727	724-1161	1127-6670	2310-2585	6310-6717	4215-4502	
	BLASTn similarity	66-99%	73-97%	73-99%	68-96%	71-94%	71-99%	
	BLASTp similarity	26-99%	28-96%	38-98%	54-92%	26-93%	98%	
CI135	Amino acid size	143	145	1847	57	131	ND	7030
	Molecular mass (kDa)	16.78	15.81	212.42	7	14.43		
	Nucleotide coordinates	296-727	724-1161	1127-6670	2412-2585	6310-6705		
	BLASTn similarity	66-99%	74-98%	73-99%	71-96%	66-96%		
	BLASTp similarity	26-99%	28-97%	37-98%	54-89%	24-92%		
CI275	Amino acid size	143	143	1871	ND	130	ND	7172
	Molecular mass (kDa)	16.58	15.59	216.12		14.3		
	Nucleotide coordinates	266-697	694-1125	1088-6703		6330-6777		
	BLASTn similarity	65-75%	66-75%	66-76%		65-68%		
	BLASTp similarity	23-75%	27-61%	38-71%		30-65%		
CIS2	Amino acid size	143	145	1845	89	131	ND	7022
	Molecular mass (kDa)	1677	15.87	212.42	10.67	14.3		
	Nucleotide coordinates	295-726	723-1160	1126-6663	2303-2572	6303-6698		
	BLASTn similarity	65-98%	74-97%	73-99%	72-94%	70-96%		
	BLASTp similarity	25-97%	30-97%	38-97%	57-91%	29-95%		
CIS3	Amino acid size	143	145	1811	90	131	ND	6920
	Molecular mass (kDa)	16.78	15.79	212.45	10.81	14.3		

	Nucleotide coordinates	296-727	724-1161	1127-6560	2310-2582	6200-6595		
	BLASTn similarity	68-98%	75-98%	73-99%	69-95%	70-97%		
	BLASTp similarity	26-99%	29-98%	38-98%	56-90%	30-95%		
CI44	Amino acid size	143	145	1837	68	130	ND	7030
	Molecular mass (kDa)	16.85	15.88	211.5	8.14	14.09		
	Nucleotide coordinates	298-729	726-1163	1129-6642	2372-2578	6285-6677		
	BLASTn similarity	66-94%	75-87%	73-99%	68-84%	69-85%		
	BLASTp similarity	24-94%	28-90%	38-92%	43-67%	24-82%		
GH64	Amino acid size	162	142	1879	ND	130	ND	7115
	Molecular mass (kDa)	18.67	15.41	215.99		14.38		
	Nucleotide coordinates	215-703	700-1128	1091-6730		6361-6753		
	BLASTn similarity	71-80%	67-84%	66-83%		64-68%		
	BLASTp similarity	25-71%	28-62%	38-72%		30-65%		
GH67	Amino acid size	143	145	1841	ND	130	ND	7020
	Molecular mass (kDa)	16.66	15.82	212.55		14.46		
	Nucleotide coordinates	276-707	704-1141	1104-6629		6260-6652		
	BLASTn similarity	69-80%	65-92%	65-90%		65-80%		
	BLASTp similarity	22-73	29-58%	37-74%		31-61%		
GH75	Amino acid size	143	145	1847	91	131	99	7024
	Molecular mass (kDa)	16.74	15.83	212.31	10.99	14.22	11.47	
	Nucleotide coordinates	294-725	722-1149	1125-6668	2308-2583	6308-6703	4213-4512	
	BLASTn similarity	66-99%	74-99%	73-99%	69-98%	70-98%	71-99%	
	BLASTp similarity	25-99%	29-99%	37-99%	57-95%	25-98%	95%	
CI134	Amino acid size	143	145	1846	90	131	ND	7012
	Molecular mass (kDa)	16.77	15.82	212.12	10.82	14.34		
	Nucleotide coordinates	283-714	711-1148	1114-6654	2297-2569	6294-6689		
	BLASTn similarity	68-99%	74-97%	73-99%	70-95%	70-95%		

	BLASTp similarity	26-97%	29-97%	38-98%	60-91%	31-93%		
CI215	Amino acid size	143	145	1845	89	131	ND	7004
	Molecular mass (kDa)	16.79	15.84	212.08	10.67	14.26		
	Nucleotide coordinates	278-709	706-1143	1109-6646	2292-2561	6286-6641		
	BLASTn similarity	66-98%	74-98%	73-99%	71-93%	70-95%		
	BLASTp similarity	26-97%	29-97%	38-98%	58-88%	30-92%		
CI286	Amino acid size	143	145	1869	ND	130	ND	7122
	Molecular mass (kDa)	16.43	15.84	215.48		14.48		
	Nucleotide coordinates	275-706	703-1140	1103-6712		6346-6738		
	BLASTn similarity	70-80%	65-69%	66-96%		65-67%		
	BLASTp similarity	24-71%	26-59%	36-71%		27-66%		
CIT5	Amino acid size	143	145	1843	87	131	ND	7016
	Molecular mass (kDa)	16.77	15.84	211.98	10.58	14.29		
	Nucleotide coordinates	295-726	723-1160	1126-6657	2309-2572	6297-6692		
	BLASTn similarity	66-97%	74-97%	73-99%	72-94%	71-97%		
	BLASTp similarity	25-97%	30-98%	38-98%	59-90%	29-95%		
CI301	Amino acid size	143	145	1846	90	131	95	7006
	Molecular mass (kDa)	16.78	15.81	212.32	10.91	14.26	11.03	
	Nucleotide coordinates	296-727	724-1161	1127-6667	2310-2582	6307-6702	4212-4499	
	BLASTn similarity	66-99%	73-97%	73-99%	68-96%	71-94%	71-99%	
	BLASTp similarity	26-99%	28-96%	38-98%	54-92%	26-93%	98%	

Table 3.3. Functional conserved domains predicted using the NCBI open reading frame (ORF) Finder tool from the ORFs of the sequenced CSSV genomes and the GenBank references. The letter codes CI (Cote d'Ivoire), GH (Ghana), or TG (Togo) indicate the country from which the cacao samples were collected.

Sequence name	ORF1	ORF2	ORF3	ORF4	ORFX	ORFY
CI311	DUF1319	ND	Zn, Pep, RT, RNase H	ND	ND	ND
CI135	DUF1319	ND	Zn, Pep, RT, RNase H	-	ND	ND
CI275	DUF1319	ND	Zn, Pep, RT, RNase H	-	-	ND
CIS2	DUF1319	ND	Zn, Pep, RT, RNase H	-	ND	ND
CIS3	DUF1319	ND	Zn, Pep, RT, RNase H	-	ND	ND
CI44	DUF1319	ND	Zn, Pep, RT, RNase H	-	ND	ND
GH64	DUF1319	ND	Zn, Pep, RT, RNase H	-	-	ND
GH67	DUF1319	ND	Zn, Pep, RT, RNase H	-	-	ND
GH75	DUF1319	ND	Zn, Pep, RT, RNase H	ND	ND	ND
CI134	DUF1319	ND	Zn, Pep, RT, RNase H	-	ND	ND
CI215	DUF1319	ND	Zn, Pep, RT, RNase H	-	ND	ND
CI286	DUF1319	ND	Zn, Pep, RT, RNase H	-	-	ND
CIT5	DUF1319	ND	Zn, Pep, RT, RNase H	-	ND	ND
CI301	DUF1319	ND	Zn, Pep, RT, RNase H	ND	ND	ND
TG_AJ781003	DUF1319	ND	Zn, Pep, RT, RNase H	-	ND	ND
CI_JN606110	DUF1319	ND	Pep, RT, RNase H	-	ND	DUF3187
TG_L14546	DUF1319	ND	Zn, Pep, RT, RNase H	-	ND	ND
TG_AJ534983	DUF1319	ND	Zn, Pep, RT, RNase H	-	ND	ND
GH_AJ609020	DUF1319	ND	Zn, Pep, RT, RNase H	-	ND	ND
GH_AJ609019	DUF1319	ND	Zn, Pep, RT, RNase H	ND	ND	ND
GH_AJ608931	DUF1319	ND	Zn, Pep, RT, RNase H	-	ND	ND
CaMMV	DUF1319	ND	Zn, BAR, Pep, RT, RNase H	-	-	ND
CYVBV	DUF1319	ND	Zn, Trim, Pep, RT, RNase H	-	-	HTH

Legend: DUF – domain of unknown function, Zn – zinc knuckle finger, Pep – pepsin-like aspartate protease, RT – reverse transcriptase, RNase H – ribonuclease H, BAR - GTPase regulator associated with focal adhesion 2 (Bin/Amphiphysin/Rvs), Trim – trimeric-dUTPase, HTH – helix-turn-helix, ND – no domain predicted.

Table 3.4. Percentage pairwise nucleotide identity for the RT-RNase H region of *Cacao swollen shoot virus* GenBank reference sequences (indicated by country code and GenBank Accession number) and the 14 sequenced genomes (indicated by country code and sample number) in relation to other badnavirus sequences using MUSCLE alignment implemented in SDTv1.2 (Muhire *et al.*, 2014). The shaded boxes indicate at least 80% identities. The letter codes CI (Cote d'Ivoire), GH (Ghana), or TG (Togo) indicate the country from which the cacao samples were collected.

TG_AJ781003	**100**																				
CI_JN606110	**76**	**100**																			
GH67	**74**	**72**	**100**																		
CI286	**74**	**71**	**82**	**100**																	
GH64	**74**	**71**	**82**	**84**	**100**																
CI275	**75**	**72**	**80**	**81**	**84**	**100**															
TG_L14546	**76**	**72**	**76**	**75**	**74**	**74**	**100**														
TG_AJ534983	**76**	**72**	**75**	**75**	**75**	**75**	**98**	**100**													
CI44	**74**	**71**	**75**	**76**	**75**	**75**	**89**	**89**	**100**												
CIS3	**73**	**70**	**71**	**72**	**72**	**72**	**86**	**86**	**89**	**100**											
GH_AJ609020	**74**	**72**	**74**	**76**	**74**	**74**	**89**	**90**	**92**	**93**	**100**										
GH_AJ609019	**74**	**72**	**74**	**75**	**74**	**74**	**90**	**90**	**93**	**94**	**98**	**100**									
CI301	**74**	**72**	**75**	**76**	**74**	**74**	**89**	**89**	**92**	**94**	**97**	**97**	**100**								
CI311	**74**	**72**	**75**	**76**	**75**	**74**	**90**	**90**	**93**	**94**	**98**	**98**	**99**	**100**							
CI135	**74**	**73**	**75**	**76**	**75**	**74**	**90**	**90**	**93**	**94**	**98**	**98**	**97**	**99**	**100**						
GH_AJ608931	**74**	**72**	**75**	**76**	**74**	**73**	**90**	**90**	**93**	**94**	**98**	**99**	**97**	**99**	**99**	**100**					
GH75	**74**	**72**	**75**	**76**	**74**	**73**	**90**	**90**	**93**	**94**	**98**	**98**	**97**	**99**	**99**	**100**	**100**				
CIT5	**75**	**73**	**75**	**76**	**74**	**74**	**90**	**90**	**93**	**94**	**98**	**98**	**98**	**99**	**99**	**99**	**99**	**100**			
CIS2	**75**	**73**	**75**	**76**	**75**	**74**	**90**	**90**	**93**	**94**	**98**	**98**	**97**	**99**	**99**	**99**	**99**	**100**	**100**		
CI215	**74**	**72**	**75**	**76**	**74**	**74**	**90**	**91**	**94**	**94**	**98**	**98**	**98**	**99**	**99**	**99**	**99**	**99**	**99**	**100**	
CI134	**74**	**72**	**75**	**76**	**74**	**74**	**90**	**90**	**93**	**94**	**99**	**99**	**98**	**99**	**99**	**99**	**99**	**99**	**99**	**100**	**100**

Table 3.5. Percentage pairwise nucleotide identity for the complete genomic sequences of *Cacao swollen shoot virus* GenBank reference sequences (indicated by country code and GenBank Accession number) and14 sequenced genomes (indicated by country code and sample number) in relation to other badnavirus sequences using MUSCLE alignment implemented in SDTv1.2 (Muhire *et al.*, 2014). The shaded boxes indicate at least 80% identities. The letter codes CI (Cote d'Ivoire), GH (Ghana), or TG (Togo) indicate the country from which the cacao samples were collected.

CI286	**100**																			
GH64	**78**	**100**																		
CI275	**78**	**82**	**100**																	
GH67	**77**	**80**	**81**	**100**																
CI_JN606110	**71**	**71**	**71**	**71**	**100**															
TG_AJ781003	**72**	**70**	**72**	**72**	**76**	**100**														
TG_L14546	**71**	**71**	**72**	**72**	**75**	**80**	**100**													
TG_AJ534983	**71**	**71**	**71**	**72**	**75**	**80**	**98**	**100**												
CI44	**71**	**71**	**71**	**71**	**75**	**76**	**83**	**83**	**100**											
GH_AJ609020	**72**	**72**	**71**	**71**	**75**	**77**	**85**	**85**	**89**	**100**										
GH75	**72**	**71**	**72**	**71**	**75**	**77**	**85**	**85**	**89**	**97**	**100**									
GH_AJ608931	**72**	**71**	**72**	**71**	**75**	**77**	**85**	**85**	**89**	**97**	**99**	**100**								
GH_AJ609019	**72**	**71**	**71**	**71**	**75**	**77**	**86**	**85**	**89**	**97**	**98**	**98**	**100**							
CIS3	**72**	**71**	**71**	**71**	**75**	**77**	**85**	**85**	**89**	**97**	**97**	**97**	**97**	**100**						
CI311	**72**	**72**	**71**	**71**	**75**	**77**	**86**	**85**	**89**	**97**	**98**	**98**	**98**	**98**	**100**					
CI301	**72**	**72**	**71**	**72**	**75**	**77**	**86**	**85**	**89**	**97**	**97**	**98**	**98**	**98**	**99**	**100**				
CI135	**72**	**72**	**71**	**71**	**75**	**77**	**86**	**85**	**89**	**98**	**98**	**98**	**98**	**98**	**98**	**98**	**100**			
CIT5	**72**	**72**	**71**	**71**	**75**	**77**	**86**	**85**	**90**	**97**	**98**	**98**	**98**	**98**	**98**	**98**	**98**	**100**		
CIS2	**72**	**71**	**71**	**71**	**75**	**77**	**85**	**85**	**89**	**97**	**98**	**98**	**98**	**98**	**98**	**98**	**98**	**100**	**100**	
CI215	**72**	**72**	**72**	**71**	**75**	**77**	**86**	**85**	**89**	**97**	**97**	**97**	**98**	**98**	**98**	**98**	**98**	**98**	**98**	**100**
CI134	**72**	**72**	**71**	**71**	**75**	**77**	**86**	**85**	**89**	**97**	**97**	**97**	**97**	**98**	**98**	**98**	**98**	**98**	**98**	**98**

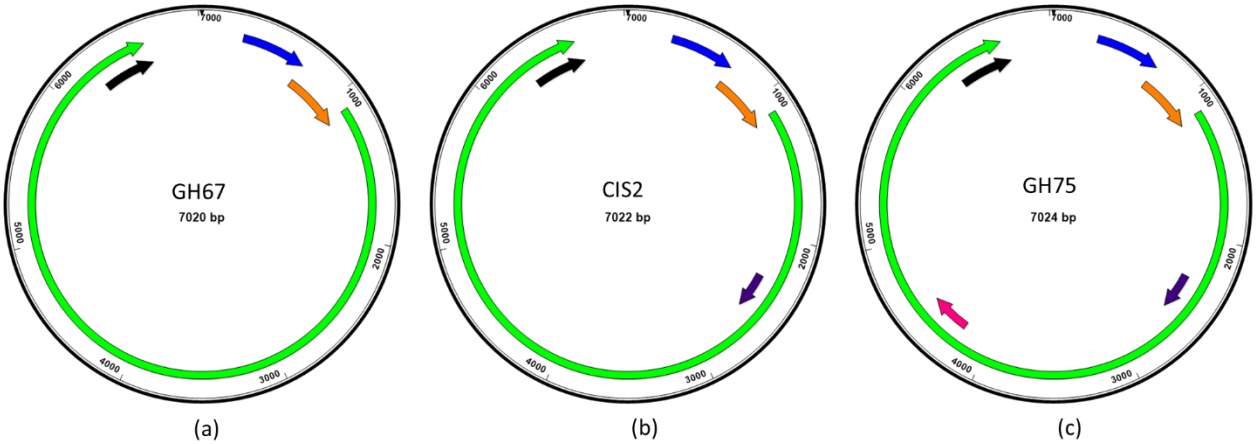

Fig. 3.1. The representative genome maps of the three types of open reading frame (ORF) arrangement of the 14 *Cacao swollen shot virus* (CSSV) genome sequences determined in this study. Four, five or six ORFs were predicted using the NCBI ORF Finder tool on each genome, and the putative coding regions for the ORFs are indicated by filled arrows; blue - ORF1, orange - ORF2, green - ORF3, pink – ORF4, purple - ORFX and black - ORFY. The sizes of all the genomes were between 6,920 and 7,172 bp. The GH67 genomic map (a) is shown as an example for the four ORF arrangements, CI275, GH64 and CI286 had similar maps. Genomes with five ORFs, as exemplified by the CIS2 map (b), were also observed on CIS3, CI134, CI135, CIT5, CI44 and CI215 genome. The only three genomes with six ORFs, represented by the GH75 map (c), were also present in CI311 and CI301. The small black arrow head at the top of each genome map indicates the coordinates of the first nucleotide, also the 5' end of the predicted $tRNA^{met}$ binding site. The letter codes CI (Cote d'Ivoire), GH (Ghana), or TG (Togo) indicate the country from which the cacao samples were collected.

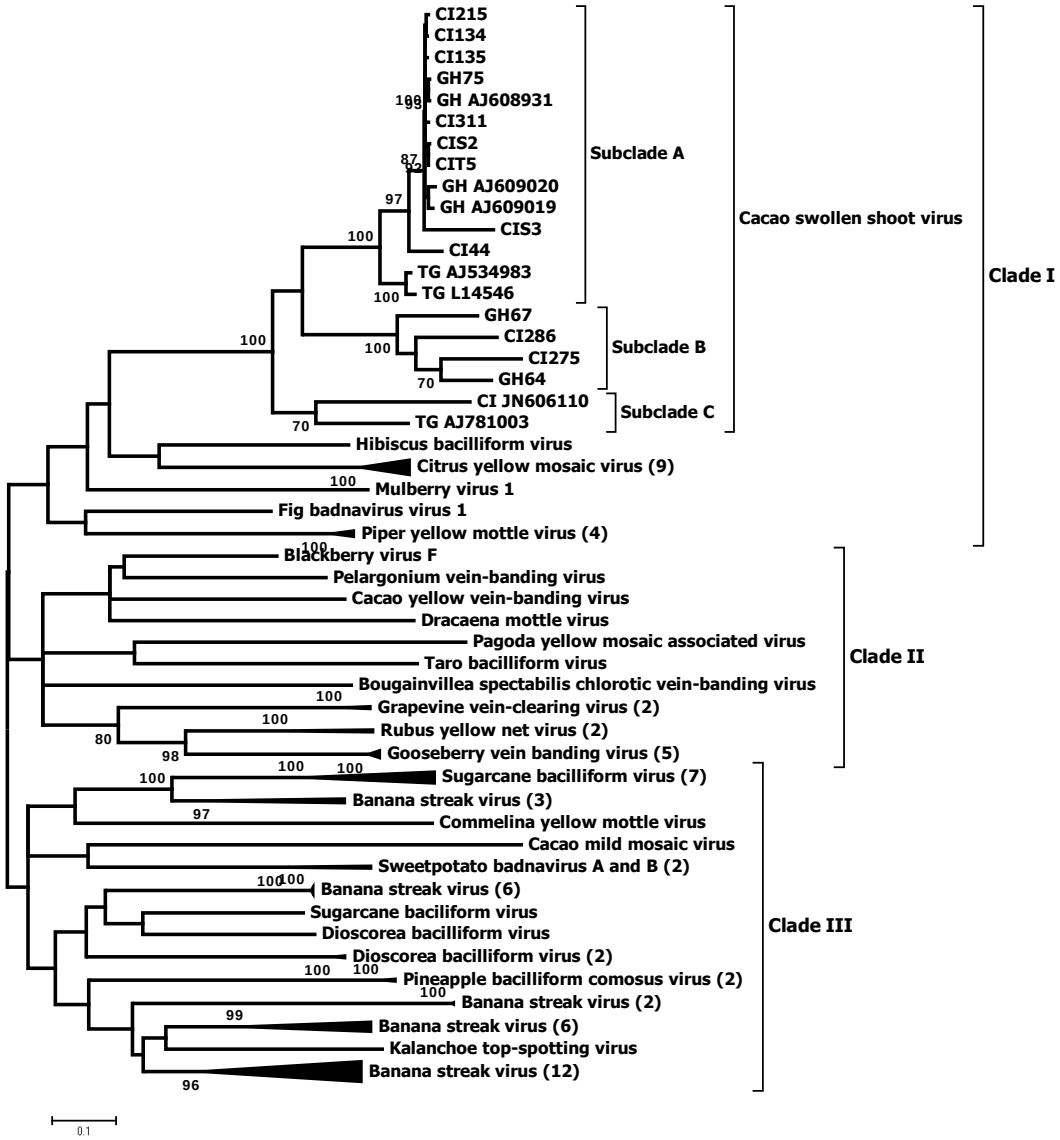

Fig. 3.2. Maximum Likelihood phylogenetic tree of the 580 bp RT-RNase H region of 14 *Cacao swollen shoot virus* (CSSV) genomic sequences and 87 published badnavirus genomes. Analyses were done using the MEGA6 software (Tamura *et al.*, 2013). The horizontal branch lengths are proportional to the genetic distance, and numbers shown at branch points indicate bootstrap values from 1000 replicates, at >70%. Each node was collapsed into one taxon, where more than one sequence is available for each species, and the number of sequences used is indicated in parentheses. The sequences determined from this study are indicated by country code followed by sample number, and the CSSV GenBank reference are indicated by the country code and the GenBank Accession number. The letter codes CI (Cote d'Ivoire), GH (Ghana), or TG (Togo) indicate the country from which cacao samples were collected.

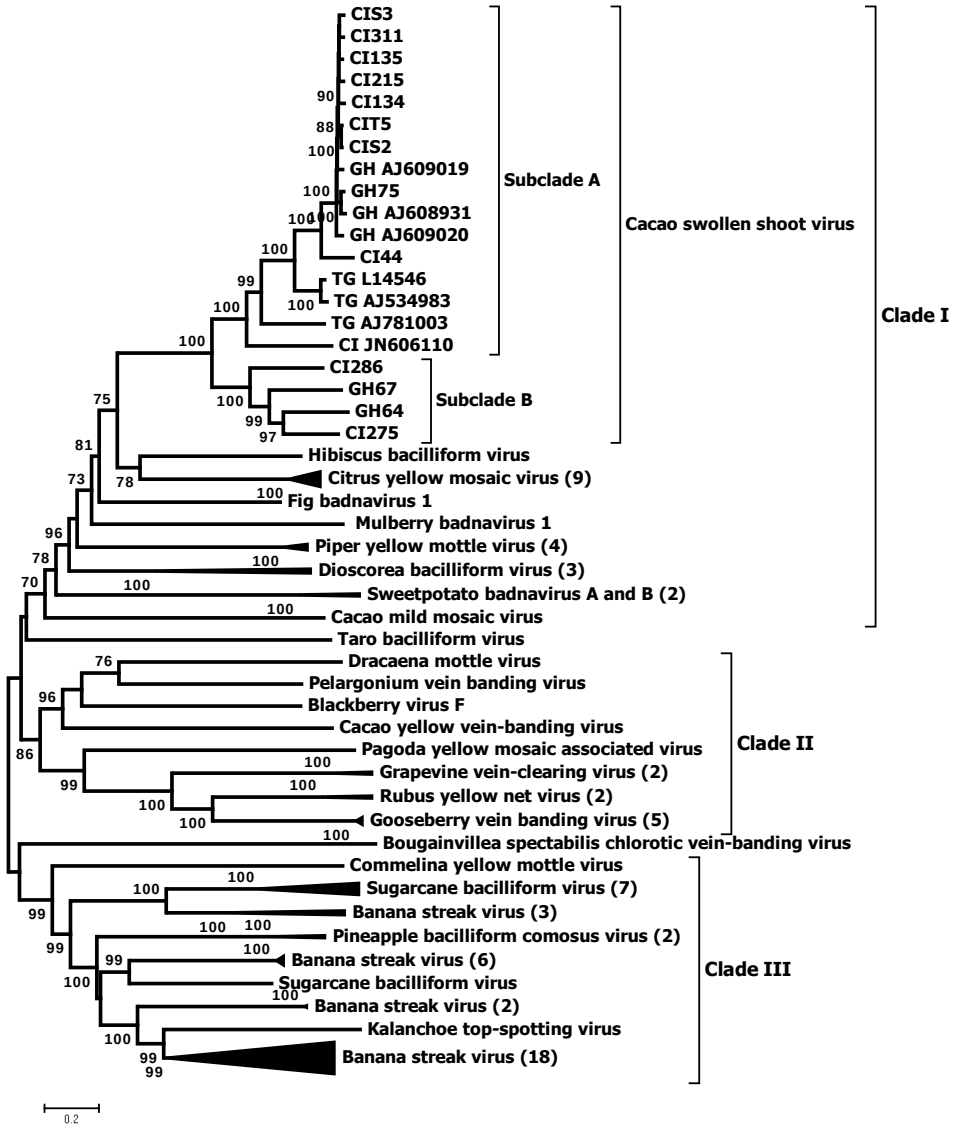

Fig. 3.3. Maximum Likelihood phylogenetic tree of the 14 *Cacao swollen shoot virus* (CSSV) complete genome sequences and 87 published badnavirus genomes. Analyses were done using the MEGA6 software (Tamura *et al.*, 2013). The horizontal branch lengths are proportional to the genetic distance, and numbers shown at branch points indicate bootstrap values >70% from 1000 replicates. Each node was collapsed into one taxon, where more than one sequence is available for each species, and the number of sequences used is indicated in parentheses. The sequences determined from this study are indicated by country code followed by sample number, and the CSSV GenBank reference are indicated by the country code and GenBank Accession number. The letter codes CI (Cote d'Ivoire), GH (Ghana), or TG (Togo) indicate the country from which cacao samples were collected.

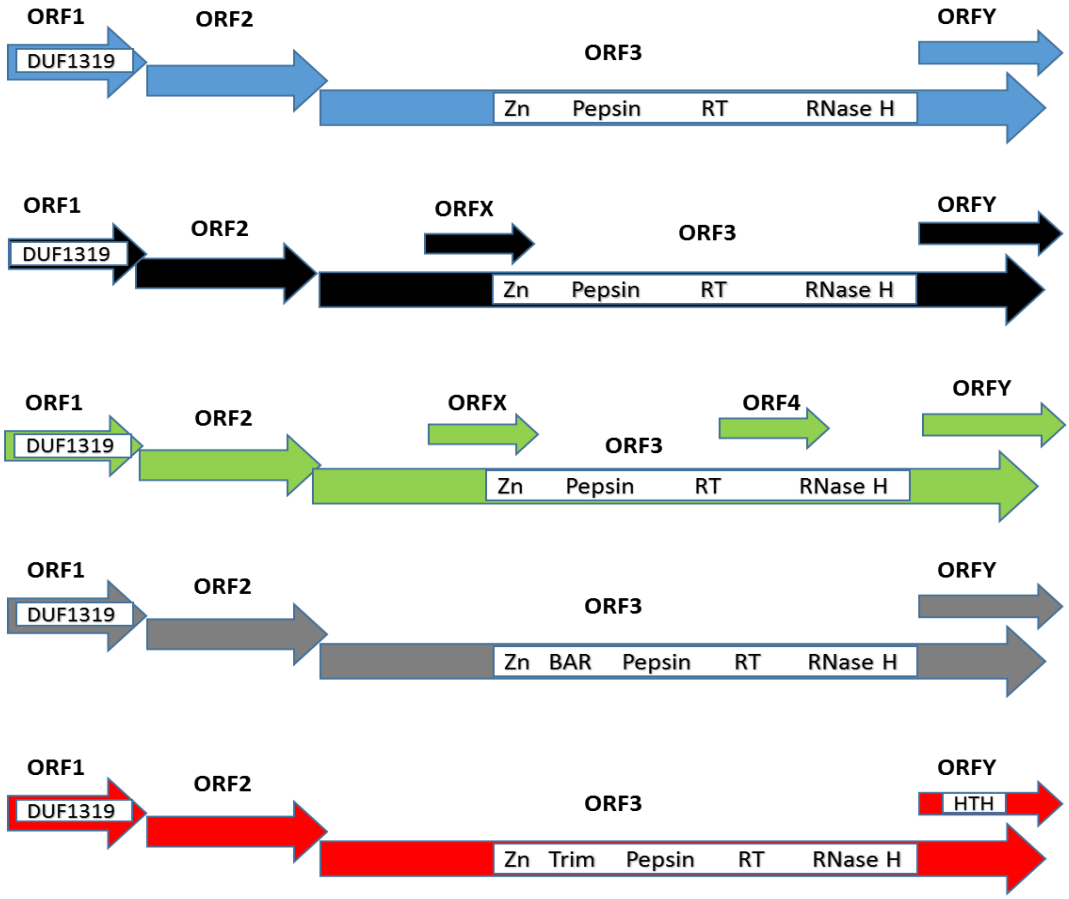

Fig. 3.4. The predicted conserved functional domains for the three types of CSSV genome arrangement are shown in linear genome maps (not to scale) together with the genomes of two other cacao-infecting badnaviruses; Cacao mild mosaic virus (CaMMV) and Cacao yellow-vein banding virus (CYVBV), for comparison. Linear genome maps for the CSSV genomes with four, five and six ORFs are shown in blue, black and green, respectively, and the genome maps for CaMMV and CYCBV are shown in gray and red, respectively. DUF1319 – domain of unknown function 1319, Zn – zinc knuckle finger, Pepsin – pepsin-like aspartate protease, RT – reverse transcriptase, RNase H – ribonuclease H, BAR - GTPase regulator associated with focal adhesion 2 (Bin/Amphiphysin/Rvs), Trim – trimeric-dUTPase, HTH – helix-turn-helix.

CHAPTER 4

IDENTIFICATION AND CHARACTERIZATION OF PREVIOUSLY ELUSIVE BADNAVIRUS SPECIES ASSOCIATED WITH SYMPTOMATIC *THEOBROMA CACAO* IN TRINIDAD

Abstract

Suspect virus-like symptoms were observed in cacao plants in Trinidad during 1943, and designated as the strain A and B isolates of Cacao Trinidad virus (CTV), however, viral etiology has not been demonstrated for either phenotype. Total DNA was isolated from symptomatic cacao leaves exhibiting the CTV A and B phenotypes, and subjected to Illumina HiSeq and Sanger DNA sequencing validation. Based on the *de novo* assemblies, two apparently full-length badnavirus genomes of 7,533 and 7,454 nucleotides (nt) were associated with CTV strain A and B, respectively. The Trinidad badnaviral genomes encoded four predicted open reading frames (ORFs), three of which are characteristic of other known badnaviruses, and a fourth that is encoded by some species. Both badnaviral genomes harbored hallmark caulimovirus-like features including a $tRNA^{met}$ priming site, a TATA box, and a polyadenylation-like signal. Pairwise comparisons of the RT-RNase H region using the Sequence Demarcation Tool software indicated the Trinidad isolates shared 57-71% nt identity with other known badnaviruses. Based on the badnavirus species demarcation for this genomic region, at <80% nt identity, the isolates represent two previously unidentified badnaviruses, herein named Cacao mild mosaic virus and Cacao yellow vein-banding virus, making them the first cacao-infecting badnaviruses identified in the Western Hemisphere.

Introduction

Theobroma cacao L. (cacao) trees produce the beans from which chocolate is made, making it an economically important crop in its New World center of diversity in the Amazon Basin and the nearby neotropical countries of domestication, and in Africa where it was introduced for commercial production over one hundred years ago. West African countries market mostly bulk

cocoa, whereas cocoa beans produced in the Caribbean, Central and South America, and the Pacific region are used for bulk cocoa or for specialty cocoa products (Gray, 2001).
A major constraint to cocoa production in West Africa is the *Cacao swollen shoot virus* (CSSV) (genus, *Badnavirus*; family, *Caulimoviridae*), which causes significantly reduced yields, and decline and death within 3-5 years after CSSV infection. In Ghana alone, during 2007-2010, outbreaks of swollen shoot disease resulted in losses of more than 100,000 hectares. The removal of diseased trees practiced to reduce virus spread has been estimated at $84.9 million US dollars (World Bank. 2013; http://documents.worldbank.org/curated/en/2013/01/17694705/ghana-cocoa-supply-chain-risk-assessment). Thus far, among the cacao producing regions of the world, CSSV occurs only in West Africa.

Badnaviruses are plant pararetroviruses classified in the *Caulimoviridae* family. They are characterized by having a circular double-stranded (ds) DNA genome of 7.2 - 9.2 kilobases (kb) in size. Badnavirus particles (virions) are non-enveloped and have a bacilliform morphology. Virions are approximately 30-50 nm wide by 60–900 nm long, and have a mean length of 130 nm (King *et al.*, 2012). Most badnaviral genomes have three open reading frames (ORFs 1-3), encoded on the viral sense strand (King *et al.*, 2012). The function of the predicted protein encoded by the badnaviral ORF1 is not known. In CSSV, the ORF2 encodes a predicted protein of approximately 14 kDa that has been associated with DNA and RNA binding activity (Jacquot *et al.*, 1996). The largest predicted protein, ORF3, encodes a polyprotein of approximately 25 kDa, and contains domains located at the 5'end attributed to within-plant viral movement. The remainder of ORF3 encodes the capsid, the aspartic protease, the viral reverse transcriptase (RT), and the ribonuclease H (RNase H) proteins. In CSSV there are two additional predicted coding regions that overlap with ORF3, referred to as ORFX and ORFY, at 13 and 14 kDa in size, respectively. The function of the ORFs X and Y predicted proteins is not known (Jacquot *et al.*, 1999a). Several members of the genus, *Badnavirus* have a predicted ORF that shares homology with the ORFY of CSSV that is referred to as ORF4 or ORF6 in the respective viruses (Borah *et al.*, 2009).

In Trinidad, virus-like disease symptoms were first reported in *T. cacao* trees by Posnette during 1943 (Posnette, 1944). The disease symptoms were manifest either as 'red mottling' or 'vein-

clearing' (Posnette, 1944), and thereafter the isolates were referred to as strain A and strain B, respectively, (Baker & Dale, 1947a, b) of Cacao Trinidad virus (CTV) (Thorold, 1975). The strain B symptoms were later corrected from 'vein-clearing' to 'yellow vein-banding' after the observation that the veins appeared more yellow than 'cleared' (Kirkpatrick, 1950). Cacao plants affected by CTV strain A developed a feather-like red banding on several or all of the main veins of newly developing (flushing) leaves, but as leaves matured, the red vein-banding symptom disappeared (Fig. 4.1). In some cacao clones mosaic symptoms developed along the main veins and persisted in mature leaves. Other clonal varieties exhibited mosaic and red-mottling symptoms on newly developing leaves, whereas, others developed a mosaic symptom but lacked the mottling phenotype. Symptoms were reported to develop primarily on the youngest leaves developing during and after the rainy season, commonly referred to as 'flush growth', whereas, the older leaves of the same tree were asymptomatic but harbored the putative virus, based on results of mealybug vector transmission studies (Sreenivasan, 2009). In addition, the CTV strain A caused red discoloration symptoms on young pods of clones ICS 6 and ICS 8. Although symptoms caused by these two strains shared certain similarities among the genetically diverse cacao clones grown in Trinidad, strain B was reported to cause persistent yellow vein-banding symptoms in the major and minor veins of mature leaves, sometimes accompanied by red vein-banding (Baker & Dale, 1947a). The ICS 6 clone has been identified as a differential indicator plant for the A and B strain in that, ICS 6 inoculated with strain A developed red vein-banding, and in contrast, red vein-banding and mosaic symptoms were observed when inoculated with strain B, phenotypes previously reported as mild and severe, respectively (Baker & Dale, 1947a).

Experimental transmission studies have shown that the mealybugs *Pseudococcus citri* (Risso), *P.brevipes* (Ckll.), *P. comstocki* (Kuw.), and *Ferrisia virgata* (Cockerell) transmitted both strains of CTV (Baker & Dale, 1947b; Kirkpatrick, 1950). Transmission of the putative virus (es) by the suspect mealybug vectors using an acquisition access period of 33 min, followed by an inoculation access period of 90 min. Given these characteristics, the suspect viruses has been considered to be transmitted by the mealybugs in a non-persistent manner (Kirkpatrick, 1950).

When the disease was first discovered in the Diego Martin Valley, which is located in the northwest region of the Northern Range in Trinidad, its distribution was thought to be restricted

there (Posnette, 1944). However, island-wide surveys later revealed that symptoms of both strains were prevalent in the Santa Cruz and Maracas Valleys (Baker & Dale, 1947b; Posnette, 1944), and in Blanchisseuse and Toco (Baker & Dale, 1947b; Swarbrick, 1961), suggesting that the putative viral pathogens were present in Trinidad and Tobago for some time. As a result, the government mandated an eradication program during the 1950's that destroyed all cacao trees exhibiting CTV symptoms. Thereafter, strain A and B virus-like symptoms were not observed in cacao trees in Trinidad and Tobago until fourteen years ago when several trees maintained in the Trinidad International Cocoa Genebank (ICGT) germplasm collection developed symptoms characteristic of the strain A- and B-like suspect viruses (Sreenivasan, 2009).

The Cocoa Research Centre (CRC, formerly, the Cocoa Research Unit) regularly transfers cacao germplasm to other countries from the ICGT, and is the recipient of germplasm through the International Cocoa Quarantine Centre-Reading (ICQCR), located in Reading, UK. There germplasm is quarantined and indexed by grafting to CSSV-susceptible 'Amelonado' seedlings. Grafted scions are inspected for symptom development over a period of two years before re-distribution to cocoa-growing countries. In 2002, the ICQCR reported the occurrence of virus-like symptoms on cacao accession ICS 76, which had been transferred from Trinidad for routine virus indexing. In 2005, virus-like symptoms were detected on ICS 76 budwood from Trinidad provided to the ICQCR (Sreenivasan, 2009). A follow-up inspection of the ICS 76 mother tree at the ICGT revealed foliar red mottling symptoms on flush growth reminiscent of those reported for strain A of CTV (Baker & Dale, 1947b) (Fig. 4.1). In addition, cacao clone ICS 27 showed yellow vein-banding symptoms like those associated with CTV strain B. Despite suspect badnavirus etiology, all attempts have failed to detect a suspect virus in symptomatic cacao leaves (strains A and B) by polymerase chain reaction (PCR) amplification using primers designed for detection of West African CSSV isolates, or general badnavirus primers (Yang *et al.*, 2003). In addition, no virus-like particles were observed by transmission electron microscopy (TEM) in negatively-stained leaf dip preparations made from leaves showing strain A and strain B symptoms (P. Umaharan, unpublished data).

The inability to identify the putative viral causal agent(s) associated with CTV symptoms in cacao in Trinidad or in the ICQCR (UK) has hindered the development of molecular diagnostic

testing of quarantined cacao germplasm and undermined the safe distribution of certified, virus-free germplasm to recipient countries. Also, graft-indexing to detect cacao-infecting plant viruses is time-consuming, and further, virus-like symptoms can be confused with nutritional imbalance in cacao. To determine the identity of the suspect badnavirus (es) associated with the symptomatic cacao trees in Trinidad, total DNA was isolated from leaves showing symptoms of CTV strain A and B. The purified DNA was subjected to DNA sequencing using the Illumina HiSeq platform and for follow-up validation, to PCR amplification and Sanger DNA sequencing of cloned viral fragments. Results indicated that a distinct badnaviral genome was associated with each of the symptom phenotypes, herein named as unique badnaviral species, making this the first report of cacao-infecting viruses in Trinidad and in the American Tropics.

Materials and Methods

Plant samples. The symptomatic ICS 76 and ICS 27 cacao trees maintained at ICGT were established successively through vegetative propagation from an earlier collection that had existed since the 1930s. Bud wood from ICS 76 showing symptoms of strain A was collected and grafted onto the genotype SCA 6 rootstocks to experimentally transmit the suspect virus. Infected leaves showing symptoms typical of strain B infection were collected from ICS 27 plants. Characteristic disease symptoms of cacao leaves associated with strain A and B are shown in Fig. 4.1. Leaves were washed with water, rinsed three times in sterile distilled water, and preserved in 100% glycerol prior to shipment to the School of Plant Sciences, University of Arizona, Tucson, AZ, USA.

Total DNA isolation. Prior to DNA isolation, the cacao leaf samples were washed to remove the 100% glycerol storage solution, and blotted dry. For each sample, 100 mg of leaf tissue was ground in liquid nitrogen using a sterile plastic pestle in a 1.6 mL Eppendorf tube, followed by transfer to a 2-mL tube containing 1.2 mL of cetyl trimethylammonium bromide (CTAB) (Doyle & Doyle, 1990) with 2 % β-mercaptoethanol (Sigma-Aldrich, St. Louis, MO, USA), and four 3.2 mm stainless steel beads (Next Advance, NY, USA). The samples were pulverized for 5 min by placing tubes in a Mini Beadbeater^{TM} (Biospec Products, OK, USA). The DNA was purified according using the procedure of Doyle and Doyle (1990). The resultant DNA pellet was dissolved in 300 µL low TE buffer, containing 10 mM Tris-HCL and 0.1 mM EDTA (pH 8.0).

To minimize potential carry over of non-viral circular DNA (e.g. plant mitochondrial and chloroplast), purified DNA was filtered through a 0.1 µm pore Ultrafree-MC column (Millipore, MA, USA) and collected by microcentrifugation at 12,000 x g for 4 min. The filtrate was precipitated in 100% ethanol, with the addition of sodium acetate to a final concentration of 0.3 M. The DNA was pelleted by centrifugation at 12,000 x g for 10 min, washed with 70% ethanol, and redissolved in 100 µL low TE buffer, pH 8.0, prior to 20 °C storage.

Illumina sequencing, assembly of paired-end reads, and bioinformatics. Purified DNA was used to prepare paired-end libraries using a TruSeq PE cluster kit, with a mean insert size of 350 bp, and individually tagged. Libraries were sequenced using the Illumina HiSeq 2500 platform at the University of Arizona Genomics Core (UAGC, Tucson, Arizona, USA). Reads were de-multiplexed and quality was assessed with FASTQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). Adapters were removed and reads were trimmed with TRIMMOMATIC v0.32 (Bolger *et al.*, 2014). The *de novo* assemblies were carried out using DNASTAR SeqMan NGen v.12 (DNASTAR, WI, USA) using the option to filter background sequences. The sequences were filtered to remove host plant sequences using the *T. cacao* genome sequences (Argout *et al.*, 2011; Motamayor *et al.*, 2013) (GenBank Accessions FR7222157.1, CM001879.1 to CM001888.1, and KE132922.1), the cacao chloroplast genome (GenBank Accession NC014676.2), and the mitochondrial genome of a related malvaceous species, *Gossypium hirsutum* L. (GenBank Accession JX065074.1), all downloaded from NCBI GenBank database. The assembled contigs for each sample were separately annotated using BLAST2GO software (Conesa & Gotz, 2008) and then subjected to a BLASTn search (Altschul *et al.*, 1990) using the algorithm available at the NCBI GenBank database website (https://blast.ncbi.nlm.nih.gov/Blast.cgi). The two apparently full-length badnaviral genome sequences were arranged according to the precedent established for badnaviral genomes in that the conserved, first nt of the predicted $tRNA^{met}$ primer binding site is designated as number 1.

Sanger DNA sequencing validation of badnavirus genomes. Circular viral DNA was preferentially enriched using rolling circle amplification (RCA) that employs phi29 DNA polymerase (Dean *et al.*, 2001), available in the Templiphi RCA kit, (GE Healthcare Bio-Sciences Corp, NJ, USA), according to the manufacturer's instructions with modifications, as

previously described (Rector *et al.*, 2004; Stevens *et al.*, 2010). RCA reactions contained 2µL purified DNA in 10 µL of sample buffer. The preparation was denatured at 95°C for 3 min, and cooled on ice for 3 min. The RCA enzyme mixture, which contained 10 µL of the reaction buffer, 0.4 µL of the Templiphi enzyme mix, and 450 µM of dNTP mix (Sigma-Aldrich, MO, USA), was added to the denatured DNA. The reaction mixture was incubated at 30°C for 18 hrs, followed by enzyme deactivation at 65°C for 15 min. After cooling to 4°C, the product was used as template for PCR amplification of badnaviral genomic fragments.

Based on the resultant putative badnavirus genome sequences determined by Illumina sequencing from cacao leaves showing typical strain A and B symptoms (Fig. 4.2), four pairs of primers were designed and used to PCR-amplify two approximately 4000 base pair (bp) fragments of each genome, with overlapping fragments of 250-500 bp. The virus-specific primer (underlined), and expected amplicon sizes for strain A, herein, Cacao mild mosaic virus (CaMMV), were CaMMV_1F 5'-ATCACTAGT*GCGGCCGC*TGGTATCAGAGCTATGTTTG-3' and CaMMV_1R 5'-GACCTGCAG*GCGGCCGC*TCGTCGAATTGGTCATCCT-3' (4,000 bp), CaMMV_2F 5'-ATCACTAGT*GCGGCCGC*GTCTCACGCTGGATTGGG-3' and CaMMV_2R 5'- GACCTGCAG*GCGGCCGC*TCCTATTCAAAGCTCATACAA-3' (4,083 bp). For strain B, herein, Cacao yellow vein banding virus (CYVBV), the primers were CYVBV_1F 5'- ATCACTAGT*GCGGCCGC*TGGTATCAGAGCAAGGTTAT-3' and CYVBV_1R 5'-GACCTGCAG*GCGGCCGC*AAATTCCTCCCCTGTACAAT-3' (4,000 bp) and CYVBV_2F 5'- ATCACTAGT*GCGGCCGC*TGGGATGTAGTCTCAGTTG-3' and CYVBV_2R 5'-GACCTGCAG*GCGGCCGC*ATAGCTCACCTTATCGCCT-3' (3,955 bp). The *Not* I restriction site was added to each primer (italics) to facilitate cloning.

Each PCR amplification reaction was carried out using the Invitrogen CloneAmp^{TM} HiFi PCR Premix (Clonetech Laboratories, CA, USA) according to the manufacturer's instructions. The reaction mixture contained 1X CloneAmp^{TM} HiFi PCR Premix, 0.2 µM each of the reverse and forward primer, 2 µL of the RCA product as template, and nuclease-free water to a total volume of 50 µL. PCR conditions were: initial denaturation at 98°C for 2 min, followed by 40 cycles of denaturation at 98°C for 20 sec, annealing at 55°C for 15 sec, extension at 72°C for 4 min, with a final extension at 72°C for 10 min. The PCR products were fractionated by electrophoresis on a

0.8% agarose gel (90 min, 100 V) stained with GelGreen (10 µL/mL) stain (Biotium, Aurora, CO, USA), in 1X Tris-acetate EDTA (TAE) buffer, pH 8.0. The bands of ~4 kbp in size were excised and gel-purified using the purification kit available from GE Healthcare Bio-Sciences (NJ, USA), and the concentration of the DNA was determined using the Nanodrop 2000 UV-Vis spectrophotometer (Thermo Scientific, DE, USA). The purified DNA was ligated into the pGEM5 plasmid vector (Promega, WI, USA) previously digested using the restriction endonuclease *Not* I (New England Biolabs, CA, USA). Ligation and transformation were carried out using the In-Fusion HD Cloning kit (Clontech laboratories, CA, USA), according to the manufacturer's instructions. Insert sizes were confirmed by isolating plasmids from white colonies using the GeneJET Plasmid Miniprep Kit (ThermoFisher Scientific, USA), according to the manufacturer's instructions, followed by *Not* I digestion. The resultant products were fractionated by agarose gel electrophoresis, as described. At least three clones with the correct insert size were selected for each isolate and subjected to bi-directional capillary (Sanger) DNA sequencing using primer walking at Eton Bioscience (San Diego, CA, USA).

The overlapping DNA sequences obtained for each clone were assembled using DNASTAR SeqMan Pro v.12 (DNASTAR, WI, USA) and the ORFs were analyzed and rearranged as previously described. The ORFs, functional domains and full-length nt sequences obtained by Sanger sequencing were then compared with those obtained using the Illumina platform using NCBI ORF Finder, CDD database and SDT, respectively, as outlined above.

Badnavirus genome characterization. The genome sequences for 120 badnavirus isolates, representing 35 species, were downloaded from the NCBI GenBank database. To eliminate redundant sequences, a haplotype search was implemented using the FaBox tool (Villesen, 2007). The resultant 84 haplotypes were used for genome characterization, and for pairwise and phylogenetic analyses.

The open reading frames (ORFs) for coding regions comprising >100 amino acids were identified for each of the two apparently full-length badnaviral genomes determined herein using the ORF finder tool (http://www.ncbi.nlm.nih.gov/gorf/gorf.html), available at the NCBI GenBank website. The ORF predictions implemented the standard genetic code.

The BLASTp search tool (Altschul *et al.*, 1990) was used to identify homologous ORFs and to compare ORFs between the previously unidentified CaMMV and CYCCV genomic sequences from cacao plants in Trinidad, and the full-length badnavirus genome sequences available in GenBank. The conserved badnaviral domains were predicted using the CDD tool on NCBI (Marchler-Bauer *et al.*, 2015).

Pairwise nucleotide comparisons and phylogenetic analyses. To determine the pairwise nt identities, the badnaviral partial genome segment encoding the badnavirus conserved RT-RNase H region, and the full-length genomic sequences were aligned for 84 haplotypes using the MUSCLE algorithm, followed by pairwise nt analysis using the Sequence demarcation tool (SDTv1.2) (Muhire *et al.*, 2014).

The RT-RNase H region (580 bp) corresponds to the genome segment delimited by the BadnaFP/RP primers (Yang *et al.*, 2003), while also representing the partial genome locus accepted for species demarcation of badnaviruses by the International Committee on Taxonomy of Viruses (ICTV) (http://www.ictvonline.org) at <80% shared nt identity (King *et al.*, 2012). The RT-RNase H and full-length genomic sequences, respectively, were aligned using the MUSCLE algorithm (Edgar, 2004) implemented in CLC Sequence Viewer 7.5, available at http://www.clcbio.com/products/clc-sequence-viewer.

Similarly, the 84 partial and complete genome sequences were subjected to phylogenetic analysis using the Maximum Likelihood (ML) algorithm implemented in MEGA 6. Trees were reconstructed using the General Time Reversible model and Gamma distribution with Invariable sites, for 1000 bootstrap replicates (Tamura *et al.*, 2013).

Results

Illumina sequencing, PCR amplification and Sanger DNA sequencing

Two apparently full-length badnaviral genome sequences were obtained from the total DNA isolated from symptomatic cacao leaves showing characteristic CTV strain A (CaMMV) and B (CYVBV) symptoms, respectively, using the Illumina HiSeq and capillary Sanger DNA sequencing platforms. The Illumina sequencing resulted in 2,111,947 and 3,664,734 total

sequence reads, of which 1,084,938 and 15,355 were assembled for CaMMV and CYVBV, respectively. Assembly of CaMMV sequences resulted in 796 contigs, while 17 contigs were obtained from CYVBV sequence assembly. Of the assembled contigs, only one viral sequence contig, was identified for each strain. The contig from the CaMMV sample was 7630 bp long, and had a total of 7591 sequence reads with a depth of coverage ranging from 60 – 160 sequences. On the other hand, the contig from CYVBV was 7502 bp in length and the depth of coverage for the corresponding 2627 sequence reads ranged from 20 – 75 sequences. Both contigs had repeated and overlapping sequences on each end, suggesting that the genomes were full-length and circular in nature.

The CaMMV and CYVBV genome sequences determined using the parallel platforms were identical, at 7533 bp or nearly so, at 7454-7458 bp, in size and each contained the same hallmark domains and coding regions, respectively, regardless of the sequencing platform. The PCR-amplification using the CaMMV-specific primers, CaMMV_1F/1R and CaMMV_2F/2R, and the CYVBV-specific primers, CYVBV_1F/1R and CYVBV_2F/2R, all designed based on the respective Illumina sequence, yielded the expected size amplicons of approximately 4 kbp for each fragment, including an overlap of 500 bp, with about 250 bp on each end, respectively.

Assembly of the two ~4.0 kbp amplicons obtained for each of the Trinidad badnavirus isolates yielded an apparently, full-length genome sequence, compared to the sequences obtained using the Illumina approach. Specifically, the CaMMV genome sequences determined by both Illumina and Sanger sequencing were each 7533 bp (Fig. 4.2). Pairwise nt comparisons using SDT analysis (Muhire *et al.*, 2014) indicated that the putative CaMMV PCR-amplicon shared 99.3% nt identity with the CaMMV sequence determined using the Illumina platform. The CYVBV genome was 7454 bp in size based on Illumina sequencing (Fig. 4.2), whereas, the Sanger-determined sequence was larger, by one amino acid, at 7458 bp. The two CYVBV sequences shared 98.8% nt identity with each other. The genome sequences assembled from contigs determined using the Illumina platform were deposited in the NCBI GenBank database. The strain A isolate was given the name, CaMMV and assigned the GenBank Accession No. KX276640, whereas, the strain B isolate was named CYVBV, and assigned the GenBank Accession No. KX276641.

Comparison of badnaviral genome and coding region sequences

The ORF size and arrangement for the CaMMV and CYVBV Illumina- and Sanger-determined badnaviral genome sequences were compared with respect to sequence length, nt alignment, arrangement of coding and non-coding regions, and by pairwise comparisons. Based on all the criteria considered the genome sequences did not deviate from one another in any discernible way, except with respect to pairwise nt divergence, at 2% or less, regardless of the sequencing approach. Thus, the remainder of the analyses have considered only the Trinidad badnaviral genome sequences that were determined using the Illumina platform.

ORF analysis predicted four open reading frames on the positive sense strand of both the CaMMV and CYVBV genomes (Fig. 4.2). The comparisons of the nt and deduced amino acid (aa) size, the molecular weight (mass), and the percentage nt identity and aa similarity for predicted ORFs for the badnavirus haplotypes available from the GenBank database and the two Trinidad cacao badnaviruses are summarized in Table 4.1. The CaMMV genome contained a 3-nt overlap in ORF1 located between nt coordinates 284 – 715, and in ORF2, between the nt coordinates 712 – 1104. Another overlap of 3-nt was identified between ORFs 2 and 3, located at the nt coordinates 1101 – 6989. Pairwise nt comparisons for all ORFs between CaMMV and the other badnavirus sequences from the GenBank database indicated that they ranged from 52–71%.

The BLASTp analysis of each CaMMV ORF against other badnaviruses indicated that the aa similarities ranged from 27–57% for ORF1, 25–40% for ORF2, 35–64% for ORF3, and 24–45% for ORFY. The fourth predicted coding region, ORFY (6632–7027), had a 357-nt overlap with ORF3 at the 5'-end. The ORFY shared 56 and 63% nt identity with the CSSV ORFY. The estimated molecular weight (mass) for ORFs1, 2, 3 and Y was 16.47, 14.33, 225.92, and 14.85 kDa, respectively.

For CYVBV, four predicted ORFs were arranged like those of CaMMV, and they both had ORF arrangements like that found in most previously studied badnaviral genomes (GenBank). The coding regions for CYVBV were located at the nt coordinates 295–816 for ORF1, 817–1215 for

ORF2, and 1212-7088 for ORF3. The ORF3 had a 3-nt overlap with ORF2. The pairwise nt identities ranged from 52–67% for all ORFs between CYVBV, and with the badnavirus genome sequences from the GenBank database, analyzed here.

The BLASTp comparison of the CYVBV ORFs with the analogous coding region for previously studied badnaviruses indicated they shared aa similarities of 23–43% for ORF1, 22–34% for ORF2, and 34–69% for ORF3. In contrast to ORFY of CaMMV, the CYVBV ORFY was located at nt coordinates 6788–7165 and overlapped with its ORF3 by 300 nt. The estimated molecular weight for the predicted coding regions of CYVBV was 20.32, 14.45, 221.87 and 14.15 kDa for ORF1, ORF2, ORF3 and ORFY, respectively. The sizes of the latter ORFs were within range found for the other members classified in the badnavirus genus, respectively, except for the CYVBV ORF1, which was slightly larger (molecular mass) than other known badnaviruses. Also, badnaviruses known thus far usually have a 3-nt overlap between ORF1 and ORF2, however, these ORFS did not overlap for CYCCV, and instead, were immediately adjacent to each other. In contrast, ORF2 and ORF3 of the same strain overlapped by 3 nt, also typical of badnavirus genomes. The BLASTp analysis of the ORFY of CYVBV indicated that it did not share homology with any viral protein for which sequences are available in the GenBank database.

Badnavirus genome conserved domains and motifs

Using the NCBI Conserved Domain Database (CDD; http://www.ncbi.nlm.nih.gov/cdd/) (Marchler-Bauer *et al.*, 2015) to identify badnavirus conserved domains in each ORF (Table 4.1), similar domains were predicted in ORFs 1 and 2 for the CaMMV and CYVBV genomes. The ORF1 contained a predicted domain of unknown function belonging to a previously reported badnavirus-specific, uncharacterized protein superfamily, referred to as 'domains of unknown function', or DUF1319 (Bateman *et al.*, 2010), while no such domain was identified in ORF2 or ORFY.

The ORF3 and ORFY domain predictions were slightly different between the two genomes. The CaMMV ORF3 contained five domains, a zinc finger-like RNA-binding domain associated with viral coat protein and having the motif CXCX2CX4HX4C (Medberry *et al.*, 1990b) located at nt

coordinates 791–808, the BAR_GRAF2 domain predicted to have an outer membrane protein H (OmpH)-like domain (1052–1162), a pepsin-like aspartate protease domain (1199–1289), a RT-LTR domain (1419–1606), and an RNase H domain (1702 – 1830). Unexpectedly, in CaMMV, the badnavirus movement protein domain, which is the 5'-most domain on the ORF3, and found in most other badnaviruses, was not identified in the CaMMV ORF3. Of the five CaMMV ORF3 predicted domains, all but one was similar to those present in ORF3 of CYVBV. Domains that were shared between the two Trinidad viruses were a zinc finger-like RNA-binding domain located at nt coordinates 810-825, a pepsin-like aspartate protease domain (1196–1284), an RT-LTR domain (1417–1603), and an RNase H domain (1699–1826). The BAR_GRAF2-like domain was absent in CYVBV ORF3, however a trimeric-dUTPase-like domain, with a predicted function in catalyzing hydrolysis of the dUTP-Mg complex into dUMP and pyrophosphate in bacteria and eukaryotes (Marchler-Bauer *et al.*, 2015), was found between nt coordinates 1073–1160.

While the CaMMV ORFY did not contain identifiable predicted domains, the CYVBV ORFY had a helix-turn-helix superfamily (HTH) domain, whose function has been implicated in sequence-specific DNA-binding (Marchler-Bauer *et al.*, 2015). The ORFY, or homologs thereof, of other badnaviruses, in particular, those of *Citrus yellow mosaic virus* (CYMV) ORF6, CSSV ORFY, *Hibiscus bacilliform virus* (HBV) ORF4, *Piper yellow mottle virus* (PYMoV) ORF4, and *Yacon necrotic mottle virus* ORF4, did not contain detectable predicted, conserved domains.

In addition to domains and motifs searched for in the coding regions, certain motifs are conserved among members of the genus, *Badnavirus*, including the intergenic (non-coding) region, were also predicted based on existing genomic sequences. The plant $tRNA^{met}$ primer binding site homolog was identified in the CaMMV genome, as TGGTATCAGAGCTATGTT, and in CYVBV, as TGGTATCAGAGCAAGGTT, both located between nt coordinates 1–18. Also, predicted were a TATA box and poly A signal, at coordinates 7240-7252, as tacTATAAAAgga, and at 7415-7420, as AATAAA, respectively, for CaMMV. Similarly, conserved motifs were identified in the CYVBV genome, at the nt coordinates 7376–7388, as tatTATAAAAtaa, and 7380-7385, as ATAAAA, having the predicted functions as a TATA box and poly A signal, respectively.

Pairwise nucleotide comparisons

To evaluate the species-level status of the CaMMV and CYVBV genomes, sequences from the RT-RNase H region (Table 4.2), or the full-length genome sequence (Table 4.3), were subjected to pairwise nt comparisons with previously reported badnavirus sequences. Results indicated that the Trinidad cacao-associated viruses shared 58-62% nt identity with their counterpart badnaviral relatives, and 62% nt identity with each another. Similarly, the pairwise SDT analysis of the Trinidad viruses and representative, full-length badnavirus genome sequences, indicated that CaMMV and CYVBV shared 58-62% nt identity with their badnavirus counterparts, and 60% nt identity with each other (Table 4.3). Overall, the pairwise nt distance analyses indicated that CaMMV and CYVBV were equally as divergent from each other as they are from all other badnaviruses. Based on the ICTV approved <80% nt identity threshold for badnavirus species demarcation, regardless of whether the RT-RNase H locus or the full-length genome sequence is considered, CaMMV and CYVBV would be considered unique species (Table 4.2, shaded boxes).

Phylogenetic analysis

The genomic relationships between CaMMV and CYVBV and other known badnaviruses analyzed here were examined by re-constructing a ML phylogenetic tree based on the ICTV taxonomically-informative 580 bp RT-RNase H region (King *et al.*, 2012) for 84 badnaviruses and 84 full-length badnavirus genome sequences. Results of the ML analysis indicated that CaMMV and CYVBV from Trinidad grouped with previously reported members assigned to the genus, *Badnavirus*, regardless of whether the partial or complete genome sequence was used to inform the phylogeny (Fig. 4.3). However, the two phylogenetic trees were not equally well-supported with respect to topology or branch locations for certain taxa. Based on phylogenetic tree reconstructed for the RT-RNase H region sequence, three groups were resolved, however, none were statistically supported at a >70% bootstrap confidence level (Fig. 4.3a). Also, on the RT-RNase H tree, CaMMV did not cluster with the CSSV isolates from West Africa, as was predicted by the well-supported, full-length genome tree (3b). In contrast, for CYVBV both the RT-RNase H and genome sequences were grouped with the clade containing the same badnavirus species, respectively, however, none of the groups resolved on the RT-RNase H tree were supported at a 70% confidence level. Based on these results, phylogenetic analysis using

the complete badnaviral genome sequence was far more robust with respect to taxonomic classification, compared to the currently accepted RT-RNase H locus.

Phylogenetic analysis of the full-length badnaviral genomes resolved three well-supported clades 1-3, at 79-100%, based on a >70% bootstrap value (Fig. 4.3b), with clades 1 and 2 being more closely related to each other than to clade 3. The two Trinidad viruses diverged substantially from one another in that CaMMV grouped with nine other badnavirus species in clade 2, which also contained the previously characterized cacao-infecting CSSV species from West Africa, the only other cacao-infecting badnaviruses identified thus far. In contrast, CYVBV was placed in clade 3, together with nine other badnavirus species (Fig. 4.3b) associated with diverse host plant species, none of which are *T. cacao*. Clade 1 contained no cacao-infecting badnavirus species, even though the badnaviruses in this clade are associated with plant families originating from Africa, as well as Asia and the Pacific region. In contrast, the families of host plants of the badnaviruses in clade 2 are native mostly to Asia and Central or South America, while clade 3 badnavirus associated plant families have a predicted origin primarily in Africa and Europe (Hancock & Miller, 2014; Motamayor *et al.*, 2002).

Although based on the full-length genome analysis, it appears that there may be some concordance between host plant origin and extant badnavirus phylogeographic distribution. This scenario may not take into account the movement of plants from their centers of origin for cultivation elsewhere or the potential for host-shifting of endemic viruses. The latter scenario applies to the CSSV complex that followed rather immediately after *T. cacao* was introduced for cultivation into West Africa. Thus, although the two badnaviruses described here infecting cacao in Trinidad, where cacao has been recently introduced (albeit, earlier than occurred in West Africa), they may possibly have evolved initially as viruses of other endemic species, or were co-introduced with a non-endemic host such as banana, root and tuber crops, or sugarcane, among others, making their adaptation to cacao potentially rather recent. It is also possible that they accompanied cacao introductions from South America, however, to our knowledge, badnaviruses have not previously been reported infecting cacao in the New World.

Discussion

Here, two apparently full-length badnavirus genome sequences were determined using the Illumina HiSeq sequencing platform for discovery, by sequencing and assembling the primary genome sequence for each, and by validation using PCR-amplification using specific primers designed based on the Illumina sequence, followed by cloning and capillary Sanger DNA sequencing and primer walking. For PCR amplification an approximately ~ 4 kbp genome fragments were amplified using 250-500 base overlaps between fragments. Here, we report the discovery of two previously, elusive badnaviruses of cacao associated with symptomatic cacao plants in Trinidad, previously known as CTV strains A and B. The CaMMV and CYVBV are herein, proposed as the species names for the (formerly) CTV strains A and B, respectively.

Based on similarities in symptom phenotype described previously in cacao in Trinidad, these two previously undescribed badnaviruses probably represent the pathogens that caused the virus-like symptoms observed in cacao plants in Trinidad during the 1940's. The etiology of the disease manifest in cacao trees, recognizable by foliar "red mottling" and "vein-clearing" symptom phenotypes, has been suspected for many years to be of viral etiology (Kirkpatrick, 1950). However, the identity of the causal agent(s) has been elusive until now, when it has been possible to determine two badnaviral genome sequences from symptomatic cacao plants using next generation Illumina sequencing, a revolutionary 'surveillance' tool that requires no *a priori* knowledge of a suspect virus-like agent. The virus-like pathogens were thought to have been eradicated from Trinidad until fourteen years ago when characteristic strain A and B symptoms were observed in cacao trees in the ICGT.

The identification of these two badnaviruses associated with cacao in Trinidad establishes the hypothesis of their involvement in causality, and now requires that causality be demonstrated. Prior to this report, only one member of the badnavirus genus, CSSV, the causal agent of cacao swollen shoot disease, has been reported to infect *T. cacao* plants, and the virus is thus far restricted to West Africa. Using an infectious clone to a CSSV isolate from Togo (Hagen *et al.*, 1994; Jacquot *et al.*, 1999a), CSSV causality has been previously demonstrated. These results have confirmed those from experimental transmission studies using the mealybug vector or

grafting, for which characteristic disease symptoms also have been reproduced in cacao plants (Box, 1945; Kirkpatrick, 1945).

The arrangement of the CaMMV and CYVBV ORFs is consistent with that found in other badnaviruses, albeit, the number of ORFs range widely from three to seven among the members of the genus (Hagen *et al.*, 1993; Harper *et al.*, 2005), with species identified thus far encoding minimally, ORFs 1-3. Consistent with other badnavirus genomes, CaMMV and CYVBV encode the 'badnavirus core' ORFs 1-3, and a fourth predicted coding region, referred to as ORFY. CSSV, the only other cacao-infecting badnavirus known so far, also encodes a predicted ORFY that overlaps with the 3'-terminus of ORF3, making its position analogous to ORFY in CaMMV and CYVBV. Also, two functionally uncharacterized predicted homologs of CSSV ORFY, referred to as ORF 4 or 6 depending on the virus, have been identified in three other badnavirus species. They are CYMV that infects *Citrus* species (Rutaceae) (Borah *et al.*, 2009) {Accession NC_003382.1}, and HBV, which infects *Hibiscus* spp. (Malvaceae){Accession NC_023485.1}, and PYMoV, a virus of black pepper (Piperaceae) (Hany *et al.*, 2014) {Accession NC_022365.1}. It may be argued that presence of additional ORFs in certain badnaviruses or in others could influence genomic diversity estimates, but because they are embedded within other coding regions making them likewise constrained, the likelihood that they misrepresent divergence estimates or phylogenetic relationships would seem minimal.

The BLASTp and BLASTn searches conducted to characterize ORFY of CYVBV at the nt and aa level, respectively, revealed no evidence of homology with other badnavirus coding regions, however, at the aa level, several shared some homology with certain hypothetical proteins and enzymes. A fifth ORF, ORFX, was found within ORF3 in six of the seven CSSV genomes available in the GenBank database. The presence of an analogous ORFX, however, was neither predicted in either of the viruses from Trinidad, CaMMV and CYVBV, nor did a BLASTp search using the CSSV ORFX produce a match to any other viral ORFs, suggesting that among the known cacao-infecting badnaviruses CSSV is the only one that encodes a predicted ORFX, albeit of unknown function. Thus, the genome structure and arrangement of predicted ORFs for the CaMMV and CYVBV are like those described previously for other members of the genus,

Badnavirus, with the exception of ORFY, which is only occasionally present (Borah *et al.*, 2009; Hany *et al.*, 2014).

Analysis of the intergenic region (IR) of CaMMV and CYVBV predicted a number of potentially biologically interesting functionally conserved motifs. Both viral genomes contained a sequence that is complementary to the $tRNA^{met}$ binding site in plants, at which initiation of viral RNA transcription is known to occur (Medberry *et al.*, 1990b). Also evident in the IR is a polyadenylation signal and a TATA box, both conserved badnavirus motifs that are retained within terminally redundant full-length, badnavirus transcripts (Muller & Sackey, 2005). These badnavirus-specific genomic features further support the identity of these two previously unreported viruses of cacao as badnaviruses. Using the CDD search tool [27] for a comparison of CaMMV and CYVBV ORFs to conserved domains in the analogous ORFs of well-studied badnaviruses, indicated that ORF length and corresponding molecular weight (mass) for each were comparable to previous reports (Table 4.1). Although most of the domains conserved in CaMMV and CYVBV were identifiable in other badnaviruses, the ORF3 domain of CaMMV and CYVBV, and the ORFY domain of CYVBV differed in several respects. Also, in the CaMMV ORF3, a predicted GTPase regulator, annotated as a focal adhesion 2 - Bin/Amphiphysin/Rvs domain (BAR_GRAF2), and containing an outer membrane protein H (OmpH)-like domain (Marchler-Bauer *et al.*, 2015), was found in ORF3 located between the zinc finger and the aspartate protease domains. By comparison, only two other badnaviruses, Banana streak virus and HBV, contain the OmpH-like domain, but not the complete BAR_GRAF2 domain. The best-studied BAR domains function in dimerization, lipid binding, and curvature sensing, however this domain is also found in many other functionally-diverse proteins (Peter et al., 2004).

In *E. coli* OmpH functions as a periplasmic chaperone that interacts with unfolded proteins during translocation from the cytosol across the cytoplasmic membrane (Korndörfer et al., 2004). At least one other bacterial OmpH homolog, annotated as a histone-like protein (HLP-I), is localized in the outer membrane of Salmonella typhimurium, and has been linked to pathogenicity (Hirvas et al., 1990; Walton & Sousa, 2004). Also, some phages and other viruses also encode similar domains, mostly of unknown function. This putative domain is uncommon

among badnaviruses, however, it could is of interest because of the inferred potential role in pathogenicity. Third, the CYVBV ORF3 a predicted, trimeric-dUTPase domain occurs between the zinc finger and aspartate protease domains. Other trimeric dUTPases catalyze the hydrolysis of dUTP-Mg complexes into dUMP and pyrophosphate, and certain viral dUTPases that are instead monomeric, have been shown to mimic the trimeric enzyme (Marchler-Bauer *et al.*, 2015). Many viruses, including some bacteriophages, herpesviruses, poxviruses, retroviruses, encode a dUTPase, even though the host encodes a homologous enzyme. Thus, a dUTPase regulatory function [45] could feasibly apply to badnavirus-host interactions that govern pathogenicity or virulence, including resistance breaking.

Although the function of the predicted trimeric-dUTPase domain in CYVBV is unknown its features suggest a potentially undiscovered role in the badnaviral infection cycle. Indeed, based on a search of the CDD database [27], this domain is present in the genome of PYMoV (Hany et al., 2014), Dioscorea bacilliform virus, and Taro bacilliform virus (TBV). The CYVBV ORF3 size was found to be similar to that of CaMMV, however, it is about 1,000 bases larger than that of CSSV. The larger size could be accounted for by the presence of additional domains found on both of these viral genomes that are likewise absent from CSSV. Another coding region, ORFY (Hagen et al., 1993), is predicted to occur on the CYVBV genome, and it is also present in the seven CSSV genomes from West Africa, and several other available badnaviral genomes that do not infect cacao. A survey of ORFY (referred to as ORF4, or as ORF6) among badnaviruses revealed no identifiable conserved domains, however, CYVBV contains an HTH domain, which has been implicated in DNA binding, a function that is consistent with a predicted role as a transcriptional regulator (Marchler-Bauer *et al.*, 2015).

Pairwise nt comparisons, using the SDT algorithm for taxonomic consideration, indicated that CaMMV and CYVBV shared 57-71% nt identity at the RT-RNase H locus, and 58 – 62% nt identity at the complete genome level with previously characterized badnaviruses. The two genomes shared 62% nt identity for the RT-RNase H locus and 60% at the complete genome level, indicating that CaMMV and CYVBV genomes are about equally divergent with one another and all other badnaviruses known to date. Also, their shared nt identity in the RT-RNase H locus, at the genome level, or at any of the four ORFs, with one another and all other

badnaviruses was lower than the <80% threshold recommended by the ICTV for badnavirus species demarcation (King *et al.*, 2012). Results indicate that CaMMV and CYVBV are two distinct, previously undescribed badnavirus species. Until now CSSV-like genomes reported from West Africa were the only previously known cacao-infecting badnaviruses. Based on the results of the global nt and aa pairwise sequence analyses, the seven available CSSV genome sequences do not share high identity or similarity at the genome level (or with respect to the RT-RNase H locus), respectively, with either of the cacao-infecting viruses from Trinidad. Despite their association with similar symptoms in cacao, these three cacao-infecting badnaviruses, CaMMV and CYVBV from Trinidad, and CSSV from West Africa, are 60% divergent (nt), and are different species.

Based on the phylogenetic analyses (ML) CaMMV and CYVBV genomic sequences grouped with their most closely related, previously described members of the genus, *Badnavirus*. A comparison of the RT-RNase H fragment and full-length genome phylogenies indicated that they clustered in different groups, with CaMMV having its closest relatives in clade 2, and CYVBV in clade 3 (Figs. 3a and 3b). Although the RT-RNase H locus is the approved informative genomic region used for badnavirus taxonomy, the topology of the RT-RNase H tree was not supported, but instead formed a polytomy (70% bootstrap). In contrast, the ML tree for full-length nt genomes resolved three well-supported clades (70% bootstrap), and was incongruent with the RT-RNase H tree (Figs. 3a and 3b), indicating that the RT-RNase region is not taxonomically informative. Consequently, the evolutionary relationships of the badnavirus genus are most accurately reflected at the badnavirus genome level.

Foliar symptoms associated with CaMMV- and CYVBV-infected cacao trees in Trinidad are similar to those observed in CSSV-infected cacao in West Africa, but shoot swelling occurs uniquely in West Africa. Infection by cacao of these viruses in Trinidad is suggestive of several possible scenarios of viral origin. On one hand, CaMMV and CYVBV could have been introduced together with cacao plants from Central and/or South America. However, no other cacao-infecting viruses have been reported from the New World. Another possibility is that ancestral CaMMV and CYVBV were present when or shortly after cacao was introduced to Trinidad, and made a host-shift to cacao from an endemic species, or perhaps from another

exotic cultivated species also introduced into the region. Many agricultural plants cultivated for fiber, food, and medicinal purposes in the Americas and Caribbean Basin are endemic to Africa, Asia-Pacific, and Europe, with which many other badnaviruses have been associated. It is known that CSSV has a number of hosts that are endemic to West Africa (Posnette *et al.*, 1950; Tinsley, 1971c; Todd, 1951), and that it infected cacao shortly after the plant was introduced there during the late 1880's [22]. Endemic CSSV hosts include *Adansonia digitata* L., *Ceiba pentandra* L., *Cola chlamydantha* K.Schum., *Cola gigantean* A. Chev. and *Sterculia tragacantha* Lindl. (Posnette *et al.*, 1950; Tinsley, 1971c; Todd, 1951), not too distant malvaceous relatives. Thus, a similar scenario could have occurred in Trinidad, in that endemic (or previously introduced) badnaviruses host-shifted to cacao, an exotic species, shortly after its introduction there. In terms of potential vector transmission barriers, all three cacao-infecting badnaviruses are transmitted by one or more mealybug vectors that occur in Africa, the Americas, and/or Asia/Pacific (Kirkpatrick, 1950). Some are endemic while others are exotic introductions, and most are relatively or highly polyphagous. Also these mealybug vector species are already known to transmit many other badnaviruses, often between closely- and distantly-related, or unrelated species (Kirkpatrick, 1950; Selvarajan *et al.*, 2016). These features suggest that transmission of badnaviruses by multiple vector species that themselves have a broad, collective host range, provides an effective vehicle for promoting viral host-shifting and thereby rapid diversification in new host niches.

Table 4.1. Analyses of predicted open reading frames identified on the plus-strand of Cacao mild mosaic virus (formerly, strain A) and Cacao yellow vein-banding virus (formerly, strain B) genome, respectively, using the NCBI BLAST algorithm (1) and ORF Finder tools (http://www.ncbi.nlm.nih.gov/gorf/gorf.html).

	ORF1		ORF2		ORF3		ORFY	
Strain	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**
Coordinates (nucleotides)	284 - 715	295 - 816	712 - 1104	817 - 1215	1101 - 6989	1212 - 7088	6632 - 7027	6788 - 7165
Size (nucleotides)	432	522	393	399	5589	5877	396	378
Size (amino acids)	143	173	130	132	1962	1958	131	125
% nucleotide identity	52 - 63%	53 - 67%	53 - 71%	54 - 63%	59 – 63%	58 – 63%	52 – 64%	52 – 63%
BLASTp % amino acid similarity[a]	27 - 57%	23 – 43%	25 – 40%	22 – 34%	35 – 64%	34 – 69%	24 – 45%	[b]NV
Calculated molecular mass (kDa)	16.47	20.32	14.33	14.45	225.92	221.87	14.85	14.15
Functional conserved domain	DUF1319	DUF1319	None	None	Zn, BAR, Pep, RT, RNase H	Zn, Trim, Pep, RT, RNase H	None	HTH

Table 4.1 legend: DUF1319 – domain of unknown function 1319, Zn – zinc knuckle finger, Pep – pepsin-like aspartate protease, RT – reverse transcriptase, RNase H – ribonuclease H, BAR - GTPase regulator associated with focal adhesion 2 (Bin/Amphiphysin/Rvs), Trim – trimeric-dUTPase, HTH – helix-turn-helix. [a]BLASTp results are based on the first 100 viral sequence hits in the GenBank database. NV[b] - No viral sequences determined

Table 4.2. Percentage pairwise nucleotide identity (nt) for the RT-RNase H locus of Cacao mild mosaic virus (formerly, strain A) and Cacao yellow vein-banding virus (formerly, strain B) in relation to other badnavirus type species using MUSCLE alignment implemented in SDTv1.2. The nt identity values are shown for the sequences obtained from the GenBank database, and those shaded share >80% nt identity and represent distinct species.

	PYMaV	TBV	SPBVB	SPBVA	MBV1	DMV	CoYMV	BSCVBV	SCBIMV	SCBMorV	BSUMV	BSUIV	BSULV	CSSV	PYMoV	GVBV	RYNV	HBV	FBV1	CYMV	PVBV	BVF	PBCV	GVCV	BSGFV	BSMYV	DBV	SCBV	KTSV	BSUAV	BSOLV	BSCAV	BSIMV	BSV	BSVAcV	CaMMV	CYVBV
NC_024301_PYMaV	100																																				
NC_004450_TBV	62	100																																			
NC_012728_SPBVB	59	61	100																																		
NC_015655_SPBVA	64	63	82	100																																	
NC_026020_MBV1	62	61	64	63	100																																
NC_008034_DMV	63	61	62	65	62	100																															
NC_001343_CoYMV	65	61	61	61	66	61	100																														
NC_011592_BSCVBV	62	63	66	67	62	66	61	100																													
NC_003031_SCBIMV	60	65	63	64	62	63	65	62	100																												
NC_008017_SCBMorV	58	64	62	64	64	62	62	64	79	100																											
NC_015505_BSUMV	61	65	64	64	63	60	63	67	69	72	100																										
NC_015503_BSUIV	60	64	64	67	61	63	63	64	73	73	76	100																									
NC_015504_BSULV	61	64	66	64	61	60	64	65	72	72	77	79	100																								
NC_001574_CSSV	62	65	64	66	67	61	63	63	60	65	63	63	62	100																							
NC_022365_PYMoV	63	64	66	66	66	62	63	63	64	63	64	64	66	64	100																						
NC_018105_GVBV	63	62	62	65	66	65	62	62	62	59	61	64	61	60	63	100																					
NC_026238_RYNV	62	62	61	60	60	63	61	65	61	60	61	59	59	63	61	71	100																				
NC_023485_HBV	64	62	65	66	67	61	64	65	62	63	64	63	64	67	67	62	61	100																			
NC_017830_FBV1	63	64	65	68	66	66	66	65	60	63	64	66	66	67	68	65	64	70	100																		
NC_003382_CYMV	62	60	64	65	66	62	63	66	64	66	61	62	64	65	66	63	62	68	70	100																	
NC_013262_PVBV	61	62	61	63	63	67	60	66	63	63	63	64	63	63	66	64	65	66	65	69	100																
NC_029303_BVF	63	63	65	68	67	67	63	67	64	65	68	65	63	63	66	67	67	66	68	70	73	100															
NC_014648_PBCV	64	61	65	65	64	67	63	65	61	62	62	64	64	65	64	61	62	65	64	67	64	66	100														
NC_015784.2_GVCV	62	67	63	63	67	63	66	67	62	61	64	63	62	63	65	67	68	65	67	66	64	65	64	100													
NC_007002_BSGFV	61	62	63	63	67	63	65	64	63	63	65	63	66	63	67	63	61	68	67	66	64	64	64	66	100												
NC_006955_BSMYV	62	62	64	66	64	63	62	68	60	61	63	66	65	65	67	62	62	65	66	66	65	66	65	63	65	100											
NC_009010_DBV	64	61	66	66	65	63	63	66	63	64	62	65	65	64	66	63	62	70	67	66	66	64	64	66	63	68	100										
NC_013455_SCBV	62	66	66	66	65	62	64	62	65	64	66	66	64	63	68	65	63	66	66	64	67	69	69	65	67	71	67	100									
NC_004540_KTSV	64	60	61	64	64	61	65	64	60	61	63	61	63	62	65	65	64	63	64	65	65	66	64	67	64	67	66	67	100								
NC_015502_BSUAV	63	64	64	65	64	66	64	65	66	66	64	68	67	62	64	65	65	64	65	65	67	67	67	66	68	70	66	71	70	100							
NC_003381_BSOLV	64	66	64	67	65	67	66	64	63	63	66	67	65	66	67	64	62	67	67	70	64	68	67	66	67	69	66	68	70	74	100						
NC_015506_BSCAV	63	64	67	69	66	65	65	67	65	65	66	68	68	67	66	67	64	63	67	66	67	67	69	66	70	69	66	70	70	79	77	100					
NC_015507_BSIMV	64	62	65	64	64	65	65	63	62	62	66	65	67	63	67	64	62	65	66	65	67	67	66	69	70	67	66	69	69	71	71	75	100				
NC_008018_BSV	63	63	65	66	65	63	63	65	62	60	65	65	65	64	66	62	63	66	66	64	67	68	66	66	68	68	65	72	69	71	71	70	77	100			
NC_007003_BSVAcV	64	63	65	65	66	62	66	65	60	63	66	65	65	65	66	63	65	68	67	64	67	68	67	67	69	71	68	72	72	71	70	73	77	89	100		
KX276640_CaMMV	60	61	65	67	59	64	62	60	62	63	59	62	61	63	65	60	58	62	63	64	62	64	63	61	65	63	64	62	59	64	61	64	63	60	61	100	
KX276641_CYVBV	60	62	63	63	63	65	64	63	64	65	61	63	64	63	63	63	62	65	65	65	67	71	64	63	65	64	63	64	63	67	67	66	63	65	63	62	100

Table 4.2 legend: PYMaV – *Pagoda yellow mosaic associated virus;* TBV – *Taro bacilliform virus*; SPBVA- *Sweetpotato badnavirus A*; SPBVB – *Sweetpotato badnavirus B*; MBV1 – *Mulberry badnavirus 1*; DMV – *Dracaena mottle virus*; CoYMV – *Commelina yellow mottle virus*; BSCVBV – *Bougainvillea spectabilis chlorotic vein-banding virus*; SCBIMV – *Sugarcane bacilliform IM virus*; SCBMorV – *Sugarcane bacilliform Mor virus*; BSUMV – *Banana streak UM virus*; BSUIV – *Banana streak UI virus*; BSULV – *Banana streak UL virus*; CSSV – *Cacao swollen shoot virus*; PYMoV – *Piper yellow mottle virus*; GVBV – *Gooseberry vein banding virus*; RYNV – *Rubus yellow net virus*; HBV – *Hibiscus bacilliform virus*; FBV1 – *Fig badnavirus 1*; CYMV – *Citrus yellow mosaic virus*; PVBV – *Pelargonium vein banding virus*; BVF – *Blackberry virus F*; PBCV – *Pineapple bacilliform comosus virus*; GVCV – *Grapevine vein clearing virus*; BSGFV – *Banana streak GF virus*; BSMYV – *Banana streak MY virus*; DBV –*Dioscorea bacilliform virus*; SCBV – *Sugarcane bacilliform virus*; KTSV – *Kalanchoe top spotting virus*; BSUAV – *Banana streak UA virus*; BSOLV – *Banana streak OL virus*; BSCAV – *Banana streak CA virus*; BSIMV – *Banana streak IM virus*; BSVAcV – *Banana streak Vietnam Acuminata virus*, CaMMV – Cacao mild mosaic virus; CYVBV – Cacao yellow vein-banding virus

Table 4.3. Percentage pairwise nucleotide identity (nt) for Cacao mild mosaic virus (formerly, strain A) and Cacao yellow vein-banding virus (formerly, strain B) complete genome sequences in relation to badnavirus sequences using MUSCLE alignment implemented in SDTv1.2. The nt identity values are shown for the sequences obtained from the GenBank database.

	PYMaV	GVCV	GVBV	RYNV	TBV	DMV	PVBV	CoYMV	SCBMorV	SCBIMV	BSUMV	BSULV	BSUIV	PYMoV	FBV1	SPBVB	SPBVA	MBV1	CYMV	HBV	CSSV	PBCV	DBV	KTSV	BSMYV	SCBV	BSGFV	BSCVBV	BSIMV	BSV	BSVAcV	BSUAV	BSOLV	BSCAV	SCBV	BVF	CaMMV	CYVBV
NC_024301_PYMaV	100																																				
NC_015784_GVCV	60	100																																			
NC_018105_GVBV	60	63	100																																		
NC_026238_RYNV	60	63	66	100																																	
NC_004450_TBV	59	60	60	59	100																																
NC_008034_DMV	60	60	60	59	60	100																															
NC_013262_PVBV	61	61	62	61	61	63	100																														
NC_001343_CoYMV	61	59	59	58	61	60	60	100																													
NC_008017_SCBMorV	58	59	59	58	61	60	60	60	100																												
NC_003031_SCBIMV	59	59	60	59	60	59	59	61	75	100																											
NC_015505_BSUMV	60	59	59	58	60	59	59	60	65	64	100																										
NC_015504_BSULV	59	60	58	59	61	59	59	61	66	64	72	100																									
NC_015503_BSUIV	59	60	60	59	60	59	59	61	66	65	73	75	100																								
NC_022365_PYMoV	61	59	60	59	62	60	61	60	61	62	62	62	61	100																							
NC_017830_FBV1	60	62	61	60	60	59	61	62	61	60	60	61	60	63	100																						
NC_012728_SPBVB	58	58	59	59	60	59	60	59	59	60	59	59	60	61	62	100																					
NC_015655_SPBVA	59	59	59	59	59	59	60	59	59	59	60	60	61	61	62	81	100																				
NC_026020_MBV1	60	61	61	60	60	60	62	61	60	61	59	61	60	62	63	60	61	100																			
NC_003382_CYMV	60	59	59	59	60	60	62	60	61	60	60	61	60	62	64	61	60	64	100																		
NC_023485_HBV	60	60	60	59	61	60	62	61	60	60	60	60	60	63	64	61	60	63	64	100																	
NC_001574_CSSV	60	60	58	60	60	59	61	61	60	59	60	60	60	62	63	62	61	62	64	64	100																
NC_014648_PBCV	60	59	60	59	59	60	59	60	60	60	61	61	60	59	61	59	60	60	60	60	61	100															
NC_009010_DBV	61	60	60	59	61	60	61	61	61	61	61	61	61	62	63	61	61	62	63	63	62	60	100														
NC_004540_KTSV	59	59	60	59	59	61	60	61	59	60	60	60	61	60	60	59	59	60	60	59	59	62	60	100													
NC_006955_BSMYV	60	60	60	59	60	60	60	62	60	60	61	62	61	62	63	59	60	62	62	61	60	63	61	63	100												
NC_013455_SCBV	61	59	60	60	60	60	61	62	62	62	63	63	63	62	62	61	61	62	62	62	61	65	62	63	66	100											
NC_007002_BSGFV	59	60	60	59	60	59	60	62	62	61	61	61	61	60	61	59	60	61	60	61	60	62	61	64	64	64	100										
NC_011592_BSCVBV	60	60	60	60	62	61	61	61	61	60	60	61	60	61	61	59	60	61	61	61	61	59	61	60	61	61	60	100									
NC_015507_BSIMV	60	61	59	59	60	60	60	62	61	61	62	61	62	61	61	61	59	62	61	60	61	63	61	64	63	65	64	60	100								
NC_008018_BSV	60	60	60	59	60	60	61	62	61	60	62	62	62	62	62	61	60	61	61	61	62	62	62	64	65	65	65	60	75	100							
NC_007003_BSVAcV	60	60	60	59	60	60	60	62	61	61	61	61	62	61	62	60	60	61	62	61	61	63	61	65	64	65	66	61	75	88	100						
NC_015502_BSUAV	60	59	61	60	60	60	61	62	61	62	63	63	63	61	62	60	60	62	62	61	61	62	62	65	63	65	65	61	68	67	68	100					
NC_003381_BSOLV	59	60	59	60	61	60	60	61	61	61	62	61	63	62	61	60	61	61	61	62	60	62	61	65	64	64	64	60	66	67	67	70	100				
NC_015506_BSCAV	58	60	59	59	60	61	61	62	61	62	62	63	63	61	61	60	61	61	62	62	60	63	62	64	63	66	64	60	67	67	67	72	77	100			
FJ824814_SCBV	59	60	60	59	60	60	60	61	62	62	63	63	63	61	62	60	61	62	62	61	61	61	62	65	63	65	64	60	67	67	67	71	79	79	100		
KJ413252_BVF	60	60	62	61	60	62	64	61	61	60	61	60	61	60	61	60	60	62	60	61	60	61	61	61	61	61	60	61	61	61	61	62	61	60	62	100	
KX276640_CaMMV	60	61	60	60	59	60	60	61	60	59	60	60	60	61	61	60	61	60	61	62	61	60	61	60	60	61	61	60	60	61	60	61	61	61	62	60	100	
KX276641_CYVBV	59	60	59	59	59	61	61	60	59	60	60	60	59	60	61	60	59	60	60	61	60	59	61	60	60	61	59	60	60	60	60	61	61	61	60	62	60	100

Table 4.3 legend: PYMaV – *Pagoda yellow mosaic associated virus;* TBV – *Taro bacilliform virus*; SPBVA- *Sweetpotato badnavirus A*; SPBVB – *Sweetpotato badnavirus B*; MBV1 – *Mulberry badnavirus 1*; DMV – *Dracaena mottle virus*; CoYMV – *Commelina yellow mottle virus*; BSCVBV – *Bougainvillea spectabilis chlorotic vein-banding virus*; SCBIMV – *Sugarcane bacilliform IM virus*; SCBMorV – *Sugarcane bacilliform Mor virus*; BSUMV – *Banana streak UM virus*; BSUIV – *Banana streak UI virus*; BSULV – *Banana streak UL virus*; CSSV – *Cacao swollen shoot virus*; PYMoV – *Piper yellow mottle virus*; GVBV – *Gooseberry vein banding virus*; RYNV – *Rubus yellow net virus*; HBV – *Hibiscus bacilliform virus*; FBV1 – *Fig badnavirus 1*; CYMV – *Citrus yellow mosaic virus*; PVBV – *Pelargonium vein banding virus*; BVF – *Blackberry virus F*; PBCV – *Pineapple bacilliform comosus virus*; GVCV – *Grapevine vein clearing virus*; BSGFV – *Banana streak GF virus*; BSMYV – *Banana streak MY virus*; DBV –*Dioscorea bacilliform virus*; SCBV – *Sugarcane bacilliform virus*; KTSV – *Kalanchoe top spotting virus*; BSUAV – *Banana streak UA virus*; BSOLV – *Banana streak OL virus*; BSCAV – *Banana streak CA virus*; BSIMV – *Banana streak IM virus*; BSVAcV – *Banana streak Vietnam Acuminata virus*, CaMMV – Cacao mild mosaic virus; CYVBV – Cacao yellow vein-banding virus

Fig. 4.1. Characteristic symptoms of the formerly, Cacao Trinidad virus (CTV) strain A and B isolates, on younger and mature cacao leaves. The Cacao mild mosaic virus (CaMMV) genome was isolated from the indicator cacao genotype SCA 6, exhibiting symptoms characteristic of the formerly, CTV strain A, whereas, the Cacao yellow vein-banding virus was isolated from cacao genotype ICS 27, which exhibited a more severe symptom phenotype, compared to CaMMV in SCA 6.

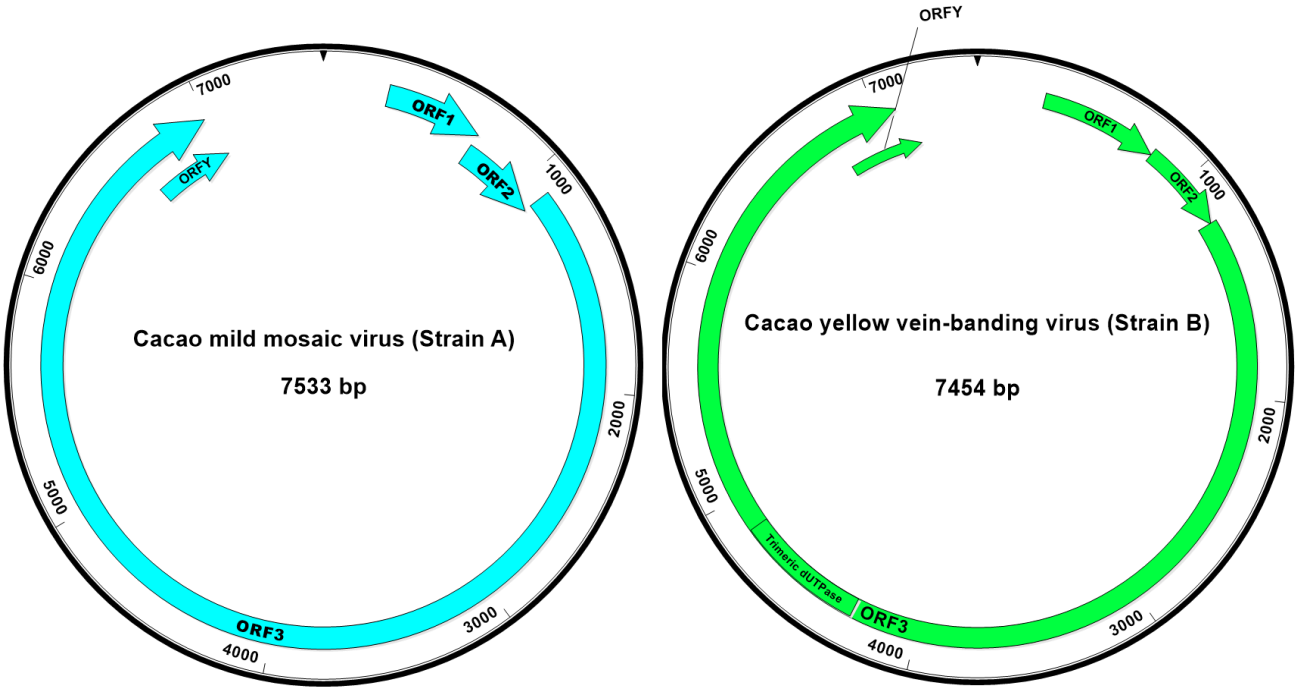

Fig. 4.2. The genome map of the newly identified Cacao mild mosaic virus (CaMMV) and Cacao yellow vein-banding virus (CYVBV) species, respectively. Putative coding regions for open reading frames (ORFs) 1, 2, 3 and Y, predicted using the NCBI ORF Finder tool, are indicated by filled arrows. The CaMMV and CYVBV genomes are 7,533 and 7,454 base pairs in length, respectively. The black arrow head indicates the coordinate of the first nucleotide, and of the 5'-end of the predicted $tRNA^{met}$ binding site, required for priming reverse transcription of the negative strand.

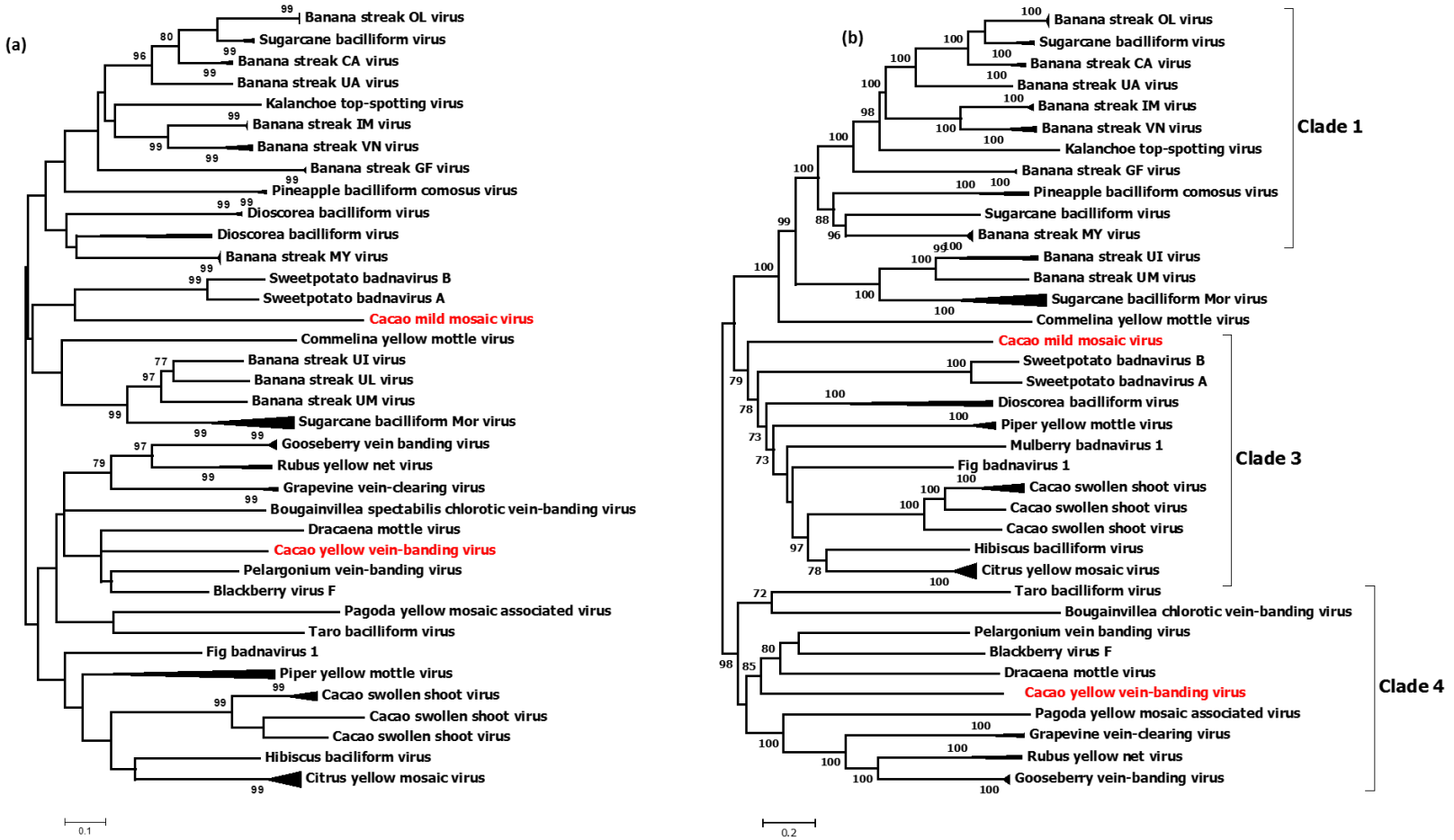

Fig. 4.3. Maximum likelihood phylogenetic tree of 35 badnavirus species genomic nucleotide sequences, including Cacao mild mosaic virus (strain A) and Cacao yellow vein-banding virus (strain B) using the 580 bp RT-RNase H region (a) and the complete badnaviral genomes (b). Analyses were done using MEGA 6 software and only bootstrap values that were greater than 70% are shown. The horizontal branch lengths are proportional to the genetic distance, and numbers shown at branch points indicate bootstrap values for 1000 replicates. Each node was collapsed into a single taxon when more than one sequence was available for each species

REFERENCES

Ackonor, J. B. (1995). Preliminary studies on breeding and predation on Scymnus (Pullus) sp. and Hyperaspis egregia Mader on Planococcoides njalensis (Laing). In *Proc 1st Cocoa Pests Dis Semin*, pp. 238–241. Accra, Ghana.

Adegbola, M. O. K. (1975). The reaction of four Nigerian isolates of CSSV on the germination and growth of cacao seedlings under different environmental conditions. In *Proc 5th Int Cocoa Res Conf*, pp. 338–343. Ibadan, Nigeria.

Adomako, D. & Owusu, G. K. (1974). Studies on the mechanical transmission of cocoa swollen shoot virus : Some factors affecting virus multiplication and symptom development in cocoa. *Ghana J Agric Sci* **7**, 7–15.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**, 403 – 410.

Alverson, W. S., Whitlock, B. A., Nyffeler, R., Bayer, C. & Baum, D. A. (1999). Phylogeny of the core Malvales: Evidence from ndhF sequence data. *Am J Bot* **86**, 1474–1486.

Ameyaw, G. A., Domfeh, O., Dzahini-Obiatey, H. K., Ollennu, L. A. A. & Owusu, G. K. (2016). Appraisal of Cocoa swollen shoot virus (CSSV) mild isolates for cross protection of cocoa against severe strains in Ghana. *Plant Dis* **100**, 810–815.

Ameyaw, G. A. & Ollennu, L. A. (2006). Control of cocoa swollen shoot disease by eradicating infected trees in Ghana : A survey of treated and replanted areas **25**, 647–652.

Ameyaw, G. A., Wetten, A., Dzahini-Obiatey, H., Domfeh, O. & Allainguillaume, J. (2013). Investigation on Cacao swollen shoot virus (CSSV) pollen transmission through cross-pollination. *Plant Pathol* **62**, 421–427.

Ameyaw, G. a., Dzahini-Obiatey, H. K. & Domfeh, O. (2014). Perspectives on cocoa swollen shoot virus disease (CSSVD) management in Ghana. *Crop Prot* **65**, 64–70. Elsevier Ltd.

Argout, X., Salse, J., Aury, J., Guiltinan, M. J. & Droc, G. et al. **(2011).** The genome of Theobroma cacao. *Nat Genet* **43**, 101 –108.

Attafuah, A., Blencowe, J. W. & Brunt, A. A. (1963). Swollen shoot disease of cocoa in the

Sierra Leone. *Trop Agric Trinidad* **40**, 229–232.

Baker, R. E. D. & Dale, W. T. (1947a). Notes on a virus disease of cacao. *Ann Appl Biol* **34**, 60–65.

Baker, R. E. D. & Dale, W. T. (1947b). Virus diseases of cacao in Trinidad II. *Trop Agric* **24**, 127–130.

Bateman, A., Coggill, P. & Finn, R. D. (2010). DUFs: Families in search of function. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **66**, 1148–1152. International Union of Crystallography.

Bhattacharjee, R. & Kumar, P. L. (2013). Technical crops: Cacao. In *Genome Mapp Mol Breed plants*, pp. 1689–1699. Edited by S. Schreck & C. Kole. Heidelberg, Germany: Springer.

Bigger, M. (1972a). Recent work on the mealybug vectors of Cocoa swollen shoot disease in Ghana. *Trop Pest Manag* **18**, 61–70.

Bigger, M. (1972b). Recent work on the mealybug vectors of Cocoa swollen shoot disease in Ghana. *Trop Pest Manag* **18**, 61–70.

Bolger, A. M., Lohse, M. & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120.

Borah, B. K., Johnson, A. M. A., Sai Gopal, D. V. R. & Dasgupta, I. (2009). Sequencing and computational analysis of complete genome sequences of Citrus yellow mosaic badna virus from acid lime and pummelo. *Virus Genes* **39**, 137–140.

Bowers, J. H., Bailey, B. A., Hebbar, P. K., Sanogo, S. & Lumsden, R. D. (2001). The impact of plant diseases the impact of plant diseases on world chocolate production. *Plant Heal Prog* 1–15.

Box, H. E. (1945). Insect transmission of the ‘Swollen-Shoot’ virus in West African cacao. *Nature* **155**, 68–609.

Briddon, R. W. (2003). Cotton leaf curl disease, a multicomponent begomovirus complex. *Mol Plant Pathol* **4**, 427–434.

Brown, J. K., Zerbini, F. M., Navas-Castillo, J., Moriones, E., Ramos-Sobrinho, R., Silva, J. C. F., Fiallo-Olivé, E., Briddon, R. W., Hernández-Zepeda, C. & other authors. **(2015).** Revision of Begomovirus taxonomy based on pairwise sequence comparisons. *Arch Virol* **160**, 1593–1619.

Brunt, A. A. & Kenten, R. H. (1963). The use of protein in the extraction virus Cacao swollen shoot virus from cacao leaves. *Virology* **19**, 388–392.

Brunt, A. A., Kenten, R. H., Gibbs, A. J. & Nixon, H. L. (1965). Further studies on Cocoa yellow mosaic virus. *J Gen Microbiol* **38**, 81–90.

Brunt, A. A. (1964). Some properties of Cacao swollen shoot virus. *J Gen Microbiol* **36**, 303–309.

Cilas, C., Muller, E. & Mississo, E. (2005). Occurrence of Cacao swollen shoot virus in Litimé, the Main Cocoa-Producing Area of Togo. *Plant Dis* **89**, 913.

Clough, Y., Faust, H. & Tscharntke, T. (2009). Cacao boom and bust: sustainability of agroforests and opportunities for biodiversity conservation. *Conserv Lett* **2**, 197–205.

Conesa, A. & Gotz, S. (2008). Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int J Plant Genomics* **2008**.

Cornwell, P. B. (1960). Movements of the vectors of virus diseases of cacao in Ghana. II.—Wind movements and aerial dispersal. *Bull Entomol Res* **51**, 175–201.

Cornwell, P. B. (1958). Movements of the vectors of virus diseases of cacao in Ghana. I.—Canopy movements in and between trees. *Bull Entomol Res* **49**, 613–630.

Dean, F. B., Nelson, J. R., Giesler, T. L. & Lasken, R. S. (2001). Rapid amplification of plasmid and phage DNA using Phi29 DNA polymerase and multiply primed rolling circle amplification. *Genome Res* **11**, 1095–1099.

Demesure, B., Sodzi, N. & Petit, R. J. (1995). A set of universal primers for amplification of polymorphic non-coding regions of mitochondrial and chloroplast DNA in plants. *Mol Ecol* **4**, 129–131.

Ding, S. W., Mackenzie, A., Torronen, M. & Gibbs, A. (1990). Nucleotide sequence of the

virion protein gene of cacao yellow mosaic tymovirus. *Nucleic Acids Res* **18**, 5886.

Dixon, L. K. & Hohn, T. (1984). Initiation of translation of the cauliflower mosaic virus genome from a polycistronic mRNA: evidence from deletion mutagenesis. *Embo J* **3**, 2731–2736.

Domfeh, O., Dzahini-Obiatey, H., Ameyaw, G. A., Abaka-Ewusie, K. & Opoku, G. (2011). Cocoa swollen shoot virus disease situation in Ghana : A review of current trends. *African* **6**, 5033–5039.

Dongo, L. N. & Orisajo, S. B. (2007). Status of cocoa swollen shoot virus disease in Nigeria. *African J Biotechnol* **6 (17)**, 2054–2061.

Doyle, J. J. & Doyle, J. L. (1990). Isolation of plant DNA from fresh tissue. *Focus (Madison)* **12**, 13 – 15.

Dunn, B. M. (2010). Overview of aspartic acid proteases. In *Aspartic acid proteases as Ther targets*. Edited by A. K. Ghosh. WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

Dzahini-Obiatey, H. & Fox, R. T. V. **(2010).** Early signs of infection in Cacao swollen shoot virus (CSSV) inoculated cocoa seeds and the discovery of the cotyledons of the resultant plants as rich sources of CSSV. *African J Biotechnol* **9**, 593–603.

Dzahini-Obiatey, H., Ameyaw, G. A. & Ollennu, L. a. (2006). Control of cocoa swollen shoot disease by eradicating infected trees in Ghana: A survey of treated and replanted areas. *Crop Prot* **25**, 647–652.

Dzahini-Obiatey, H., Domfeh, O. & Amoah, F. M. (2010). Over seventy years of a viral diseaseof cocoa in Ghana: From researcher's perspective. *African J Agric Res* **5**, 476–485.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–1797.

Entwistle, P. F. & Longworth, J. F. (1963). The relationships between cacao viruses and their vectors : the feeding behaviowr of three mealybug (Hornoptera : Pseudococcidae) species. *Ann Appl Biol* **52**, 387–391.

Entwistle, P. F., Johnson, C. G. & Dunn, E. (1959). New pests of cocoa (Theobroma cacao L.)

in Ghana following applications of insecticides.pdf. *Nature* **184**, 2040.

Frison, E. A., Diekman, M. & Nowell, D. (1999). Cacao. In *FAO/IPGRI Tech Guidel safe Mov germplasm*, pp. 1–32. Edited by E. A. Frison, M. Diekman & D. Nowell.

Geering, A. D., McMichael, L. A., Dietzgen, R. G. & Thomas, J. E. (2000). Genetic diversity among Banana streak virus isolates from Australia. *Phytopathology* **90**, 921–927.

Geering, A. D. W. (2014). Caulimoviridae (Plant pararetroviruses). *eLS*.

Ghosh, P., Ghosh, K., Simsekt, M. & Rajbhandaryt, U. L. (1982). Nucleotide sequence of wheat germ cytoplasmic initiator methionine transfer ribonucleic acid. *Nucleic Acids Res* **10**, 3241–3247.

Gray, A. (2001). *The world cocoa market outlook*. London, UK: LMC International Ltd.

Guest, D. (2007). Black pod: Diverse pathogens with a global impact on cocoa yield. *Phytopathology* **97**, 1650–1653.

Hagen, L. S., Lot, H., Godon, C., Tepfer, M. & Jacquemond, M. (1994). Infection of Theobroma cacao using cloned DNA of Cacao swollen shoot virus and particle bombardment. *Phytopathology* **84**, 1239–1243.

Hagen, L. S., Jacquemond, M., Lepingle, A., Lot, H. & Tepfer, M. (1993). Nucleotide sequence and genomic organization of Cacao swollen shoot virus. *Virology* **196**, 619–628.

Hancock, J. F. & Miller, A. J. (2014). Crop plants: Evolution. In: eLS. In *John Wiley Sons Ltd, Chichester*. Chichester: John Wiley & Sons Ltd.

Hanna, A. D. & Heatherington, W. (1957). Arrest of the swollen-shoot virus disease of cacao in the gold coast by controlling the mealybug vectors with the systemic insecticide, dimefox. *Ann Appl Biol* **45**, 473–480.

Hany, U., Adams, I. P., Glover, R., Bhat, A. I. & Boonham, N. (2014). The complete genome sequence of Piper yellow mottle virus (PYMoV). *Arch Virol* **159**, 385–388.

Harper, G., Hart, D., Moult, S., Hull, R., Geering, A. & Thomas, J. (2005). The diversity of Banana streak virus isolates in Uganda. *Arch Virol* **150**, 2407–2420.

Hoffmann, K., Sackey, S. T., Sackey, E., Maiss, D., Adomako, D. & Vetten, H. J. (1997). Immunocapture polymerase chain reaction for the detection and characterization of Cacao swollen shoot sirus 1A isolates. *J Phytopathol* **145**, 205–212.

Hohn, T., Fütterer, J. & Hull, R. (1997). The proteins and functions of plant pararetroviruses : Knowns and unknowns. *CRC Crit Rev Plant Sci* **16**, 133–161.

Hughes, J. d'A. & Ollennu, L. A. A. (1994). Mild strain protection of cocoa in Ghana against cocoa swollen shoot virus — a review. *Plant Pathol* **43**, 442–457.

Hull, R. & Covey, S. N. (1983). Does cauliflower mosaic virus replicate by reverse transcription? *Trends Biochem Sci* **8**, 119–121.

Jacquot, E., Hagen, L. S., Michler, P., Rohfritsch, O., Stussi-Garaud, C., Keller, M., Jacquemond, M. & Yot, P. (1999a). In situ localization of Cacao swollen shoot virus in agroinfected Theobroma cacao. *Arch Virol* **144**, 259–71.

Jacquot, E., Hagen, L. S., Michler, P., Rohfritsch, O., Stussi-Garaud, C., Keller, M., Jacquemond, M. & Yot, P. (1999b). In situ localization of cacao swollen shoot virus in agroinfected Theobroma cacao. *Arch Virol* **144**, 259–271.

Jacquot, E., Hagen, L. S., Jacquemond, M. & Yot, P. (1996). The open reading frame 2 product of Cacao swollen shoot badnavirus is a nucleic acid-binding protein. *Virology* **225**, 191–195.

Jeger, M. J. & Thresh, J. M. (1993). Modeling reinfection of replanted cocoa by swollen shoot virus in pandemically diseased areas. *J Appl Ecol* **30**, 187–196.

Johns, R. & Gibberd, A. V. (1951). Review of cocoa industry in Nigeria. In *6th Cocoa Conf*, p. 137. London, UK: Cocoa Chocolate Confectionary Alliance, Ltd.

Kebe, B. I., Kouakou, K., N'Guessan, N. F., Assiri, A. A., Adiko, A., Aké, S. & Anno, A. P. (2006). Le swollen shoot en Côte d'Ivoire: situation ac- tuelle et perspectives. In *Proc 15th Int Cocoa Res Conf San José, Costa Rica*, pp. 907–922.

Kenten, R. H. (1972). The purification and some properties of cocoa necrosis virus, a serotype of tomato black ring virus. *Ann Appl Biol* **71**, 119–126.

Kenten, R. H. & Legg, J. T. (1970). Methods for assessing the tolerance and resistance of different types of cocoa to cocoa swollen-shoot virus. *Ann Appl Biol* **65**, 419–424.

Kenten, R. H. & Legg, J. T. (1971). Serological relationships of some viruses from cocoa (Theobroma cacao L.) in Ghana. *Ann Appl Biol* **67**, 195–200.

King, A. M. Q., Adams, M. J., Carstens, E. B. & Lefkowitz, E. J. (2012). *Badnavirus In: Virus taxonomy. Ninth report of the international committee on taxonomy of viruses.* London: Elsevier Academic Press.

Kirkpatrick, T. W. (1945). *Insect pests of cacao and insct vectors of cacao virus disease. In: A Report of Cacao Research, 1945 - 1951*. St. Augustine, Trinidad.

Kirkpatrick, T. W. (1950). Insect transmission of cacao virus disease in Trinidad. *Bull Entomol Res* **41**, 99.

Kouakou, K., Kébé, B. I., Kouassi, N., Aké, S., Cilas, C. & Muller, E. (2012). Geographical distribution of Cacao swollen shoot virus molecular variability in Côte d ' Ivoire. *Plant Dis* **96**, 1445–1450.

Lartey, L. L. (2013). Mapping cocoa swollen shoot virus disease distribution in western region, Ghana 78.

Liu, P. S. W. & Liew, P. S. C. (1975). Transmission studies of a cocoa virus disease (yellow veinbanding) in Sabah. In *Tech Bull Dep Agric*, pp. 11–17. Sabah, Malaysia.

Lockhart, B. E. L. & Olszewski, N. E. (1993). Serological and heterogeneity of Banana streak badnavirus: Implications for virus detection in Musa germplasm. In *Breed Banan plantain Resist to Dis pests*, pp. 105–113. Edited by J. Ganry. Montepellier, France: CIRAD, INIBAP.

Lockhart, B. E. L. (1990). Evidence for a double-stranded circular DNA genome in a second group of plant viruses. *Phytopathology* **80**, 127–131.

Lot, H., Djiekpor, E. & Jacquemond, M. (1991). Characterization of the genome of cacao swollen shoot virus. *J Gen Virol* **72**, 1735–1739.

Mangenot, G., Alibert, H. & Basset, A. (1946). Sur les caracteres du swollen shoot en Cote

d'Ivoire. *Rev Int Bot Appl Paris* **283**, 13.

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., Geer, R. C., He, J., Gwadz, M. & other authors. **(2015).** CDD: NCBI's conserved domain database. *Nucleic Acids Res* **43**, D222–D226.

Medberry, S. L., Lockhart, B. E. L. & Olszewski, N. L. (1990a). Properties of Commelina yellow mottle virus's complete DNA sequence, genomic discontinuities and transcript suggest that it is a pararetrovirus. *Nucleic Acids Res* **18**, 5505–5513.

Medberry, S. L., Lockhart, B. E. L. & Oiszewski, N. E. (1990b). Properties of Commelina yellow mottle virus's complete DNA sequence, genomic discontinuities and transcript suggest that it is a pararetrovirus. *Nucleic Acids Res* **18**, 5505–5513.

Motamayor, J. C., Risterucci, A. M., Lopez, P. A., Ortiz, C. F., Moreno, A. & Lanaud, C. (2002). Cacao domestication I: the origin of the cacao cultivated by the Mayas. *Heredity (Edinb)* **89**, 380–386.

Motamayor, J. C., Mockaitis, K., Schmutz, J., Haiminen, N., Iii, D. L., Cornejo, O., Findley, S. D., Zheng, P., Utro, F. & other authors. **(2013).** The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. *Genome Biol* **14**, r53. BioMed Central Ltd.

Muhire, B. M., Varsani, A. & Martin, D. P. (2014). SDT: A virus classification tool based on pairwise sequence alignment and identity calculation. *PLoS One* **9**, e108277.

Muller, E. & Sackey, S. (2005). Molecular variability analysis of five new complete cacao swollen shoot virus genomic sequences. *Arch Virol* **150**, 53–66.

Muller, E. (2008). Cacao swollen shoot virus. In *Encycl Virol*, 3rd edn., pp. 403–409. Edited by B. W. J. Mahy & M. H. V. Van Regenmortel. France: Elsevier Ltd, Oxford.

Muller, E., Jacquot, E. & Yot, P. (2001). Early detection of cacao swollen shoot virus using the polymerase chain reaction **93**, 15–22.

Ollennu, A. (2001). Synthesis : case history of cocoa viruses. In *Plant Virol Sub-Saharan Africa*, pp. 33–49. Edited by J. d'A. Hughes & B. O. Odu. Ibadan, Nigeria: International Institute of

Tropical Agriculture.

Ollennu, L. & Owusu, G. K. (2003). Field evaluation of the protective capacity of CSSV mild strain N1 against severe strain New Juaben (1A) isolate. *Ghana J Agric Sci* **36**, 3 – 12.

Ollennu, L. A. A. & Owusu, G. K. (1989). Isolation and study of mild strains of cocoa swollen shoot virus for possible cross protection. In *Proc 4th Int Plant Virus Epidemiol Work*, pp. 119 – 122. Montpellier, France.

Ollennu, L. A. A. & Owusu, G. K. (2002). Spread of cocoa swollen shoot virus to cacao (Theobroma cacao L.) plantings in Ghana. *Trop Agric Trinidad* **79**, 224–230.

Oro, F., Mississo, E., Okassa, M., Guilhaumon, C., Fenouillet, C., Cilas, C. & Muller, E. (2012a). Geographical differentiation of the molecular diversity of cacao swollen shoot virus in Togo. *Arch Virol* **157**, 509–514.

Oro, F., Mississo, E., Okassa, M., Guilhaumon, C., Fenouillet, C., Cilas, C. & Muller, E. (2012b). Geographical differentiation of the molecular diversity of cacao swollen shoot virus in Togo. *Arch Virol* **157**, 509–14.

Owusu, G. K. (1971). Cocoa necrosis virus in Ghana. *Trop Agric Trinidad* **48**, 133–139.

Padi, F. K., Domfeh, O., Takrama, J. & Opoku, S. (2013). An evaluation of gains in breeding for resistance to the cocoa swollen shoot virus disease in Ghana. *Crop Prot* **51**, 24–31. Elsevier Ltd.

Partiot, M., Amefia, Y. K., Djiekpor, E. K. & Bakar, K. A. (1978). Le swollen shoot du cacaoyer au Togo. Investaire preliminaire et premiere estimation des pertes causees par la maladie. *Cafe-cacao Thea* **22**, 217–228.

Ploetz, R. C. (2006). Cacao diseases : Important threats to chocolate production worldwide. *Phytopathology* **97**, 1634–1639.

Posnette, A. F. (1950). Virus dieases of cacao in West Africa. VII. Virus transmission by different vector species. *Ann Appl Biol* **37**, 378–384.

Posnette, A. F. & Strickland, A. H. (1948). Virus diseases of cacao in West Africa . III. Technique of insect transmission. *Ann Appl Biol* **35**, 53–63.

Posnette, A. F. & Todd, J. M. (1955). Virus diseases of cacao in West Africa. IX. Strain variation and interference in virus 1A. *Ann Appl Biol* **43**, 433 – 453.

Posnette, A. F. (1940). Transmission of swollen shoot. *Trop Agric Trinidad* **17**, 98.

Posnette, A. F. (1944). Viruses of cacao in Trinidad. *Trop Agric* **21**, 105 – 106.

Posnette, A. F. (1947). Virus diseases of cacao in West Africa. I. Cacao viruses 1A, 1B, 1C and 1D. *Ann Appl Biol* **34**, 388–402.

Posnette, A. F., Robertson, N. F. & Todd, J. M. (1950). Virus diseases of cacao in West Africa. V. Alternative host plants. *Ann Appl Biol* **37**, 229–240.

Quainoo, A. K., Wetten, A. C. & Allainguillaume, J. (2008a). Transmission of cocoa swollen shoot virus by seeds **150**, 45–49.

Quainoo, A. K., Wetten, A. C. & Allainguillaume, J. (2008b). The effectiveness of somatic embryogenesis in eliminating the cocoa swollen shoot virus from infected cocoa trees **149**, 91–96.

Rector, A., Tachezy, R. & Van Ranst, M. (2004). A sequence-independent strategy for detection and cloning of circular DNA virus genomes by using multiply primed rolling-circle amplification. *J Virol* **78**, 4993–8.

Richard, A. & Ræbild, A. (2016). Tree diversity and canopy cover in cocoa systems in Ghana. *New For* **47**, 287–302. Springer Netherlands.

Sackey, S. T. & Hull, R. (1994). The use of dot blot hybridisation methods to detect cocoa swollen shoot virus isolates. In *Rep Cocoa Res Inst Ghana 1991/1992*, p. 126.

Sagemann, W., Paul, H. L., Adomako, D. & Owusu, G. K. (1983). The use of Enzyme-Linked Immunosorbent Assay (ELISA) for detection of Cacao swollen shoot sirus (CSSV) in Theobroma cacao. *J Phytopathol* **106**, 281–284.

Sagemann, W., Lesemann, D.-E., Paul, H. L., Adomako, D. & Owusu, G. K. (1985). Detection and some comparison of Ghanaian isolates of Cacao swollen shoot virus (CSSV) by Enzyme-linked immunosorbent assay (ELISA) and immuno-electron microscopy (IEM) using an antiserum to CSSV strain 1A. *J Phytopathol* **114**, 79–89.

Selvarajan, R., Balasubramanian, V. & Padmanaban, B. (2016). Mealybugs as vectors. In *Mealybugs their Manag Agric Hortic Crop*, pp. 123–130. Edited by M. Mani & C. Shivaraju. Tiruchirappalli: Springer.

Smulders, M. J. M., Esselink, D., Amores, F., Ramos, G., Sukha, D. A. & Butler, D. R. (2008). Identification of cocoa (Theobroma cacao L .) varieties with different quality attributes and parentage analysis of their beans. *IGENIC Newsletters* **12**, 1–13.

Sreenivasan, T. N. (2009). *The enigma of the ICS 76 plants at Reading, UK*. Reading, UK.

Steven, W. F. (1936). A new disease of cocoa in the Gold Coast. *Trop Agric Trinidad* **14**, 84.

Stevens, H., Rector, A. & Van Ranst, M. (2010). Multiply primed rolling-circle amplification method for the amplification of circular DNA viruses. *Cold Spring Harb Protoc* **2010**, pdb.prot5415.

Swarbrick, J. T. (1961). Cacao virus in Trinidad. *Trop Agric* **38**, 245 – 249.

Syller, J. (2012). Facilitative and antagonistic interactions between plant viruses in mixed infections. *Mol Plant Pathol* **13**, 204–216.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* **30**, 2725–2729.

Thorold, C. A. (1975). Diseases of cacao. *Clarendon Press Oxford* 77–78.

Thresh, J. M. (1958). The spread of virus disease in cacao. *West African Cocoa Res Inst Tech Bull* **5**, 36.

Thresh, J. M. & Tinsley, T. W. (1959). *The viruses of cocoa*. *West African Cocoa Res Inst Tech Bull*. Crown Agents Overseas Govt. Admin.

Tinsley, T. W. (1971a). The ecology of cacao viruses . I . The role of wild hosts in the incidence of swollen shoot virus in West Africa. *J Appl Ecol* **8**, 491–495.

Tinsley, T. W. (1971b). The ecology of cacao viruses. I. The role of wild hosts in the incidence of swollen shoot virus in West Africa. *Jo* **8**, 491–495.

Tinsley, T. W. (1971c). The ecology of cacao viruses. I. The role of wild hosts in the incidence

of swollen shoot virus in West Africa. *J Appl Ecol* **8**, 491–495.

Todd, J. M. (1951). An indigenous source of swollen shoot disease of cacao. *Nature* **167**, 952–953.

Tormo-Mas, M. A., Donderis, J., GArcia-Caballer, M., Alt, A., Mir-Sanchis, I., Marina, A. & Penades, J. R. (2013). Phage dUTPases control transfer of virulence genes by a proto-oncogenic G protein-like mechanism. *Mol Cell* **49**, 947–958.

Uhde, C., Vetten, H. J., Maiss, E., Adomako, D. & Paul, H. L. (1993). Studies on particle components of Cacao swollen shoot virus. *J Phytopathol* **139**, 207–216.

Villesen, P. (2007). FaBox: An online toolbox for FASTA sequences. *Mol Ecol Notes* **7**, 965–968.

Wessel-Riemens, P. C. (1965). The use of proteins Nigerian and isolates alkaloids of cocoa in mechanical transmission Cacao swollen shoot virus. *Virology* **27**, 566–570.

Yang, I. C., Hafner, G. J., Revill, P. a., Dale, J. L. & Harding, R. M. (2003). Sequence diversity of South Pacific isolates of Taro bacilliform virus and the development of a PCR-based diagnostic test. *Arch Virol* **148**, 1957–1968.

Yang, Z. N. & Mirkov, T. E. (1997). Sequence and relationships of Sugarcane mosaic and Sorghum mosaic virus strains and development of RT-PCR-based RFLPs for strain discrimination. *Phytopathology* **87**, 932–939.

Zhang, Y., Angel, C. A., Valdes, S., Qiu, W. & Schoelz, J. E. (2015). Characterization of the promoter of Grapevine vein clearing virus. *J Gen Virol* **96**, 165–169.

FUTURE DIRECTIONS

Eight primers were designed around the CSSV genome, and all were capable of amplifying CSSV sequences. However, because of the high genomic variability, they selectively amplified certain CSSV isolates. The proposed research involves alignment of the sequences determined from PCR amplification with each primer from the current study. The alignment will be used to assess regions on the genome where "universal" degenerate primers can be designed, so that the majority of CSSV isolates can be identified using one set of primers. The sequences determined here are expected to provide insight into variation of the amplified regions so that degeneracy can be incorporated as necessary. The use of one primer will be more cost-effective than using eight primers for each sample.

Additionally, species-specific primers will be designed based on the four CSSV species identified. Each species-specific primer will confirm the presence of CSSV sequences, and it will also allow identification of the specific species that is amplified. The "universal" and the species-specific primers will be valuable for rapid and reliable CSSV diagnostics in West Africa and they will assist in making informed decisions, especially during the eradication campaign where infected trees are cut down and replaced by tolerant varieties.

The number of CSSV full-length genomes in this study were only a small percentage compared to the total number of samples sequenced. There is still need for extensive study of CSSV genomic variability through sequencing of complete genomes. Young, symptomatic leaves will be collected for Illumina sequencing because they are likely to have higher viral titer than the mature leaves. Also, more locations will be sampled, preferably from all the cacao-growing regions, to allow for the study of genomic diversity based on geographic location. The availability of full-length genomes will aid in diagnostic primer development and in epidemiological studies of CSSV in West Africa.

In addition, the full-length genomes will allow for the construction of full-length infectious CSSV clones from various isolates or species. Moreover, virus inoculation procedures can be improved using the infectious clones. The current efforts to breed cacao varieties that are resistant to CSSV will most likely benefit from the full-length CSSV clones, instead of relying on mealybugs or grafting for CSSV transmission, which do not allow quantification of the inoculum.

Because the functions of some proteins encoded by the CSSV ORFs are still unknown, the full-length clones will be useful for studies on the biology and transmission of CSSV. Alternatively, the sequences of full-length clones will be used for secondary structure analysis, which allows for prediction of functions of CSSV proteins, and for comparison to known isolates or species. This information will contribute to the management of the cacao swollen shoot disease.

This study indicated that six species of non-cacao plants were infected with CSSV. To manage the disease efficiently, the full host range of CSSV needs to be determined. To this end, extensive screening of plants in and around the cacao fields, especially those used as barriers or for shade, will need to be conducted. This may be achieved by use of the universal or species-specific primers, or the full-length genome sequencing. Knowing the full host range of CSSV will determine what crops to use as barriers so as to hinder the spread of the virus.

Some of the clones sequenced in this study, and also some isolates that are previously published, showed evidence of recombination between unknown isolates. Recombination analysis will be conducted on the recently determined sequences to evaluate the presence of recombinant isolates. The results are expected to elucidate the basis of high genomic variability among the sequenced CSSV isolates, and also to assess the impact of recombination on virus species or isolate virulence.

Lastly, the cacao genome should be screened to determine if there are CSSV integrated into the genome. The genomes of at least five other badnaviruses have been found to be integrated into their plant host genome. The outcome of such an experiment will influence how the diagnostic tests will be developed because the integrated viral genome must be distinguished from the non-integrated genome. Integrated viral genomes may excise out of the host genome and cause infection, so their identification is important.