STUDIES ON THE PATHOGENESIS OF *CACAO SWOLLEN SHOOT VIRUS* TO ENHANCE *THEOBROMA CACAO* L. QUARANTINE PROCEDURES

Megan Alice Richardson

A thesis submitted in partial fulfilment of the requirements of the University of the West of England, Bristol for the degree of Doctor of Philosophy (PhD)

This research programme was carried out in collaboration with Cocoa Research UK and UWE Bristol

Faculty of Applied Sciences,
UWE Bristol

April 2024

Word count: 37,474

ABSTRACT

Theobroma cacao (cocoa) is an economically important crop and the source of one of the world's most consumed products, chocolate; an industry valued at USD 103 billion per annum. Pests and diseases pose a major threat to cocoa production, affecting yield and resulting in a loss of up to 38% of the annual global harvest. To further the understanding of these diseases, as well as developing strategies for detection and prevention, are imperative in protecting the livelihoods of small-hold cocoa farmers worldwide.

Cocoa Swollen Shoot Disease (CSSD) has been described as one of the most important diseases affecting West African cocoa production. The causative agent is a Badnavirus named Cacao Swollen Shoot Virus (CSSV). CSSV is vectored by mealybug species in a semi-persistent manner, although there are still major gaps in the understanding of modes of transmission including location of viral particles *in planta*. Symptoms of the disease on *T. cacao* are only identifiable three years post-infection; this long latency period makes it difficult for farmers to control spread and increases processing times in quarantine centres, where symptoms must be observed to confirm viral presence.

This project aimed to improve on knowledge of CSSV pathogenicity and explore new methods for detection which can be easily implemented in quarantine centres such as the International Cocoa Quarantine Centre (ICQC, Reading), which serve to stop the spread of cocoa diseases. A method of CSSV detection using samples of *T. cacao* stems has been developed. CSSV DNA can be extracted from stem exudate and detected by qPCR and end-point PCR. This method has the potential to improve quarantine processes for detection of low levels of CSSV DNA from budwood samples. During this project, novel badnaviral sequences were discovered in *T. cacao*. A High-Resolution Melt assay was developed which successfully

differentiates between four different badnaviral insert types in *T. cacao.* The mendelian inheritance of these integrated sequences was determined using this method to examine samples from multiple crosses of *T. cacao* plants containing different insert types. This assay can be implemented in quarantine centres as a quick and straightforward method of detection, and differentiation between CSSV positive and negative samples. Finally, fluorescent *in situ* microscopy was used in order to identify the location of CSSV *in planta*. Using a novel CSSV antibody and an indirect immunofluorescent staining strategy, CSSV was successfully detected in the spongy mesophyll tissues of *T. cacao* leaves. This finding provides valuable insight into the mealybug-based transmission of this virus, suggesting that the rate of transmission is linked to the extended periods of time the vectors are known to probe these cells during the feeding process.

ACKNOWLEDGEMENTS

It is going to be difficult to show my appreciation to all of the people who have supported me over the last ~four years, but I'll do my best. Firstly, I'd like to thank my supervisor Dr Andy Wetten for all of his support, and my funders, Cocoa Research UK and UWE Bristol. Thank you to all of my PGR friends; without your support I don't think I would have been able to get through this. A special thanks to Viky and Savannah; I'm so grateful to have met you both and to have become such good friends during this process. Thank you for your love and support, you've been my personal cheerleaders; I'm glad we all managed to wade through in the end. Thank you to my colleagues at Forest Research; Liv and the entire AOD team, especially Sandra who has been incredibly supportive during my transition into work.

The support of my family, my Mum (Sue) and sisters Bethan and Siân who have been especially supportive, never once doubting my ability to finish this thesis[*I think]. I'd also like to make a special mention to my partner Luke's family; Mandi, David and Katie, with whom I spent the majority of my time with during the COVID-19 pandemic. John and Nik, who have spent many an evening with me, drinking around a fire, listening to me enthuse (or not) about my PhD work. Of course, I have to thank my wonderful partner, Luke, for the huge amount of love and support throughout the last few years. Without you and the Mao I would probably (definitely) have given up a long time ago.

Finally, I want to thank my neighbour Benita and her late husband, Brian. Through you both I was exposed to so many wonderful areas of Cornwall, Devon and beyond. Without those experiences I would not be where I am today. I am forever grateful for the time I had to spend with you both as a child and the time I get to

speak to Benita about my research, and other interesting things I find out and about. I hope to continue to do you proud.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CSSV	Cacao Swollen Shoot Virus
CSSD	Cocoa Swollen Shoot Disease
ICQC	International Cocoa Quarantine Centre, Reading, UK
qPCR	Quantitative PCR
EP-PCR	End-point PCR
WBD	Witches' Broom Disease
ds	Double stranded
BSV	Banana streak virus
ORF	Open reading frame
COSHH	Control of substances hazardous to health
ICG	International Cocoa Genebank
UWE	University of the West of England
MGW	Molecular grade water
DNA	Dioxyribonucleic acid
PCR	Polymerase chain reaction
WT	Wild type
EVE	Endogenous viral element
ss	Single stranded
HGT	Horizontal gene transfer
BEV	Banana endogenous virus
TEPRV	Tobacco endogenous pararetrovirus
PVCV	Petunia vein clearing virus
HRM	High Resolution Melt
bp	Base pair
A	Adenine
T	Thymine
C	Cytosine
G	Guanine
eTcBV	Endogenous *T. cacao* bacilliform virus
DBTRV	Dioscorea bacilliform TR virus
CMV	Cucumber mosaic virus
TYLCV	Tomato yellow leaf curl virus
CP	Coat protein

RTVB	Rice Tungro bacilliform virus
PLRV	Potato leaf roll virus
CGMMV	Cucumber green mottle mosaic virus
Cy3	Cyanine
FITC	Fluorescein isothiocyanate
Ab	Antibody
BSA	Bovine Serum Albumin
CaMV	Cauliflower mosaic virus

CHAPTER ONE

Introduction

1.1 *Theobroma cacao* L.

Theobroma cacao L. (family *Malvaceae*), know colloquially as 'cocoa' or 'cocoa', is an economically important crop and the source of one of the worlds most consumed products, chocolate. *T. cacao* was first cultivated by the Maya in South America, who referred to the fruit and resulting chocolate as the "Food of the Gods"; the origin from which its Latin name was derived. While chocolate was often consumed by the Maya as a drink, cocoa beans were also historically used as currency in Mesoamerica (Montagna *et al*., 2019).

While *T. cacao* originated in South America, distribution now reaches across the equator to the Caribbean, West Africa and Indonesia. Agroforestry is often used in cocoa cultivation, whereby forests are selectively thinned allowing for cocoa and other trees can be planted beneath the remaining canopy (Franzen and Mulder, 2007). Cocoa plants favour high temperatures (minimum average of 18-21 °C, maximum average of 30-32 °C) and relatively high humidities (>70%; Montagna *et al*., 2019). Wild relatives of cultivated *T. cacao* are native to the upper amazon tributaries, and display the greatest genetic diversity (Zarrillo *et al*., 2018); conferring variability in plant size, yields, and disease resistances (dos Santos Dias *et al*., 2003). The lack of genetic diversity and close proximity of plants when grown commercially can put plants at risk of attack from pests and disease more readily.

Three genetic groups of cocoa have been defined: "Criollo" and "Forastero" were defined based on morphological traits and geographical origins. The third group named "Trinitario" consists of "Criollo" x "Forastero" hybrids (Motamayor *et al.*, 2008; Cornejo *et al*., 2018). A more detailed analysis of the genetic diversity of *T. cacao* was performed in 2008 by Motamayor *et al*. using microsatellite markers. This analysis characterised cocoa germplasm into ten major genetic groups, known as Amelonado, Contamana, Criollo, Curaray, Guianna, Iquitos, Marañon, Nacional, Nanay and Purus (Figure 1), with Criollo the most likely representative of Mesoamerican cocoa domesticated ~3,600 years ago (Cornejo *et al*., 2018).

Figure 1: Distribution of ten major genetic groups of *T. cacao* in South America: sourced from Cornejo *et al*., 2018.

1.2.1 Cocoa production in West Africa

The cocoa industry is valued at USD 103 billion per annum (Marelli *et al*., 2019). West Africa is one of the largest producers of cocoa in the world. It is forecasted that in the year 2018-2019, 76.3% of the world's cocoa will be produced in West Africa (ICCO Quarterly Bulletin of Cocoa Statistics, 2019). Six million hectares of West African forest is planted with cocoa, much of which belongs to small holder farmers who traditionally plant their cocoa randomly in sparse areas of shaded forest. The largest West African producers include Ghana and Côte d'Ivoire, followed by Nigeria and Cameroon (Wessel and Quist-Wessel, 2015). One of the major causes of economic losses in cocoa is the action of pests and diseases.

1.3 Diseases of *T. cacao*

T. cacao is under constant threat, affected by pests and diseases wherever it is grown around the world. With over 80% of cocoa grown by millions of small-hold farmers who rely on it as a source of income (Perrine-Walker, 2020), cocoa research focusses heavily on the prevention and treatment of pests and diseases, which are responsible for losses of up to 37% of the annual global cocoa harvest (Marelli *et al*., 2019).

1.3.1 Phytophthora pod rot

Phytophthora pod rot, also known as Black pod rot, is a fungal disease of *T. cacao* which can be caused by a complex of seven *Phytophthora* species: the two primary causative agents being *Phytophthora palmivora* and *P. megakarya*. It is accountable for up to 30 % of pod losses and 10 % of tree deaths in wetter cocoa growing areas of Central and West Africa (Perrine-Walker, 2020). Disease symptoms include pod rot, stem cankers, leaf and seedling blight, flower cushion infection and root infection (End *et al*., 2021). Control measures involve the removal of infected pods, reducing shade in plantations to lower humidity and the use of fungicides. Phytophthora is a soil-borne pathogen, which coupled with the fact that it can be spread through rain splash, insects and rodents makes this disease particularly difficult to control (Marelli *et al.*, 2019).

1.3.2 Witches' Broom Disease

Witches' Broom Disease (WBD) is a disease of *T. cacao* originating in the Amazon Basin, South America. The causal agent of WBD is a basidiomycete in the genus *Moniliophthora;* specifically, *M. perniciosa* (Marelli *et al*., 2019). The name 'Witches' broom' is derived from the ability of this fungal pathogen to infect all actively growing tissues, which can result in vegetative brooms which become necrotic within several weeks of infection. Young pods are also susceptible to

distortion, early ripening and necrotic lesions when infection is present. WBD rapidly spread throughout South America from its origin in the Amazon Basin over the course of 30 years to surrounding cocoa producing regions, such as Brazil, Colombia and Ecuador (End *et al.*, 2021).

1.3.3 *T. cacao* quarantine

Crop breeding programmes aiming to access an increased genetic diversity pool or research groups striving to test new plant varieties in the field are often in receipt of non-native germplasm. However, these potentially useful activities bring with them a large unintended risk involved in the international exchange of plants worldwide. The associated, unintentional introduction of pests and diseases into new areas can have a severe detrimental effect on native species. Almost all countries have put into place plant quarantine measures in an attempt to inhibit the spread of pests and diseases and regulate the movement of genetically modified organisms with the aim of protecting the future of plant heath (MacLeod *et al.*, 2010; Daymond, 2018; Dubey *et al*., 2021).

Disruptions to seasonal patterns, inconsistent rainfall, rising temperatures and drought attributed to a changing climate have been identified as contributing factors in enhancing the spread of cocoa diseases such as CSSD and Phytophthora pod rot (Anning *et al*., 2022), making the role of cocoa quarantine units such as the ICQC that bit more important.

1.4 Badnaviruses and badnavirus/plant interactions.

Badnaviruses have been characterised as having non-enveloped bacilliform particles 120-150 x 30 nm in size, containing a circular double-stranded (ds) DNA genome (Lockhart, 1990). Badnaviruses infect a broad range of plants of economic importance all over the world and include CSSV, *Banana streak virus* (BSV), *Grapevine vein clearing virus* and *Piper yellow mottle virus* (Bhat *et al*., 2016). Following host cell penetration Badnaviruses replicate through transcription/reverse transcription. The dsDNA is released from the capsid and repaired by host repair polymerase and ligase and is then transferred into the nucleus where it forms mini chromosomes; the template for transcription of mRNA, which when transported into the cytoplasm will be synthesised into new genomic molecules or viral proteins (Chiumenti *et al*., 2016). Some Badnaviruses, such as BSV (LaFleur *et al*., 1996), have been known to integrate into the plant genome without sign of infection. Some BSV associated Badnaviruses are capable of generating infection in response to abiotic stresses, whereby de novo infective particles are created, which release viral genome into plant cells and cause infection. Others are incapable of producing infections due to insufficient amounts of integrated viral DNA (Bhat *et al*., 2016). Badnavirus disease mechanisms can be attributed to the production of RNA silencing molecules, which interfere with the host plant's post-transcriptional products, resulting in their inactivation or degradation (Chiumenti *et al*., 2016).

1.4.2 *Cacao Swollen Shoot Virus* (CSSV)

CSSV belongs to the family Caulimoviridae, genus Badnavirus (Hagen *et al*., 1993; Büchen-Osmond, 2004). As with other members of the badnavirus genus, CSSV virions consist of a non-enveloped, elongated, icosahedral capsid enclosing the double stranded DNA viral genome (Figure 2). Virions have a length of 130 nm and a width of 28 nm (Büchen-Osmond, 2004). The CSSV genome is about 7.5 kb long

and consists of 5 open reading frames (ORF; Figure 3) (Abrokwah *et al*., 2016). ORF I encodes a 17-kDa protein with an unknown function. ORF II encodes a 14-kDa nucleic acid-binding protein. ORF III is the largest and encodes a 211-kDa polyprotein which is post-transcriptionally cleaved into smaller functional units by the viral aspartic protease, which is also encoded by ORFIII. Other proteins produced by this ORF include those required for virus movement, assembly and replication. ORF X and OFR Y, 13-kDa and 14-kDa respectively, are of unknown function (Hagen *et al*., 1993). ORF X is diverse among different CSSV strains, differing considerably in both size and sequence (Bhat *et al*., 2016). Localisation of CSSV virions within host plant and vector are still under investigation. Theories based on feeding patterns of the CSSV vector, the mealybug tend to suggest that CSSV is localised in the vascular system, specifically the phloem tissues (Jacquot *et al.,* 1999).

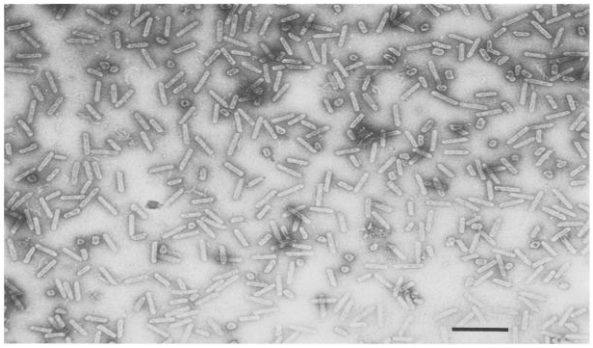


Figure 2: TEM image of purified CSSV virions. Scale bar represents 200nm. Image sourced from Lot *et al*., 1991.

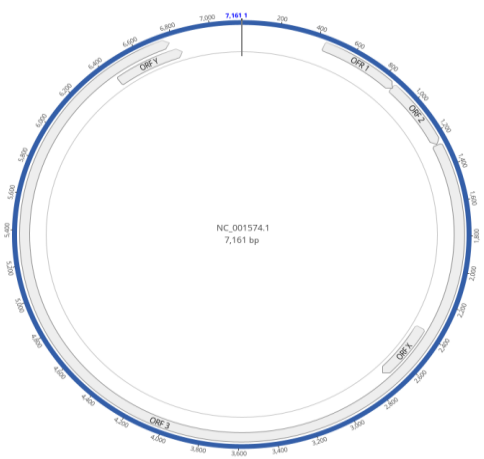


Figure 3: Full, 7kb CSSV genome with ORFs annotated. Accession number: NC_001574, Hagen *et al*., 1993.

1.5 Cocoa Swollen Shoot Disease

Cacao Swollen Shoot Disease (CSSD) was discovered in *T. cacao* in major cocoa producing regions of South East Ghana in 1936 (Steven, 1936). Posnette (1940) identified that the causative agent of this disease is a virus, which was subsequently named Cacao Swollen Shoot Virus (CSSV). CSSD is endemic to West Africa and has not yet been described in any other cocoa producing countries (Muller *et al*., 2018). CSSD has been described as one of the most important diseases affecting West Africa cocoa production (Muller *et al*., 2018), with reports of the virus causing substantial crop losses in these cocoa producing West African countries (Ameyaw *et al*., 2016). Average annual loss in Ghana between 1946 and 1974 was estimated to be over £3,650,000 and as of 2008, in the Eastern Region of Ghana the estimated annual yield losses attributed to CSSV ranged from 20,000 tonnes to 120,000 tonnes of cocoa (Muller, 2009). Both severe and mild virus strains exist, with the most virulent causing a severe reduction in the yield obtained from cocoa trees and eventually resulting in tree death (Posnette, 1947). Susceptible varieties of cocoa, such as Amelonado, can be killed by severe CSSV strains, such as New Juaben, within 2-3 years (Muller, 2009). Despite the severity and impact CSSD has on the livelihood of cocoa producing farmers in West Africa, there are still major gaps in the understanding of its mealybug-based transmission and the means by which the pathogen is harming the plant. Understanding CSSV pathogenicity is essential; this information has the potential to help establish new methods of CSSD control.

In 1946 a CSSV eradication campaign was launched in Ghana. Farmers were encouraged to remove all cocoa trees showing CSSD symptoms and any adjacent trees; through this campaign, more than 200 million cocoa trees have been destroyed. This CSSV eradication campaign was a failure due to disruption, discontinuation and deficient eradication procedures and is thought to be the costliest campaign of its kind. It has been suggested that the only efficient way to reduce losses caused by CSSV is to increase varietal resistance. However, the ongoing efforts to replace susceptible cocoa varieties with less-susceptible varieties are met with resistance from farmers due to the failure of the first eradication

campaign (Andres *et al*., 2017). The identification and/or development of more resistant varieties of cocoa will be aided through understanding and clarification of CSSV pathogenesis.

1.5.1 Symptoms

The principal host of CSSV is *T. cacao*. However, reservoirs of the virus have been discovered in plants native to West Africa, including species in the families of *Sterculiaceae*, *Malvaceae*, *Tiliaceae* and *Bombaceae* (Muller, 2009). CSSV symptoms can be seen on the leaves, stem, roots and pods of *T. cacao*. In younger leaves on particularly susceptible varieties of cocoa such as Amelonado, red-vein banding can occur (Figure 4a) and will disappear as the leaves mature. In mature leaves a number of symptoms can be seen: yellow clearing along veins; small and large flecks; blotches or streaks. A common symptom is chlorotic vein banding (Figure 4b). Other symptoms include swelling of the stem (Figure 4c) and roots, which results from abnormal cell proliferation (Muller, 2009). Distortion of pod shape and colour can also occur (Domfeh *et al*., 2011).

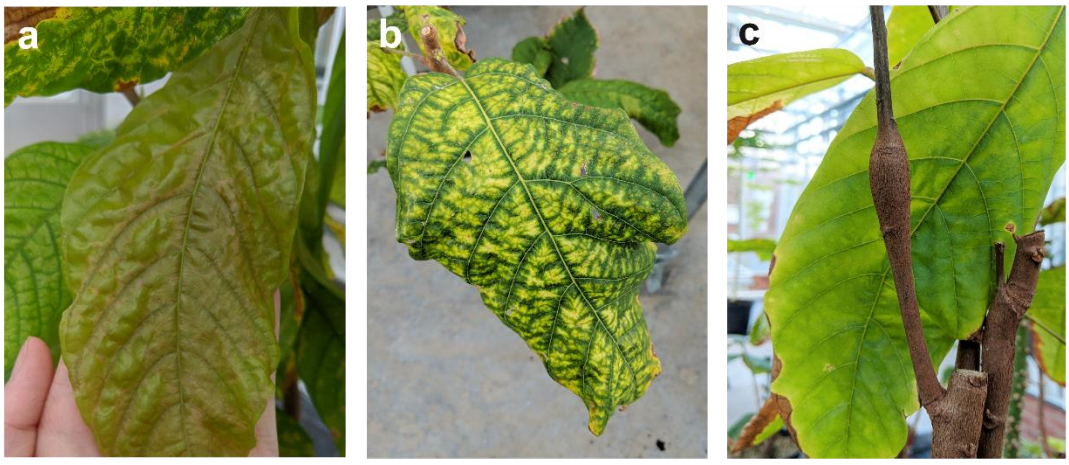


Figure 4: Typical symptoms associated with CSSD on CSSV-positive *T. cacao* plants. a) red vein banding on immature leaf; b) chlorosis on mature leaf; c) swollen shoot.

1.5.2 CSSV vector – the mealybug

Different species of mealybug, such as *Formicoccus njalensis* and *Planococcus citri*, are responsible for the transmission of CSSV (Strickland, 1951). The nymphs of the first, second and third larval stages and the adult females are capable of virus transmission (Figure 5). The younger, more mobile nymphs make more efficient transmitters than the sedentary adult females. The short-lived adult males do not transmit CSSV due to their lack of mouthparts. After feeding on infected cocoa tissues the mealybugs will acquire the virus and later transmit the virus when they come into contact with healthy hosts (Muller, 2009). Transmission of CSSV is thought to be semi-persistent (Ameyaw *et al*., 2014); virus acquisition can take from several minutes to several hours and associate with their vector on the cuticle lining the mealybug's mouthparts (Blanc *et al*., 2014). In mealybugs, CSSV acquisition has been found to take between 12 and 36 hours, with transmission to healthy plants taking as little as one hour feeding on the plant (Ameyaw *et al*., 2014). As a semi-persistent virus, vector mouthparts are thought to be the site of CSSV retention. However, the retention site has not been confirmed.

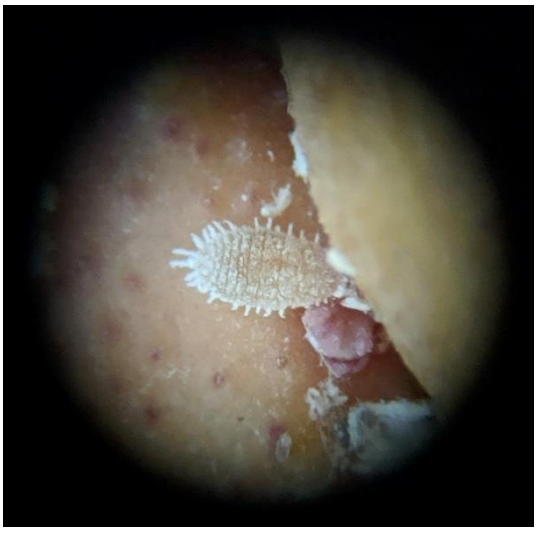


Figure 5: Adult *Planococcus citri* female (mealybug).

1.6 Aims and objectives

The aim of this project is to provide insights into this economically important virus' pathogenicity and develop novel methods for CSSV detection in the *T. cacao* quarantine processes through the following objectives:

- Development of a rapid and efficient method for the detection of CSSV DNA to aid in *T. cacao* quarantine processes.
    - Given CSSV is described as a vascular-limited virus, will disrupting the vasculature and incubation of budwood samples in water allow diffusion of CSSV DNA into exudate?
- Identification of novel Badnaviral sequences integrated into the genome of *T. cacao*.
    - Development of a HRM method for detection of different insert types.
    - Can these sequences be inherited vertically in a Mendelian fashion?
- Identify the localisation of CSSV virions *in planta*.
    - Using a CSSV specific antibody, an indirect staining approach and immunofluorescent confocal laser scanning microscopy.

CHAPTER TWO

General Materials and Methods

2.1 Introduction

All experimental work was carried out in accordance with conditions and procedures used working in a hazard level 2 laboratory with Department of Environment Food and Rural Affairs accreditation. Health and safety training, personal laboratory induction and Control of Substances Hazardous to Heath (COSHH) Risk Assessment evaluation was received and/or prepared and assessed prior to experimental work starting. Suitable personal protective equipment was used when required in accordance with COSHH Risk Assessment forms. All reagents were purchased from Fisher Scientific (Loughborough, UK) unless stated otherwise.

2.2.1 Source of materials

CSSV-infected *Theobroma cacao* Amelonado plants (G plants, Table 1) were used as a primary source of CSSV inoculum. These plants were grown from seed in a sand mix, kept in a glasshouse at and >25 °C in 80% RH and were watered for 10 minutes twice a day through an automated irrigation system. Supplementary lighting automatically activates during low-light conditions in day-time. Plants were inoculated with CSSV at seedling stage via mechanical transmission of CSSV particles through mealybug feeding and translocation. At the start of this project, these plants were ~9 years old with a high CSSV viral load. All CSSV-positive samples were selected from G plants showing symptoms of CSSV infection on either young and/or mature leaves or stem. CSSV-free *T. cacao* Amelonado and other *T. cacao* strains were provided by the International Cocoa Quarantine Centre, Reading (ICQC, R, UK; Figure 6) and the International Cocoa GeneBank, Trinidad and Tobago (ICG, T; Trinidad and Tobago; Table 1).

Figure 6: The International Cocoa Quarantine Unit at Reading, UK. Sourced from: https://research.reading.ac.uk/cocoa/international-cocoa-quarantine-centre-reading/

Table 1: List of *T. cacao* used in experimental work.

Code	Strain	Donor Collection	Country of Origin	Accession number
G#	CSSV 'New Juaben'-Infected Amelonado	ICQC, R; infected with CSSV when seedlings at Reading University, brought to UWE by Dr Andy Wetten	-	
RB1	CSSV-Free Amelonado	ICQC, R	-	
IMC60	eTcBV+ Iquitos	ICQC, R	Peru	RUQ 1617
EET183		ICG, T	Ecuador	
ICS40	eTcBV+ Admixed	ICQC, R	Trinidad	RUQ 1432
CCN51		ICQC, R	Ecuador	

ICQC, R: International Cocoa Quarantine Centre, Reading UK
ICG, T: International Cocoa GeneBank, Trinidad and Tobago
eTcBV+: Endogenous *T. cacao* Bacilliform Virus positive.
CSSV 'New Juaben'-Infected: infected with CSSV strain 'New Juaben'
CSSV-Free: free of CSSV and eTcBV.

2.2.2 *T. cacao* maintenance

All *T. cacao* used in this project was kept in the UWE Envirotron (Figure 7) at >25 °C in 80% RH and were watered for 10 minutes twice a day through an automated irrigation system. Supplementary lighting automatically activates during low-light conditions in day-time. Collections of CSSV-infected, CSSV-free and Integrated Sequence Positive (IS+) *T. cacao* were kept in separate compartments to avoid accidental contamination through mealybug feeding. On occasion, leaf disks were sampled from ICQC and transported to UWE where they were stored at -20 °C before processing.

Figure 7: The Envirotron at UWE Bristol, UK.

2.3 Molecular methods

Molecular techniques such as DNA extraction, end-point PCR (EP-PCR) and quantitative PCR (qPCR) were used to identify presence of CSSV in *T. cacao* tissues, quantify CSSV concentration in different samples and distinguish between different integrated sequences. All procedures described were used as standard unless stated otherwise. All molecular methods were performed using aseptic technique, with reagents kept in ice at all times to reduce risk of degradation, and all equipment sterilised in a UV hood for 15 minutes before and after experimental set-up. All primers and synthetic oligos were sourced from Eurofins Genomics (Eurofins Genomics Europe Shared Services GmbH, Germany) unless stated otherwise.

2.3.1 Extraction of CSSV DNA from *T. cacao* leaf discs

Samples for DNA extraction were taken using sterile Eppendorf tubes. The Eppendorf lids were used to cut a 1 cm diameter disc from the outer perimeter of *T. cacao* leaves, ensuring that a section of the secondary vein was present. Leaf discs were frozen in 1.5 mL Eppendorf tubes at -20 °C prior to extraction. The QIAGEN DNeasy Plant Mini/Pro Kits (QIAGEN, UK) were used to perform the DNA extractions. Care was taken to ensure CSSV infected samples and non-infected samples did not come into contact, to avoid accidental cross-contamination. Leaf tissue samples were homogenized in a QIAGEN TissueLyser LT according to the manufacturer's instructions. Samples were disrupted in 2-3 cycles, with each cycle lasting for 2 min at 30 Hz. Cycles were repeated until a finely macerated sample was achieved.

2.3.2 Endpoint PCR for amplification of CSSV DNA using *T. cacao* DNA extractions or CSSV extractions from *T. cacao* stems

DNA extractions from *T. cacao* tissues underwent end-point PCR using CSSV specific primers (Table 2) for the detection of CSSV DNA. A 2 µM primer working stock was made by adding 10 µL of forward primer and 10 µL of reverse primer to 480 µL of Hyclone Molecular Biology Grade Water (MGW) (Fisher Scientific, UK). A mastermix was made for a 25 µL reaction; this included 12.5 µL of BioMix (Meridian Bioscience; Scientific Laboratory Supplies Limited, UK), 2.5 µL of 2 µM primer working stock, 5 µL of MGW and 5 µL of DNA or virus extract. PCR was performed using a Prime Thermal Cycler (Techne, UK), cycling as follows: 96 °C for 5 minutes; [94 °C for 30 seconds; 56 °C for 30 seconds; 72 °C for 1 minute] x35 cycles; 72 °C for 10 minutes 30 seconds. PCR products were stored at 4 °C.

2.3.3 Agarose gel electrophoresis for detection of CSSV following Endpoint PCR using *T. cacao* DNA extractions or extractions from *T. cacao* stems

PCR product aliquots (6 µl amplicon + 2 µl loading buffer [Meridian Bioscience; Scientific Laboratory Supplies Limited, UK]) were size-fractionated by

electrophoresis through a 1 % w/v agarose gel, pre-stained with SYBR safe (Invitrogen, USA; 15 µl per gel) at 90 v (400 mA) for 45 min.

2.3.4 TaqMan qPCR for amplification and detection of CSSV using *T. cacao* DNA extractions or extractions from *T. cacao* stems

CSSV DNA present in DNA extractions from *T. cacao* leaf discs was quantified using qPCR and CSSV specific primers (Table 2). A 10 µM primer/2.5 µM TaqMan probe working stock was made using 10 µL of 100 µM forward primer, 10 µL of 100 µM reverse primer, 2.5 µL of 100 µM probe and 87.5 µL of molecular grade water. Typical master mix for a 20 µL reaction included 10 µL SensiFAST Probe No-ROX (Meridian Bioscience; Scientific Laboratory Supplies Limited, UK), 0.8 µL primer/probe working stock, 4.2 µL molecular grade water and 5 µL DNA extract. Synthetic CSSV RTase region reference samples (Table 3) at concentrations of $5x10^5$, $5x10^4$, and $5x10^3$ copies μL^{-1} were run with samples to create a CT curve for quantification of CSSV copies in each sample, and a non-template control of molecular grade water was used. Reactions were run in a Rotor-Gene Q 2-plex (QIAGEN, UK), cycling as follows: 95 °C for 10 minutes; [95 °C for 10 seconds (acquiring on green), 60 °C for 45 seconds (acquiring on green)] x45 cycles. Results were analysed using Rotor-Gene 6000 Series Software.

2.3.5 PCR clean-up, Sanger sequencing and analysis

PCR products of interest were prepared for DNA sequencing using GeneJET PCR Purification Kit (Thermo fisher, UK). Following purification, PCR products were kept at either -20 °C or 4 °C before being sent for sequencing. Prepared PCR products, and primers at a concentration of 3.2 pM, were sent to Source Bioscience (UK, Reading) for Sanger sequencing. Trimming and filtering of PCR products, multiple alignment and phylogenetic analysis was performed using Geneious Prime.

Table 2: PCR primer/probe combinations used for molecular methods.

Specificity	Primer Name	Primer sequence (5'-3')	Ta (°C)	Amplicon size (bp)
RT RNase H region				
Badnavirus	Badna1deg2	CCATCCCTTGGACHGCNTTYTGGGT	63	628
	Badna4deg2	TTACATACGGCNCCCCAHCCYTCCAT		
CSSV	RTaseN F	ACCATCCTGAAGCGAGTAGG	59.4	107
	RTaseN R	AAAGCAGTCCAAGGGATGGA	57.3	
	RTaseN P	[FAM]TCCATCAGGTTGCCATGGCA[BHQ1]	59.4	

Table 3: CSSV or Integrated Sequence synthetic reference sequences used in molecular methods.

Sequence name	Oligo sequence (5'-3')	Reference region
CSSV reference sequence	CTGAAGCGAGTAGGCAACAAAAAAATTTTCAGCAAGTTTGATTTAAAATCAGGCTTCCATCAGGTTGCCATGGCAGAAGAGTCCATCCCTTGGACTG	CSSV RTase

CHAPTER THREE

Development of a method for the extraction and detection of CSSV DNA from *Theobroma cacao* stem exudate

3.1 Introduction

3.1.1 Plant biosecurity and quarantine procedures

Agricultural and food diversification has improved through international exchange of germplasm, with up to 68 % of national food supplies now derived from non-native crops (Kumar *et al*., 2021). Crop improvement successes through international scientific partnerships and the flow of plant genetic resources have been improved through the implementation of the International Treaty for Plant Genetic Resources for Food and Agriculture of the Food and Agriculture Organisation (FAO; Enriquez *et al*., 2022). These improvements pose a threat in the form of the spread of pests and diseases through unrestricted transportation of germplasm, including botanical seeds and vegetative propagules. As such, the regulation of germplasm transportation through compliance with the FAO-International Plant Protection Convention procedures must be followed, to ensure the spread of pest and diseases are kept to a minimum, if not negated entirely (Kumar *et al*., 2021).

3.1.2 Quarantine procedures in place for the movement of *T. cacao* germplasms

Cocoa is a globally important cash crop which loses up to 38% of harvests to pest and disease (Marelli *et al*., 2019), making the production of new and improved varieties comprising genetic pest and disease resistances, while retaining qualities that increase yield and flavour, are a hugely important part of the industry. To prevent the spread of pests and diseases when transporting cocoa planting material from one region to another, germplasm must first be sent to an Intermediate Quarantine Centre in a country where cocoa is not cultivated. Germplasm is subjected to a quarantine process involving rigorous testing for many different pests and diseases, such as *Moniliophthora perniciosa* the causative agent of Witches' broom disease, to CSSV, to Cocoa pod borer infestations. Intermediate quarantine centres are currently located in Reading, UK (ICQC) and Florida, USA (United States

Department of Agriculture; End *et al*., 2021). Specific guidelines have been set out to ensure the safe movement of *T. cacao* germplasm worldwide. These guidelines include considerations for the movement of seed, budwood, whole plants, *in vitro*, pollen and open flowers and flower buds (End *et al*., 2021).

3.1.3 Cocoa quarantine processes

It is important that all constituent parts of the plant, including any *in vitro* plant material, are inspected by a quarantine centre before arriving at their intended destination. The safest way of transporting cocoa germplasm is by seed following inspection by hand lens or microscope. Some pathogens such as *M. perniciosa*, and *M. roreri* the causative agent of Frosty pod rot, are capable of being transmitted by seed, so fungicide application is recommended before transport. The transportation of pods and whole plants in soil is not recommended, as these may contain fungal pathogens, invertebrate pests and soil-borne organisms. When a genetically identical copy of a particular genotype is required, cocoa germplasm may be moved as budwood. Budwood may, however, be infected by fungal pathogens and with a number of different viruses, for example CSSV. Viral pathogens are, by their nature, challenging to detect during quarantine due to their potentially protracted latent periods on host plants (End *et al*., 2021).

3.1.3.1 Detection of common *T. cacao* pathogens in quarantine

Around 23 different pest and diseases found on *T. cacao* germplasm are monitored by the ICQC, with quarantine measures outlined for 14. Some of the highest risk infections include Witches' broom disease, Phytophthora Pod Rot and Cocoa pod borer. Quarantine measures will differ depending on the disease; for example, for fungal infections such as those caused by *M. perniciosa* and *Phytophthora spp*., newly introduced material is kept in isolation for at least a year before being released for general use. For other infections, such as by the Cocoa Fruit Borer (*Carmenta spp*.), germplasm may be released following washing and treatment with an appropriate insecticide (End *et al.*, 2021).

3.1.4 Current methods of detection of CSSV positive germplasm in cocoa quarantine centres

All budwood samples intended for transport to a different cocoa producing location must go through an intermediate quarantine station and undergo a virus indexing procedure to identify potential viral pathogens, such as CSSV. The current virus indexing procedure involves the grafting of budwood onto Amelonado cocoa seedlings (Figure 8). Amelonado seedlings are known to be sensitive to badnavirus infection, therefore any symptoms of viruses such as CSSV are likely to be conspicuous; however, some plant stress responses from nutritional deficiencies can be mistaken for biotic disease symptoms (Kaur and Guatam, 2021). Three successful budded seedlings are recommended per plant being tested. Once a seedling-bud union has formed, the leaves and stems are inspected for characteristic leaf symptoms and swellings associated with cocoa viruses. Inspections occur weekly over a period of two-years to compensate for the latency period between infection and symptom development; if viral symptoms are observed, the test plant and mother plant are destroyed. During the observation stage these grafted plants are maintained in polytunnels at the ICQC separate from the main germplasm collection (End *et al*., 2021). Limitations in this process, such as distinguishing between biotic and abiotic disease symptoms; the two-year quarantine period; and infection and symptom development latency period could be improved through the implementation of a quick, molecular method for the detection of CSSV prior to the first round of budwood-grafting.

Figure 8: Budwood graft of a *T. cacao* strain in quarantine on a *T. cacao* Amelonado root stock.

3.1.5 Potential for elimination of CSSV by somatic embryogenesis

In 2008, Quainoo *et al*. trailed the use of somatic embryogenesis, a tissue culture technique, in order to generate CSSV-free clonal propagules from CSSV-infected *T. cacao* trees, in an attempt to aid in CSSV control measures. While the experiment was successful, with all secondary somatic embryos testing negative for CSSV by PCR/agarose electrophoresis, this process of elimination is a time-consuming process with relatively high set-up costs.

3.1.6 Current methods of CSSV detection

In addition to budwood grafting and observing test plants for symptoms of CSSV, various molecular diagnostic methods are used outside of the ICQC for detection. These include Enzyme-linked immunosorbent assay (ELISA) and PCR-based detection (Dzahini-Obiatey *et al*., 2009; End *et al*., 2021).

3.1.6.1 Use of serological methods such as ELISA for the detection of CSSV

Serological tests such as ELISA are decisive in identifying presence of an unknown virus or studying relatedness of virus isolates. Antibody isolates must be obtained in

order to perform serological testing; this requires a highly purified viral preparation. (Dzahini-Obiatey *et al*., 2009). Several effective anti-sera have been raised against CSSV for use in ELISA studies (Sagemann *et al*., 1985; Muller, 2009). Although the ELISA method was fast and reliable, due to the weak immunogenicity of CSSV, high background values were detected and there were difficulties detecting CSSV in latent infections with low viral titres (Sagemann *et al*., 1985; Dzahini-Obiatey *et al*., 2009). An alternative method that makes use of anti-sera known as immunosorbent electron microscopy has also been used to successfully detect high concentrations of CSSV in *T. cacao* samples, but its ability to detect latent, low titre CSSV infections has not been tested (Dzahini-Obiatey *et al*., 2009).

3.1.6.2 DNA based methods for CSSV detection

The genetic diversity of CSSV makes detection using DNA based methods such as PCR challenging. In 2018, Muller *et al*. used next generation sequencing to compile genomes for 21 new CSSV-related species, of which 10 appear to cause disease in *T. cacao*. This diversity complicates the diagnostic process; it is unlikely that a single PCR assay can be developed for the identification of all of these CSSD responsible species. However, advances in qPCR, specifically the use of qPCR may provide a solution to this problem.

3.1.7 Comparison of Cocoa quarantine to other crop quarantine procedures: *Musa* spp.

The International Transit Centre (ITC; Bioversity International, Katholieke Universiteit Leuven, Belgium) holds the largest *Musa in vitro* collection in the world and is essential in the safe exchange of *Musa* germplasm. The major pathogen contaminants found in *in vitro Musa* germplasm re viruses, and this ITC guidelines only cover viral diseases of *Musa*. Technical procedures for the virus indexing in both vegetative material and seeds involve the use of tissue culture and virus indexing over the course of six months (Thomas, 2015); a quarter of the time dedicated to indexing Cocoa germplasm for CSSV (End *et al.,* 2021). Unlike the Technical Guidelines for the Safe Transport of Cacao Germplasm (End *et al.,* 2021),

the Technical Guidelines for the Safe Movement of *Musa* Germplasm (Thomas, 2015) suggest the use of serological and molecular diagnostics for identification of viral pathogens; several different techniques such as immunocapture PCR, rolling circle amplification and ELISA have been developed for the detection of different viral *Musa* pathogens (Thomas, 2015). However, the genetic diversity of *Musa* viral pathogens, such as *Banana mild mosaic virus*, and in the case of *Banana Streak Virus* (BSV) the sometimes-present endogenous sequences, can interfere with PCR-based detection methods; in these cases, the serological methods are preferred. The same issue is attributed to the detection of CSSV in cocoa germplasm. This, along with the lack of development of alternative molecular and serological detection methods is why visual indexing is still used (End *et al*., 2021).

3.1.8 Aims and objectives

In order to aid with cocoa quarantine processes, this chapter aims to develop a cheap and quick method for CSSV DNA extraction and qPCR detection from *T. cacao* stem exudate, to offset the high costs of the maintenance, materials and specialist facilities required for the current two-year visual indexing process used for CSSV detection in the ICQC (End *et al*., 2021). CSSV has long been characterised as a phloem limited pathogen (Jacquot *et al.*, 1999). This method could be used to extract CSSV DNA from phloem-disrupted budwood samples which are routinely sent to quarantine centres, which would significantly speed up sample processing times and therefore germplasm movement through enhanced germplasm screening, supporting future cocoa research. .

3.2 Methods

3.2.1 Using *T. cacao* stem samples for CSSV detection

The process of detection of CSSV in *T. cacao* currently involves sampling a 1 cm diameter section of leaf tissue, performing a DNA extraction followed by either endpoint PCR or qPCR. It is time consuming and labour intensive to process large sample sizes using this approach. An alternative method is proposed for CSSV detection without the need for DNA extraction.

3.2.2 *T. cacao* stem sample preparation

Samples were based on cocoa stem sections of the approximate dimensions required for international movement of budwood i.e. approx. 8 mm diameter and 15 cm long. A 2 cm sub-section of such a cutting was taken from CSSV-infected *T. cacao* plants (code G#, Table 1) maintained in the UWE Envirotron. Using sterile secateurs, a shallow cut was made down the stem lengthways and the bark was peeled back at the cambium (Figure 9). Both the peeled bark (outer layer) and heartwood core (inner layer) were submerged in 2 mL deionized water in a bijou bottle and left to incubate at room temperature for 30 minutes, 2 hours or 24 hours on a tube roller (Stuart, UK). Following incubation and rolling, 500 µL samples were taken from each bijou and kept at -20°C.

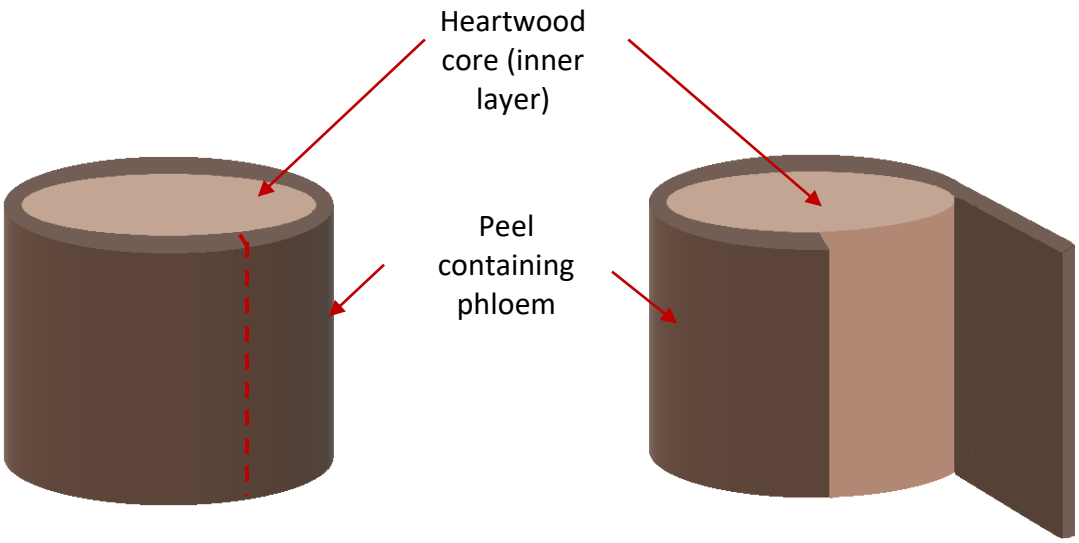


Figure 9: Preparation of stems for CSSV extraction. Using sterile secateurs, a shallow slice is made lengthways down *T. cacao* stems. The peel (outer layer) is pulled away from the heartwood core (inner layer), exposing the phloem tissue.

3.2.3 Sampling from *T. cacao* leaves, DNA extractions and detection of CSSV DNA using qPCR

Sampling of *T. cacao* leaves, DNA extractions and detection of CSSV DNA using qPCR were performed as described in 2.3.1 and 2.3.4.

3.2.4 Scanning electron microscopy for imaging of peeled *T. cacao* stem samples and intact *T. cacao* stem samples

Using a sterile one-sided razor, 1 mm long cuttings of woody stem were taken from uninfected *T. cacao* plants maintained in the UWE Envirotron in a separate compartment from that with the CSSV-infected plants. Samples were either kept intact or had a small section of the bark cut and peeled off and were imaged wet using the FEI Quanta FEG 650 SEM with the Gaseous Secondary Electron Detector attached.

3.3 Results

Samples of *T. cacao* stems were prepared as described in Figure 9. For all following qPCR experiments, RTaseN primers and probe were used (Table 2). All qPCR analyses were performed using a positive control in the form of synthetic CSSV reference samples (Table 3), and a negative water control to check that no contamination occurred during set up.

3.3.1 Presence of CSSV in undiluted and diluted stem samples

Extractions from stem samples taken from CSSV infected *T. cacao* plants G13, G20, G30 and G91 were either left undiluted, or diluted 1:10 and 1:20 in MGW before undergoing qPCR to detect the concentration of CSSV copies μL^{-1} and to establish if the presence of PCR inhibitors in the phloem extract was affecting the approach's capacity for CSSV detection. Using known concentrations of synthetic CSSV DNA fragments, a C_T curve was created, and the CSSV concentration in copies μL^{-1} was calculated. A Shapiro-Wilk test for normal distribution determined that the dataset is normally distributed (p = > .05). Multiple box and whisker plots were created in Prism Graph-Pad 9 using the Tukey method, and a one-way ANOVA with Tukey's multiple comparison test was used to compare the means of each dilution within plant samples (Figure 10). Statistical significance is indicated when p = < .05.

CSSV was not detected via qPCR in any of the samples left undiluted, while consistently detected in samples diluted 1:10 and 1:20. The detected concentration of CSSV was higher in samples diluted 1:10 than in the samples diluted 1:20. A negative technical replicate of G13 diluted 1:10 (Figure 10a) was identified as failed PCR. Similarly, a negative technical replicate was identified in one sample from G20 diluted 1:20 (Figure 10b). Overall, these two negative samples represented x% of the all the diluted samples analysed from the infected plants.

In all cases, samples that were diluted 1:10 display a higher median value than the samples diluted 1:20, indicating a higher detected concentration of CSSV present in the 1:10 diluted samples.

Large inter-quartile ranges seen in samples G13 1:10 dilution, and both 1:20 dilutions for G20 and G30 indicate a large spread of data (Figure 10a, b & c). The remaining samples have a normal distribution of data, indicated by the small inter-quartile ranges.

Irregular positioning of the median value on box plots suggests skewed data, with a higher median indicating a negative skew and lower median indicating a positive skew. G13 1:20 dilution is an exception, where the median appears in the middle of the inter-quartile range, indicating a symmetrical distribution and no skew (Figure 10a).

In all biological replicates, comparison of 1:10 dilutions to undiluted samples results in a p-value of < .0001, indicating an extremely significant difference (Figure 10). When comparing 1:20 dilutions to undiluted samples from G13, results show that the p-value < .0001, indicating an extremely significant difference. However, no significant difference was found when comparing the 1:10 dilution to 1:20 dilution ($p = \geq .05$; Figure 10a). Comparison of G20 dilutions 1:20 to undiluted samples result in a p-value < 0.001, indicating a significant difference. A p-value of < .05 is identified when comparing samples diluted 1:10 and 1:20, indicating a significant difference (Figure 10b). When comparing 1:20 dilutions to undiluted samples from G30, results show that the $p = < .0001$, indicating an extremely significant difference. Comparison of the 1:10 dilution and 1:20 dilution result in a p-value of < .01, indicating an extremely significant difference (Figure 10c). A p-value of < .0001 is identified when comparing samples from G91; diluted 1:20 undiluted; and diluted 1:10 and diluted 1:20, indicating an extremely significant difference (Figure 10d).

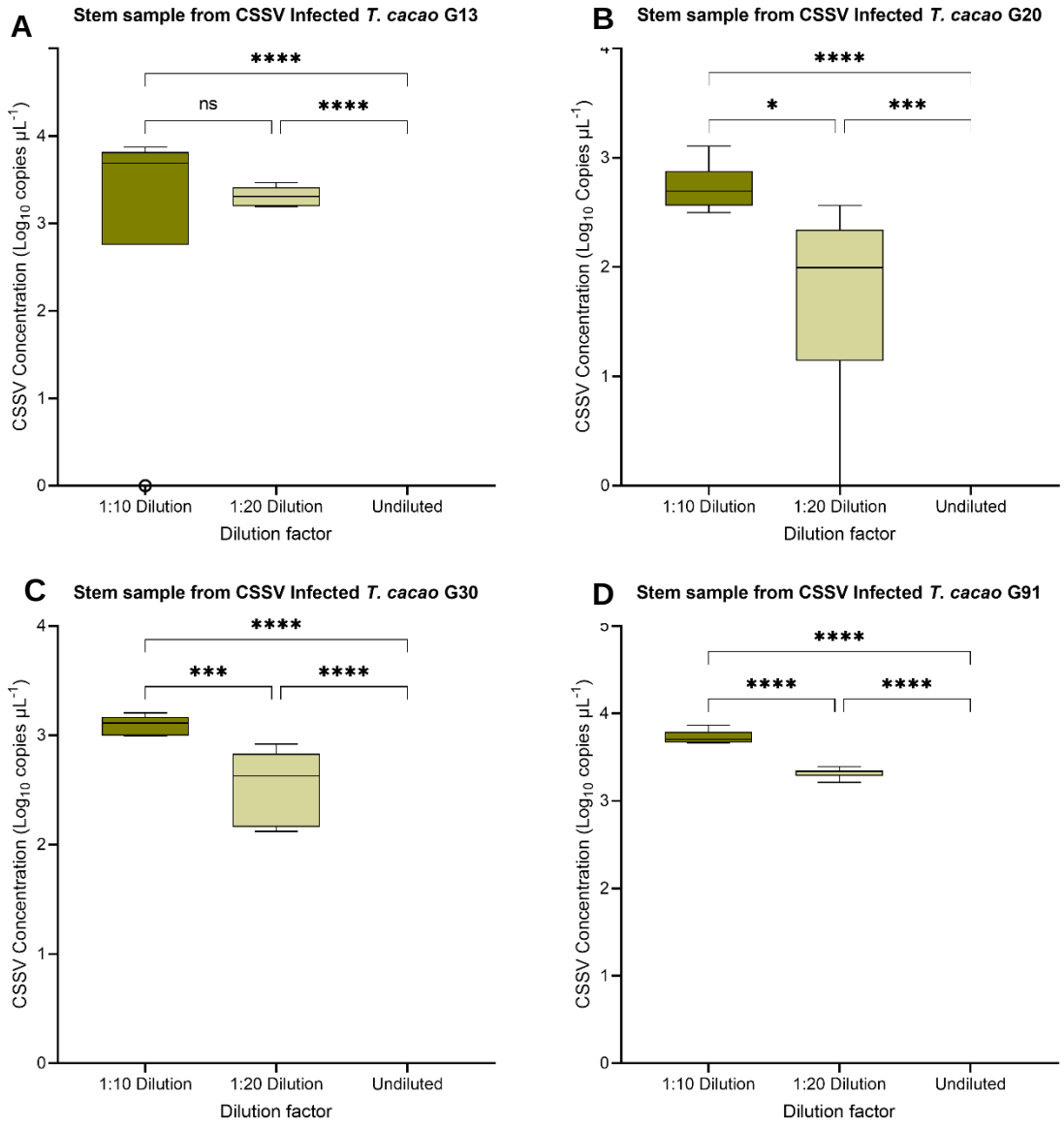


Figure 10: Comparison of CSSV Concentration (Log_{10} CSSV copies μL^{-1}) of diluted samples of CSSV infected *T. cacao* G13 (a), G20 (b), G30 (c) and G91 (d). Tukey box and whisker plots show the lowest data point, lower quartile, median, upper quartile, and highest data point. Significance is illustrated above each data point using GraphPad style: **** $p < .0001$, *** $p < .001$, ** $p < .01$, * $p < .05$, ns $p \geq .05$ (one-way ANOVA with Tukey's multiple comparison test). $n = 6$.

3.3.2 Identification of presence of qPCR inhibitors in stem samples prior to dilution

Stem samples taken from non-infected *T. cacao* (Table 1) were incubated for 2 hours at room temperature. The synthetic reference sequence (Table 3) at a concentration of $5x10^5$ copies μL^{-1} was used to 'spike' the undiluted sample. This spiked sample was then diluted 1:10, 1:20, 1:40 and 1:80. Using known concentrations of synthetic CSSV DNA fragments, a C_T curve was created, and the CSSV concentration in copies μL^{-1} was calculated.

The synthetic sequence was consistently detected in all diluted samples (Figure 11). No synthetic sequence was detected in the undiluted sample, indicating that there is potential for a PCR inhibitor to be present. The concentration of CSSV detected in the samples decreased as the sample was diluted.

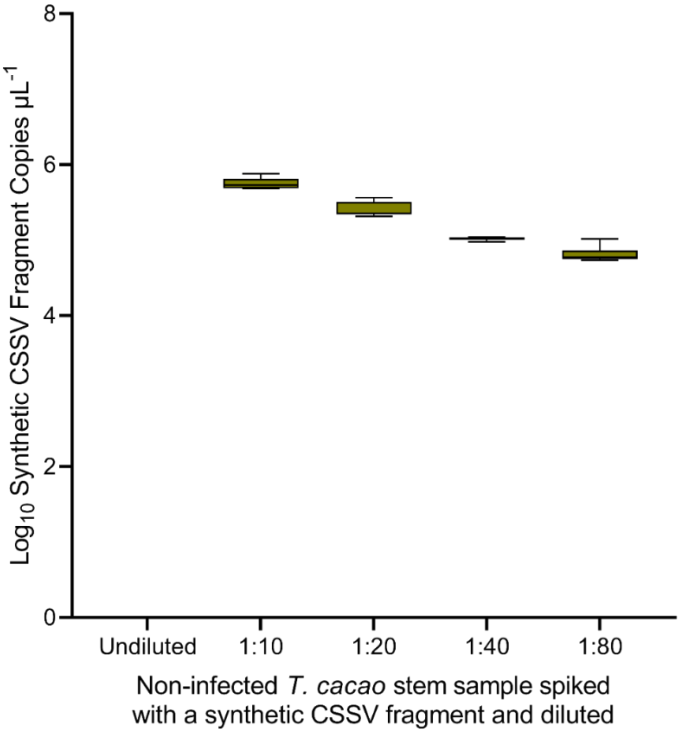
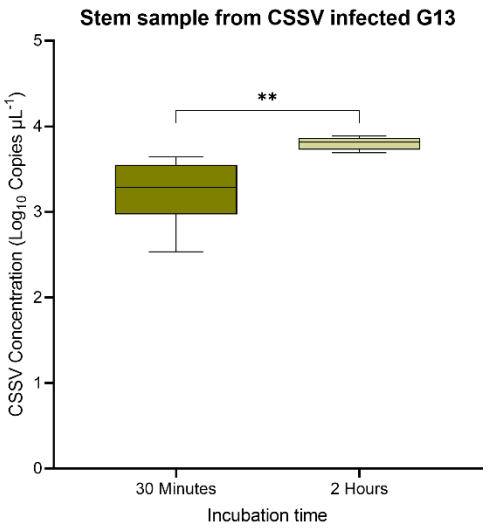


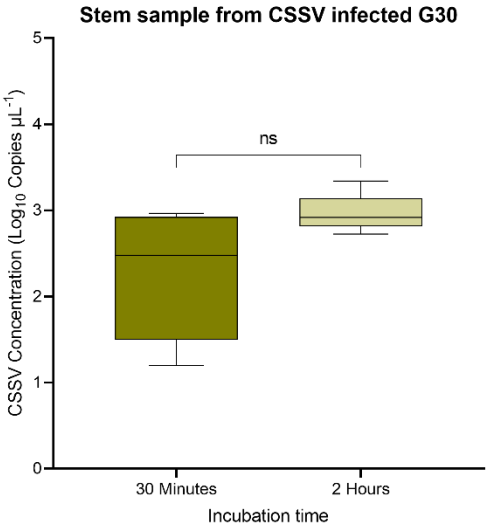
Figure 11: Log_{10} CSSV Copies μL^{-1} detected by qPCR. Synthetic reference samples at a concentration of $5x10^5$ copies μL^{-1} were mixed with stem samples taken from an uninfected *T. cacao* plant, then diluted to 1:10, 1:20, 1:40 and 1:80. Box and whisker plots show the lowest data point, lower quartile, median, upper quartile, and highest data point. *n* = 3.

3.3.3 Concentration of CSSV in stem exudate following an increase in incubation duration

Stems of CSSV infected *T. cacao* plants G13, G30 and G44 were incubated for either 30 minutes or two hours and diluted 1:10 before undergoing qPCR to determine the concentration of CSSV copies μL^{-1}. Using known concentrations of synthetic CSSV DNA fragments, a C_T curve was created, and the CSSV concentration in copies μL^{-1} was calculated (Figure 12). A Shapiro-Wilk test for normal distribution determined that the dataset is normally distributed (p = > .05). Multiple box and whisker plots were created in Prism Graph-Pad 9 using the Tukey method, and an unpaired t-test was used to compare the means of each dilution within plant samples (Figure 12). Statistical significance is indicated when p = < .05.

CSSV was detected in all three G plant samples following 30 minute and two-hour incubation steps. The 30-minute incubation data is spread about a wider range, whereas the two-hour incubation results were more closely centred around the median.





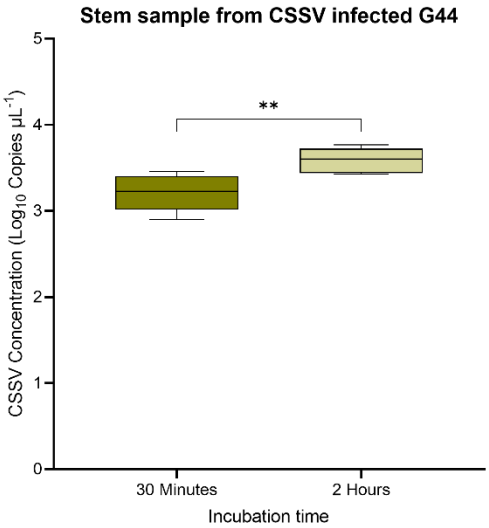


Figure 12: Effect of incubation duration on yield of CSSV (Log_{10} copies μL^{-1}) from *T. cacao* stem sections. Tukey box and whisker plots show the lowest data point, lower quartile, median, upper quartile, and highest data point. An unpaired t-test was performed on each data set, with significance illustrated above each data point using GraphPad style: * $p > .05$, ns $p \geq .05$. $n = 6$.

3.3.4 Effects of storage of stems on CSSV detection – emulating the shipping process

Stems of CSSV infected *T. cacao* plants G13, G30 and G44 were incubated for CSSV extraction either immediately following collection or stored at 20 °C ± 2 °C in a moist paper towel in plastic bag for a week prior to extraction, emulating the conditions in which a budwood sample would be transported. Samples were diluted 1:10 before undergoing qPCR to determine the concentration of CSSV copies μL^{-1}. Using known concentrations of synthetic CSSV DNA fragments, a C_T curve was created, and the qPCR results were analysed (Figure 13). A Shapiro-Wilk test for normal distribution determined that the dataset is normally distributed (p = > .05). Multiple box and whisker plots were created in Prism Graph-Pad 9 using the Tukey method, and a repeated-measures one-way ANOVA was used to compare the means of each dilution within plant samples (Figure 13). Statistical significance is indicated when p = < .05.

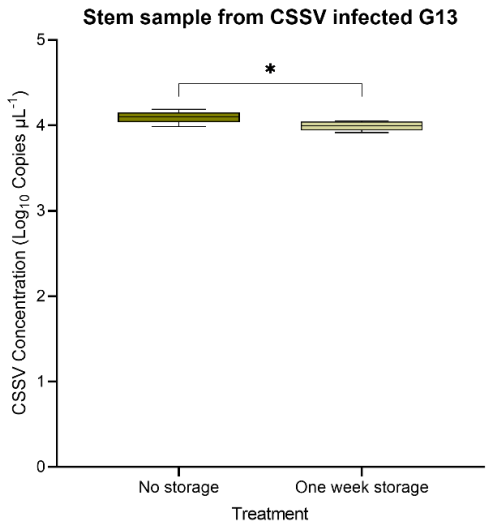
CSSV was consistently detected in all non-stored *T. cacao* samples, and in all but one of the stored sample replicates (G44, Appendix I). One replicate from the sample G30, stored, was identified as an outlier (Figure 13).

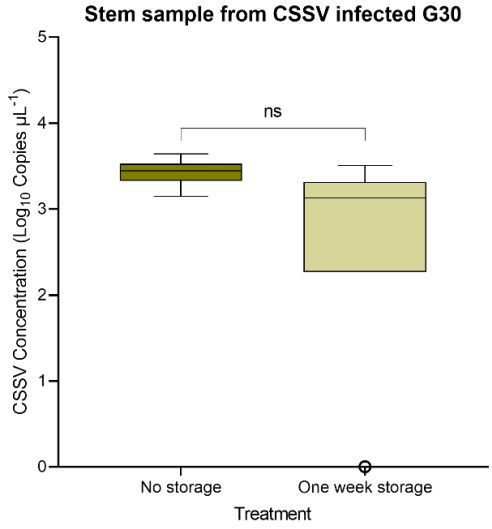
In samples from G13 and G30 the median values are lower in the stored samples compared to the samples extracted immediately, indicating a higher concentration of detectable CSSV present in the non-stored samples. This is not the case with sample G44, where the median values of the stored samples are higher than that of the non-stored samples, indicating a higher concentration of CSSV present in the stored samples (Figure 13).

Large inter-quartile ranges seen in stored samples of G13 and G30 indicate a large data spread (Figure 13a & b). The remaining samples have a normal distribution of data, indicated by the small inter-quartile ranges (Figure 13a, b & c).

Irregular positioning of the median value on box plots suggests skewed data, with a higher median indicating a negative skew. The non-stored samples are an exception, where the median appears in the middle of the inter-quartile range, indicating a symmetrical distribution and no skew (Figure 13a, b & c).

The unpaired t-test indicates that there is no significant difference between the non-stored samples and stored samples in *T. cacao* plants G30 and G44 (ns $p = \geq .05$; Figure 13b & c). However, there is a significant difference when comparing non-stored samples and stored samples of *T. cacao* plant G13 (* $p > .05$; Figure 13a).





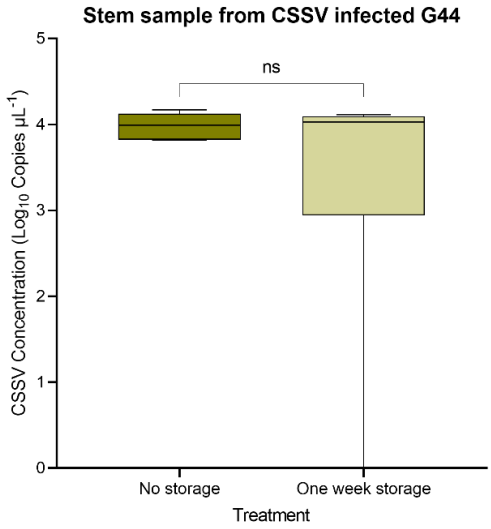


Figure 13: The Log_{10} CSSV copies μL^{-1} detected by qPCR using samples extracted from the stems of CSSV infected *T. cacao* G13, G30 and G44. Stem samples were either extracted immediately or extracted after one week in storage at 20 °C ± 2 °C. Box and whisker plots show the lowest data point, lower quartile, median, upper quartile, and highest data point. An outlier is indicated by a circle on the x-axis. An unpaired t-test was performed on each data set, with significance illustrated above each data point using GraphPad style: * $p > .05$, ns $p \geq .05$. $n = 6$.

3.3.5 Comparison of extraction types: Kit-based leaf DNA extraction vs. novel stem extractions

CSSV concentration (Log_{10} copies μL^{-1}) was measured in DNA extractions from 1 cm diameter discs of *T. cacao* leaf tissue using qPCR. These concentrations were compared to the concentrations of CSSV found in corresponding 2 cm stem extractions. Tukey box and whisker plots were created in Prism Graph Pad 9.

A higher concentration of CSSV DNA can be detected when testing DNA extractions from leaf discs compared to the 1:10 diluted stem extractions from the same CSSV infected *T. cacao* G plant (Figure 14).

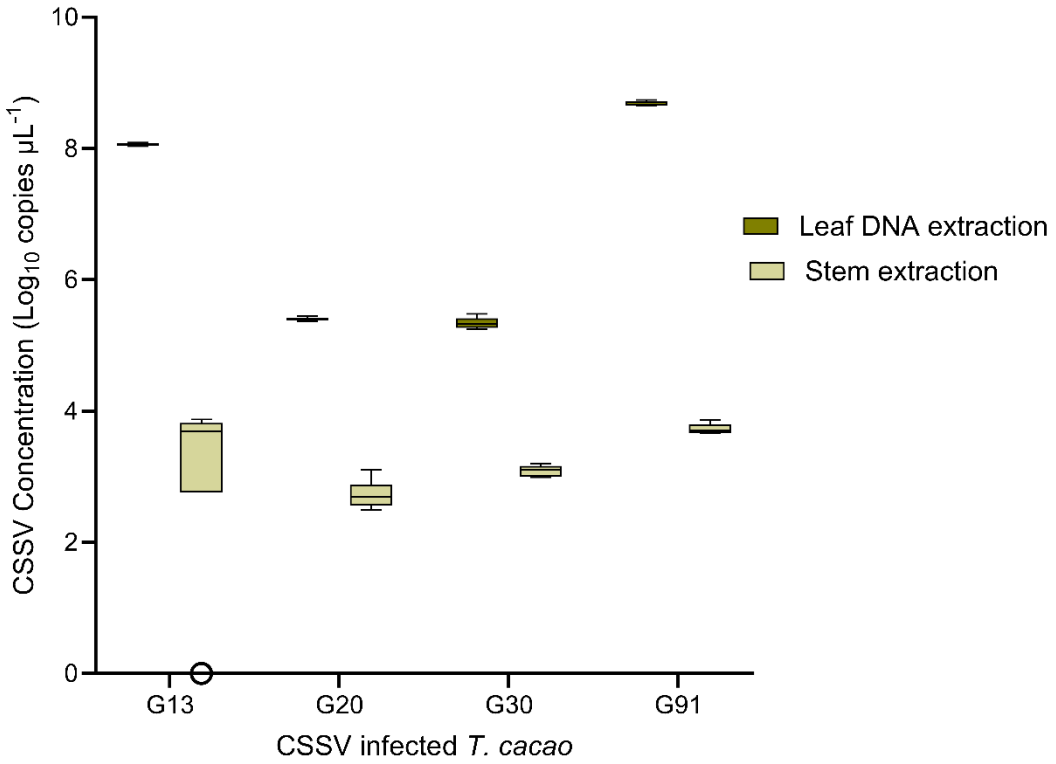


Figure 14: A comparison of the Log_{10} CSSV copies μL^{-1} of four CSSV infected *T. cacao* plants G13, G20, G30 and G91, when using DNA extractions from leaf tissue and the novel stem extraction method (with 1:10 dilution). Tukey box and whisker plots show the lowest data point, lower quartile, median, upper quartile, and highest data point. A negative result is indicated by a circle on the x-axis. *n* = 6.

3.3.6 Can stem section analysis shed light on CSSV location?

Stem samples were taken from CSSV positive *T. cacao* plant G20 and prepared as described (Figure 9). In this case, the outer peel was completely removed from the middle heartwood prior to immersion in water and incubation. The two components were incubated separately. Three biological replicates were used from stem sections of plant G20, and six technical replicates were used for each sample. Using known concentrations of synthetic CSSV DNA fragments, a CT curve was created, and the qPCR results were analysed.

CSSV was rarely detected in the middle heartwood section of stem samples. For replicate two, only three technical replicates out of six produced a positive result. Fewer than three positive results were recorded for the other middle heartwood replicates (two in biological replicate one, and one in biological replicate three) (Appendix 1). Middle heartwood data does not follow a normal distribution, with a large inter-quartile range, indicating a highly dispersed data set, and median values for both replicates one and two lying at 0, indicating a negative skew.

CSSV was consistently detected in the outer peel section of stem samples. All technical replicates were positive in both replicates one and three, while four technical replicates were positive in replicate two (Appendix I).

A Shapiro-Wilk test for normal distribution determined that the dataset is not normally distributed (p = < .05). Each replicate was analysed separately using a Mann-Whitney test to determine the statistical significance. In both biological replicates one and three, the CSSV concentration is statistically significantly different (** p < .001). However, for biological replicate two there is no significant difference in the CSSV concentration between the middle heartwood and outer peel (ns $p \geq .05$) (Figure 15 and Figure 16).

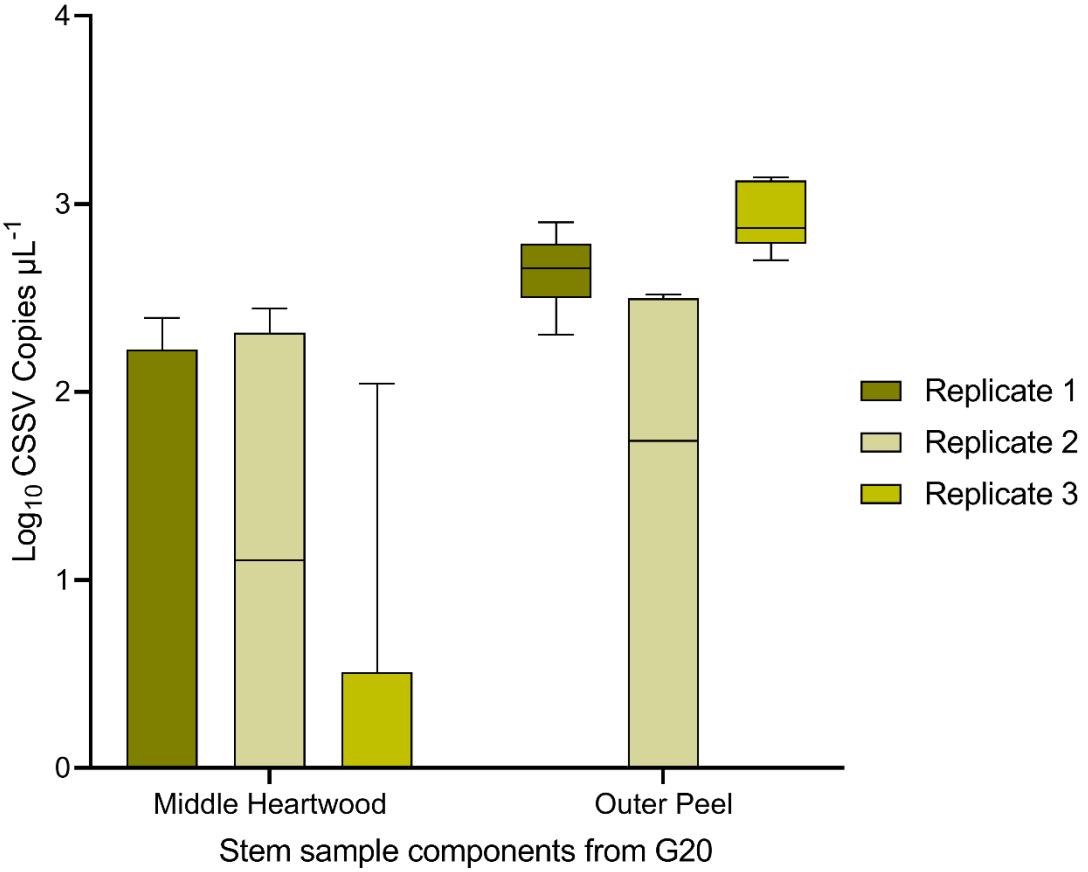


Figure 15: The Log_{10} CSSV copies μL^{-1} detected by qPCR using samples extracted from different sections of the stems of CSSV infected *T. cacao*. Box and whisker plots show the lowest data point, lower quartile, median, upper quartile, and highest data point.

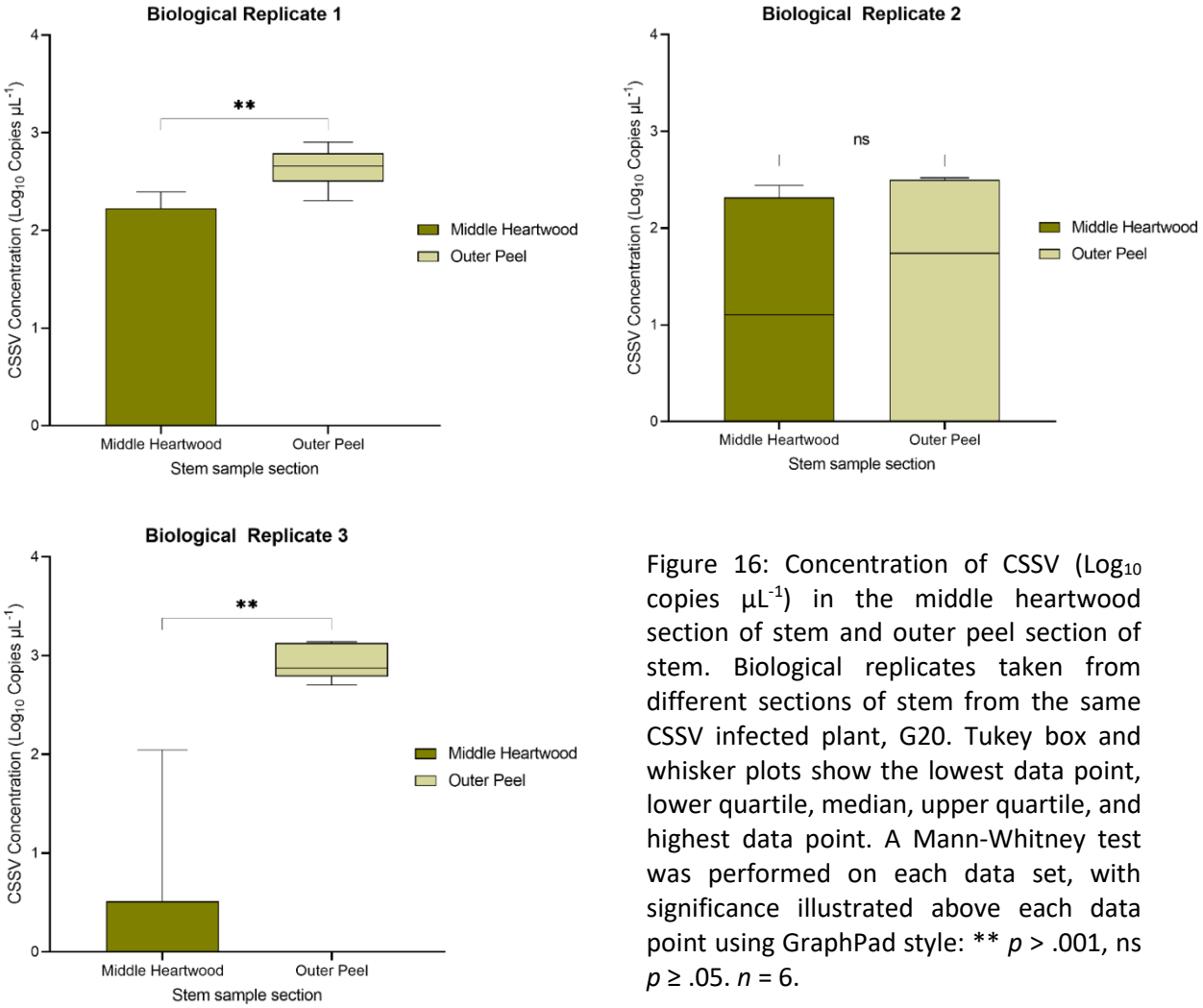


Figure 16: Concentration of CSSV (Log_{10} copies μL^{-1}) in the middle heartwood section of stem and outer peel section of stem. Biological replicates taken from different sections of stem from the same CSSV infected plant, G20. Tukey box and whisker plots show the lowest data point, lower quartile, median, upper quartile, and highest data point. A Mann-Whitney test was performed on each data set, with significance illustrated above each data point using GraphPad style: ** $p > .001$, ns $p \geq .05$. $n = 6$.

3.3.7 Use of scanning electron microscopy and light microscopy to identify tissues disrupted by the stem peeling process

Scanning electron imaging was used on fresh *T. cacao* stem samples in order to determine exactly which tissues were being disrupted during the stem peeling process. A sample of un-peeled stem (Figure 17) was compared to a sample of peeled stem (Figure 18). Figure 17 shows the intact stem and fissure at which the outer layer of stem tissue is peeled away. During stem peeling, the live tissue containing the phloem and vascular cambium are disrupted (Figure 18).

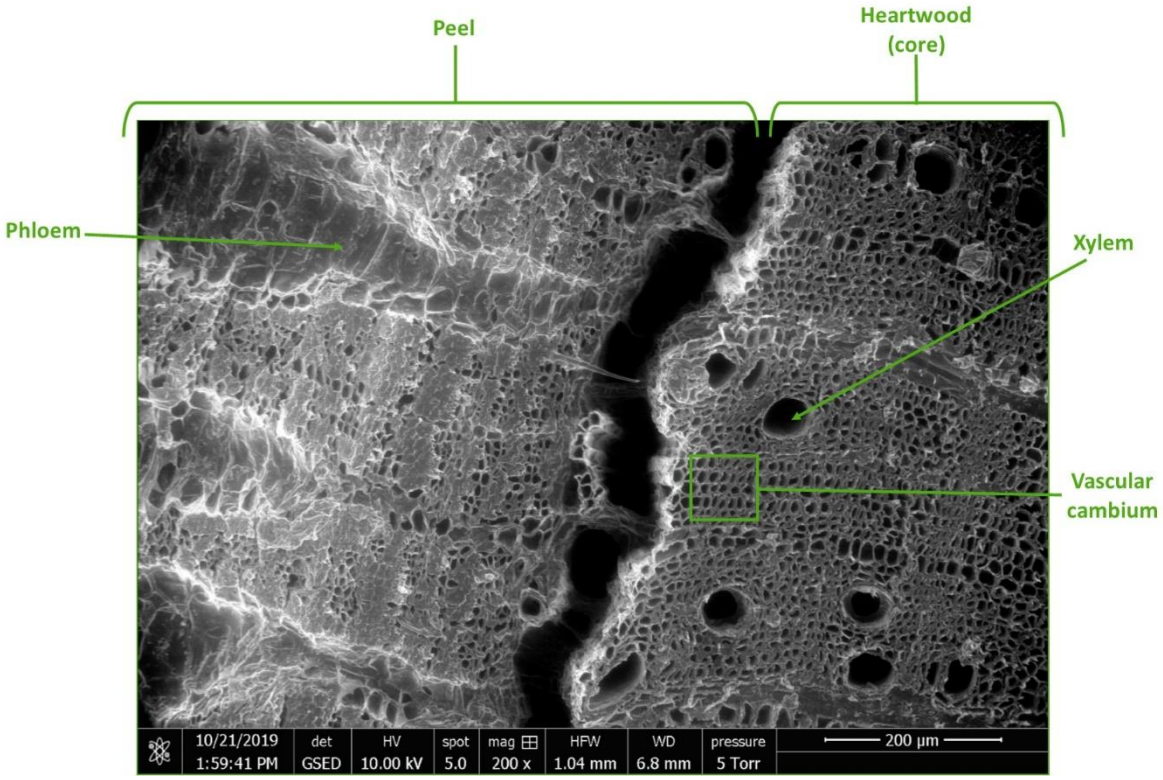


Figure 17: Scanning electron micrograph of a cross section of a freshly cut *T. cacao* stem. Tissue features are highlighted in green, with what would be the 'peel' and 'core' samples labelled, should this stem be processed using the novel stem peeling method for extraction of CSSV DNA.

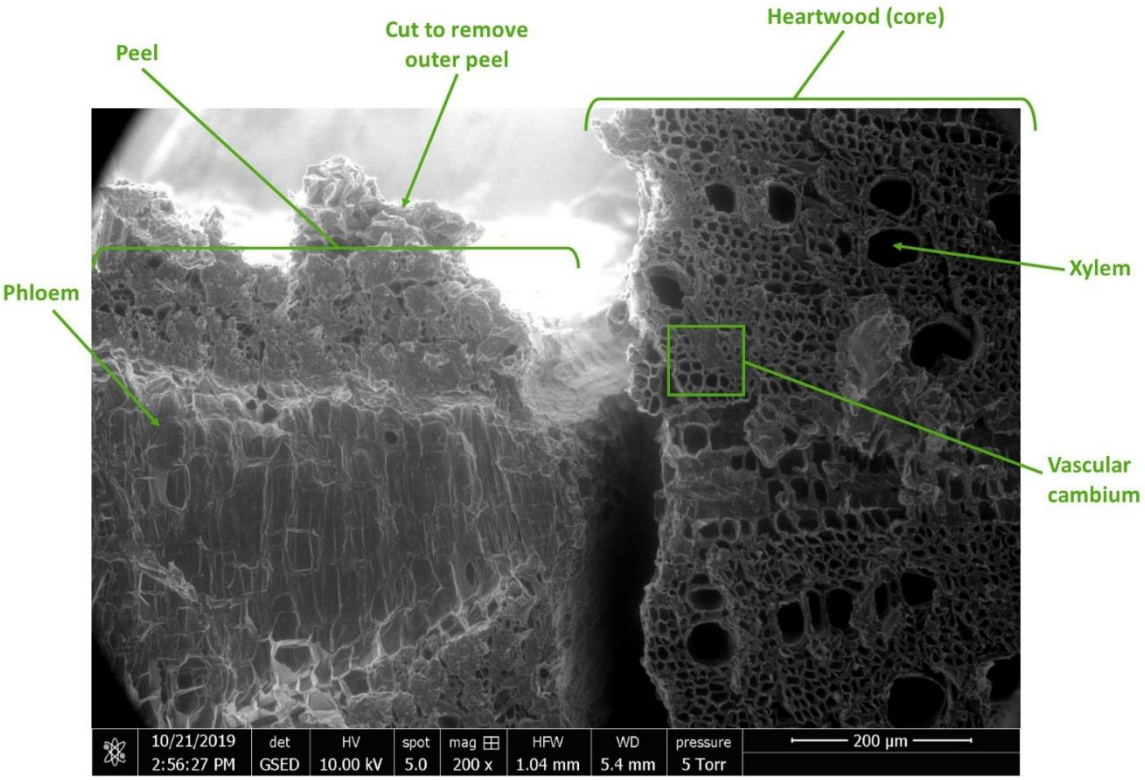


Figure 18: Scanning electron micrograph of a cross section of a freshly cut *T. cacao* stem prepared using the novel stem peeling method. Tissue features are highlighted in green, with what would be the 'peel' and 'core' samples labelled.

3.4 Discussion

The use of stems for CSSV extraction and purification has advantages over using leaves. The mucilaginous property of *T. cacao* leaves makes purification of CSSV virions challenging and poses an issue when attempting to locate virions in planta (Jacquot *et al*., 1999). Another advantage of using the stems for extraction is the ease of transportation and storage. To date, in order to move *T. cacao* germplasm from one country to another for research, crop improvement, plant breeding, exploration or conservation, it must first go through the ICQC to minimize the risk of pest introductions. Usually when transporting cocoa germplasm, a budwood sample is sent via airfreight, which can take up to seven days, and is grafted onto Amelonado cocoa indicator rootstocks. This process will produce a genetically identical copy of a particular genotype and may be infected with transmissible viruses such as CSSV. In order to determine if there is any viral presence through identification of virus associated symptoms, rootstocks are inspected weekly for two years (End *et al*., 2017). With a view to an alternative to this lengthy quarantine process, and to aid with virus purification, a study was conducted to determine if superior viral yields could be achieved using peeled stems rather than the conventional leaf tissue samples.

3.4.1 Can CSSV DNA be detected by qPCR in stem exudate samples?

qPCR was first used to determine whether CSSV presence could be detected in undiluted and diluted samples. Results described in 3.3.1 Presence of CSSV in undiluted and diluted stem samples show that CSSV DNA was consistently (across four biological replicates) undetectable in undiluted samples but could be detected in samples diluted 1:10 and 1:20 (Figure 10). The difference in the concentrations of CSSV detected were variable depending on which CSSV positive *T. cacao* G-plant was chosen; this was to be expected, as levels of CSSV within plant tissues will vary depending on the plant itself and sampling location (Sagemann *et al*., 1985), with

potential PCR inhibitors present in different concentrations depending on the age of the tissue (Ramírez *et al*., 2017). The presence of PCR inhibitors could also explain the lack of any CSSV presence in undiluted samples (Figure 10) could be explained high DNA concentrations as a result of the crude extraction process which have an effect on the efficacy of the PCR (Jansson and Hedman, 2019). Alternatively, secondary compounds found in plant tissues that have the ability to inhibit *Taq* polymerase could also be at play(Hedman and Rådström, 2015; Rezadoost *et al*., 2016; Ramírez *et al*., 2017).

3.4.2 Presence of qPCR inhibitors in undiluted stem exudate samples

An internal amplification control method (Hedman and Rådström, 2015) was used in order to determine whether inhibitory PCR molecules are present in undiluted stem exudate samples. Using synthetic oligos at known concentrations (Table 3), routinely used in qPCR as positive controls, undiluted samples of stem exudate from CSSV-free *T. cacao* (RB1; Table 1) were inoculated before dilution and qPCR. Results from this experiment indicate that there could be a qPCR inhibitory factor present in undiluted stem exudate, based on the lack of CSSV detection in undiluted samples when compared to all other diluted samples (Figure 11). Plant metabolites such as polysaccharides have a structural similarity to that of DNA and RNA, which allows them to interfere with the activation of *Taq* polymerase. In order to reduce the inhibitory effects of qPCR plant inhibitors such as polysaccharides and phenols, samples can be treated. However, this would involve the use of reagents such as a non-ionic detergent (Hedman and Rådström, 2012) or chloroform (Demeke and Jenkins, 2010), which unnecessary when dilution of samples allows adequate CSSV detection. Alternatively, the crude stem extracts could be passed through silica-based columns as a safer, cheap alternative to the use of phenols (Kemp *et al*., 2006).

3.4.3 Effect of incubation times on stem exudate samples

In order to determine the most effective incubation time for the extraction of CSSV DNA from stem samples, two different incubation times were tested: 30-minutes and two-hours (Figure 12). Results indicate that an increase in the concentration of CSSV DNA can be detected after a two-hour incubation when compared to a 30-minute incubation, with significant differences recorded in concentrations of two biological replicates (G13 and G44; Figure 12). This result could be attributed the ability of CSSV DNA to move out of stem tissues and into the surrounding water; the longer the stems are left, the more CSSV DNA is likely to diffuse out into a lower concentration environment. While the concentration of CSSV DNA is higher in samples incubated for two-hours, there are still consistent levels of detection in the 30-minute incubated samples. From this, a two-hour incubation period is suggested for the reliable detection of CSSV DNA, although a minimum 30-minute incubation is sufficient when there are time constraints with regards to sample processing.

3.4.4 Can CSSV DNA still be detected in stem samples after being stored for one-week?

The effect of one-week storage at 20 °C ± 2 °C on the stem samples to imitate stem transport was assessed. Figure 15 shows that the stems that underwent preparation and incubation immediately contain a higher concentration of CSSV than those that were stored for a week. This could be due to degradation of CSSV DNA during the storage process, or simply differences in CSSV concentration between sample sites (Sagemann *et al*., 1985). While CSSV DNA was detected in all samples prepared immediately, only five of six technical replicates were CSSV positive in two week-stored biological replicates. Given this result, it is recommended to use a minimum of three technical replicates when using this assay on week-long stored samples to achieve reliable results for presence/absence of CSSV DNA.

3.4.5 How do qPCR calculated CSSV concentrations in CSSV positive *T. cacao* leaf DNA extractions compare to stem extracted samples?

The difference in detectable CSSV DNA in *T. cacao* leaf DNA extracts when compared to stem extracts (diluted 1:10) are significant (**** $p > .0001$, figure #). This difference is likely due to the removal of any PCR inhibitors by DNA extraction kits, and the previous optimisation of DNA extraction on *T. cacao* leaf tissues by Ramírez *et al*. (2017); the optimal DNA extraction kit described in this study was used in all leaf disc DNA extractions for this project (2.3.1 Extraction of CSSV DNA from *T. cacao* leaf discs). Although the concentration of CSSV detected in these DNA extracts was higher than those detected in stem exudate samples, this method is designed as a binary presence/absence detection system for CSSV DNA, and therefore the exact quantification of CSSV concentrations is not necessary. In cases where concentration of CSSV DNA is an important consideration, DNA extractions may be used.

3.4.6 Localisation of CSSV DNA in different stem parts

The location of CSSV within *T. cacao* is something that has been assumed due to CSSV classification as a badnavirus and the characteristics of mealybug vector feeding. It has been shown that the mealybug vector acquires CSSV when feeding in the *T. cacao* vascular system (Obok *et al.,* 2018), which would suggest that CSSV is a vascular tissue limited virus. A study by Jacquot *et al.* (1999) explored the location of an artificially constructed form of CSSV within *T. cacao*. In this study, a greater-than-unit length copy of the CSSV genome was cloned into an *Agrobacterium* binary vector, which was transferred to young *T. cacao* plants via *Agrobacterium tumefaciens*. This research provided information regarding the potential location of wild-type (WT) CSSV within *T. cacao*; the artificial CSSV was detected in phloem companion cells and a few xylem parenchyma cells, suggesting that the CSSV is restricted to the vascular system. Using stem extraction methods and SEM imaging, the location of WT CSSV within infected *T. cacao* could be explored.

A stem sample from infected *T. cacao* was split into two components: the peel and the core. SEM images of stem samples confirm that when stems are peeled, vascular tissue is exposed. Phloem tissue can be seen clearly in the peeled section of the sample (Figure 18), while the majority of the xylem is left in the heartwood core.

Figure 17 shows a gap between the outer vascular tissue and the vascular cambium, which is where the peel is separated from the core. Using qPCR, the concentration of CSSV in the peel and core could be compared. Figure 15 and Figure 16 show that CSSV is limited to the peel; a result that was expected, considering the studies outlined above (Obok *et al*., 2018; Jacquot *et al*., 1999).

This method of CSSV DNA extraction through the incubation of stems in water provides a successful alternative for CSSV DNA detection, which has the potential to be implemented in quarantine centres and used directly on budwood samples, providing an alternative to the lengthy two-year grafting and monitoring procedure (Daymond, 2018; End *et al*., 2021).

CHAPTER FOUR

Identification of a novel badnaviral sequence integrated in the genome of *Theobroma cacao*: a HRM method to distinguish between insert types and modes of inheritance.

4.1 Introduction

4.1.1 Endogenous Viral Elements (EVEs)

The integration of viral elements into the genomes of eukaryotes is a widely recognised phenomenon. This process, known as endogenization, is the result of the chromosomal integration of viral DNA or cDNA in the germ cells of host organisms. These integrated viral sequences are known as endogenous viral elements (EVEs) and can be vertically transmitted through host populations, inherited from parent to offspring in a Mendelian fashion (Katzourakis and Gifford, 2010; Geering *et al*., 2010; Holmes, 2011; Feschotte and Gilbert, 2012). EVEs are often referred to as 'molecular fossils', as they are evidence of past infection events (Vassilieff *et al*., 2023). The first EVEs to be discovered were prophages and endogenous retroviruses; responsible for important human diseases such as rheumatoid arthritis and multiple sclerosis. In more recent times, non-retroviral EVEs which originate from single-stranded (ss) DNA, double-stranded (ds) DNA, ssRNA and dsRNA viruses have been discovered in a wide range of eukaryotic genomes (Vassilieff *et al.*, 2023). Integration into host genomes is not an obligatory step in the life-cycle of these viruses; unlike retroviruses, which require chromosomal integration in order to successfully replicate (Harper *et al*., 2002; Mette *et al.,* 2002). One proposed method of integration of plant viral DNA into plant host genomes is through horizontal gene transfer (HGT), the result of illegitimate recombination during double-stranded DNA repair (Staginnus and Richert-Pöggeler, 2006; Feschotte and Gilbert, 2012). However, some studies have pointed towards the presence of retrotransposon sequence signatures at the EVE-host genome junction, suggesting that enzymatic machinery encoded by retrotransposons may be implicit in these integration events (Feschotte and Gilbert, 2012). An increase in the number of available eukaryotic full-length genomes, and advances in bioinformatic analysis, have resulted in an emergence of important questions with an aim to expand on basic knowledge of eukaryotic genome

composition and the time-scales of virus evolution (Holmes, 2011; Feschotte and Gilbert, 2012).

4.1.2 EVEs located in plant genomes

Two categories of EVEs are recognised within plant genomes: endogenous non-retroviral elements which originate from dsRNA, ssRNA and ssDNA viruses (Chiba *et al*., 2011); and endogenous pararetroviruses (EPRVs) which are derived from the reverse-transcribing dsDNA viruses of *Caulimoviridae* (Boutanaev and Nemchinov, 2021), the latter of which this chapter is concerned with. EVEs belonging to five genera of the *Caulimoviridae* family have been characterised in at least 27 plant species from nine different families (Staginnus *et al*., 2009; Geering *et al.,* 2010). EVE structure and distribution within plant genomes can vary; short, dispersed, repetitive viral sequences such as the endogenous *Rice tungro bacilliform virus*-like sequences found in the genome of *Oryza sativa*, which were found to be segmented and lacking in ORF 2, are not considered infectious, while longer, near full-length viral sequences such as *Banana endogenous virus* (BEV) found in a variety of *Musa* species, can lead to systemic infections of the host plant (Staginnus and Richert-Pöggeler, 2006). Alternatively, in the case of *Tobacco endogenous pararetroviruses* (TEPRVs) in *Nicotiana tabacum*, some EVEs have been shown to serve a beneficial function to the host plant (Mette *et al.,* 2002). The first recorded instance of identification of an integrated viral sequence in a plant genome occurred in 1996, when Bejarano *et al*. discovered multiple repeats of geminiviral-related DNA integrated into the genome of *Nicotiana tabacum* (tobacco). This discovery provided evidence of the evolutionary significance of EVEs in plants and brought into question why these inserts are being retained by plants.

4.1.3 Non-symptomatic EVEs in plants

A study by Diop *et al*. in 2018 identified the presence of at least one *Caulimoviridae* EVE present in 66 tracheophyte genomes. While EVEs have been detected in almost all major groups of higher plants, except for a select few outlined below, there is

little evidence to suggest that their integration has any effect on the expression of host plant genes (Vassilieff *et al*., 2023).

4.1.3 EVE activation in plants

Activation of EVEs, resulting in a phenotype typical of infection by the corresponding virus, has been reported in at least three different genomes. These phenotypes are accompanied by viral particles capable of replication and cell to cell movement (Staginnus and Richert-Pöggeler, 2006). The activation of EVEs has been attributed to plant stresses, such as nutrition stress, tissue culture or wounding (Harper *et al*., 2002).

4.1.2.1 *Banana Streak Virus* (BSV, genus Badnavirus) episomal infection

In 1999, Ndowora *et al*. discovered that the integrated DNA relating to BSV could give rise to an episomal viral infection. Symptoms associated with a BSV infection were identified in tissue cultures resulting from both virus- and disease-free *Musa* spp. DNA recovered from these tissue culture-derived plants was highly similar to BSV DNA recovered from a natural infection. At this time, a partial badnaviral sequence integrated into the *Musa* genome had been discovered and was considered as a possible source of these infections. Two integration sites within the *Musa* genome were identified, one consisting of a partial BSV genome and the second containing a near complete genome almost identical to BSV DNA which was recovered from tissue culture-derived *Musa* plants. Infective endogenous BSV sequences have been found exclusively in the *M. balbisiana* genome. However, several reports have identified partial badnaviral sequences in *M. acuminata* and/or *M. balbisiana* genomes (Chabannes *et al.* 2021).

4.1.2.2 *Petunia vein clearing* (pararetro)virus infection of petunia

In 2003, Richert-Pöggeler *et al*. evidenced the presence of *petunia vein clearing virus* (PVCV) as a constituent of the *Petunia hybrida* genome and demonstrated infection of petunia by an endogenous pararetrovirus of episomal and

chromosomal origin. Some PVCV integrated sequences within the *P. hybrida* genome were found arranged in tandem, which allows the direct release of viral particles by transcription. Induction of these pararetroviral elements was successfully observed following wounding of PVCV positive plants, suggesting that infective forms of the endogenous EVE can form as the result of abiotic stresses.

4.1.4 Genetic diversity of CSSV and host plant *T. cacao*

Recent advances in the understanding of CSSV genetic diversity have brough into question evolution of CSSV with its host *T. cacao*. In 2018, Muller *et al*. reconstructed 21 new, complete genome sequences corresponding to eight CSSV-associated species. Along with these complete genome sequences, an incomplete sequence named species S which clustered separately from other species (Figure 19) was widely detected in asymptomatic cocoa trees, including samples that did not originate from West Africa.

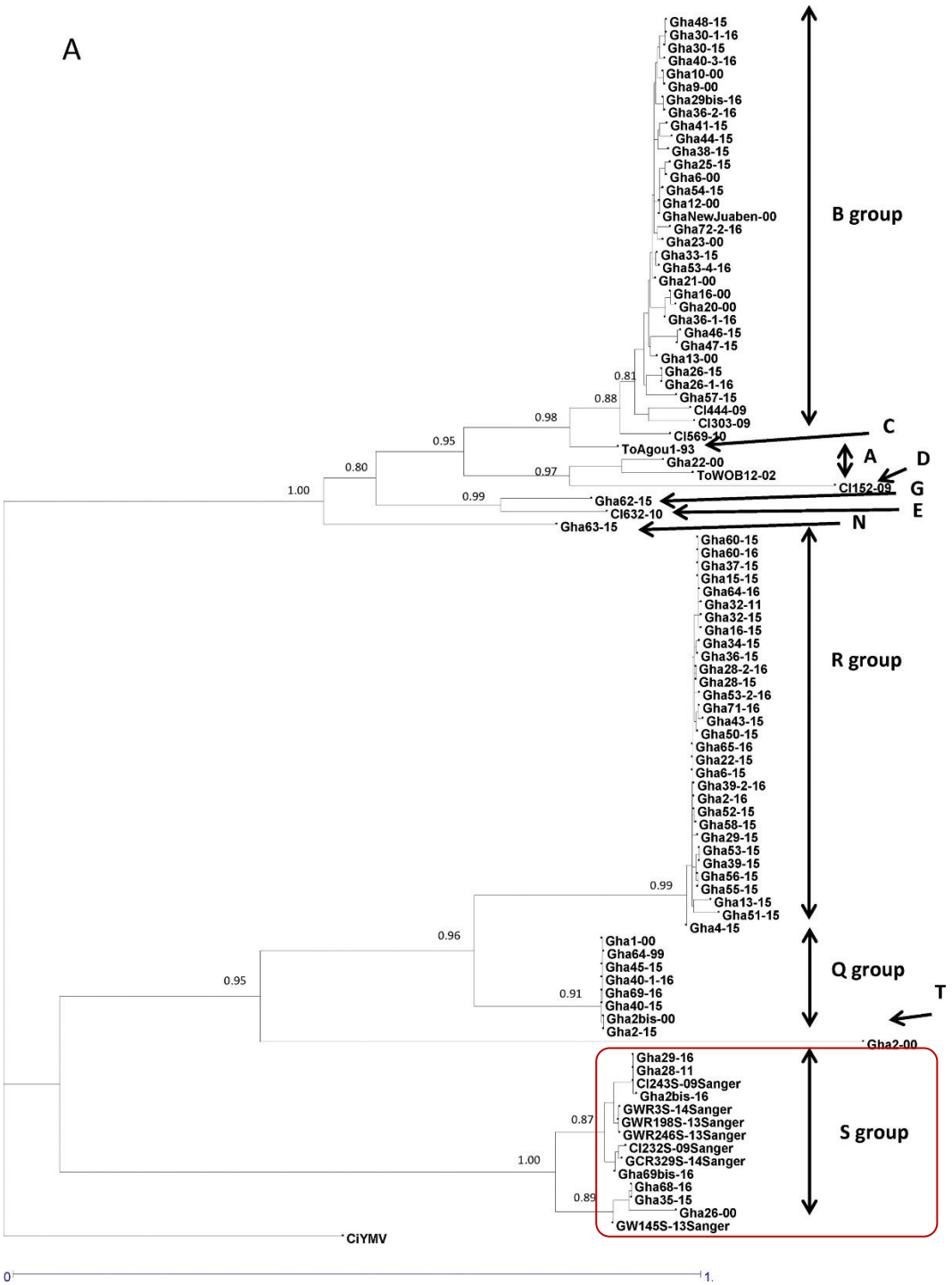


Figure 19: Maximum likelihood phylogenetic tree of CSSV sequences (obtained by Sanger sequencing) based on alignment of the RT/Rnase H region of open reading frame 3 (ORF3) from *T. cacao* trees sampled from multiple locations across Ghana. Numbers on the branches represent the SH-aLRT (approximate likelihood ratio test) branch supports over 0.7. The names of the CSSV groups A, B, C, D, E, F, G, H, J, K, L, M, N and P are indicated. The Citrus yellow mosaic virus sequence (CiYMV) (AF347695) is used as the outgroup. The names of sequences include the abbreviation of the country (CI for Côte d'Ivoire, G or Gha for Ghana, To for Togo), a sampling number along with a region code or a number corresponding to the clone number in brackets, and the year of sampling (1993–2016 coded as 93–16). S-species outgroup of interest highlighted in red. Muller *et al*., 2018.

4.1.5 High Resolution Melt (HRM) as a rapid method for genotyping

The occurrence of EVEs within plant genomes has previously been achieved using a number of approaches. The use of approaches which are dependent on DNA sequencing and analysis of subsequent alignments can be time consuming and prohibitively expensive for plant quarantine facilities and local phytopathology labs that may need to deal with a high throughput of samples.

High-resolution melting, a well-established technique used for genotyping, mutation scanning and sequence matching, was first introduced in 2002 by the University of Utah (UT, USA) and Idaho Technology (UT, USA). HRM works by gradually heating PCR products from around 50 °C to 95 °C, until the melting temperature of the amplicon is reached, and the two DNA strands denature, or 'melt' apart. Fluorescent dyes attach to the amplicon and fluoresce until melting occurs and the dye is released, resulting in a reduction in fluorescence intensity. This reduction in fluorescence intensity is captured in real-time (dF) (Reed *et al*., 2007), with the melting temperature (T_m) calculated by determining the peak of the melting curve first derivative (dF/dT; Figure 20). There is a positive correlation between the T_m and sequence length/GC content. HRM can be used to determine single base pair (bp) differences in sequences, which have an effect on the melting temperature of the amplicon; the $\boldsymbol{\Delta} T_m$ corresponding to a single G-C ↔ A-T change is 0.2 °C. These single bp changes are more difficult to detect in larger (>200 bp) fragments, whereas single bp changes in smaller fragments (<200 bp) can be consistently detected. HRM analysis can also be used in allelic discrimination through the observation of melting curve shape; regions containing different % G+C contents can result in different melt curve shapes, for example monophasic and bi- or multi- phasic (Figure 20; Tong and Giffard, 2012).

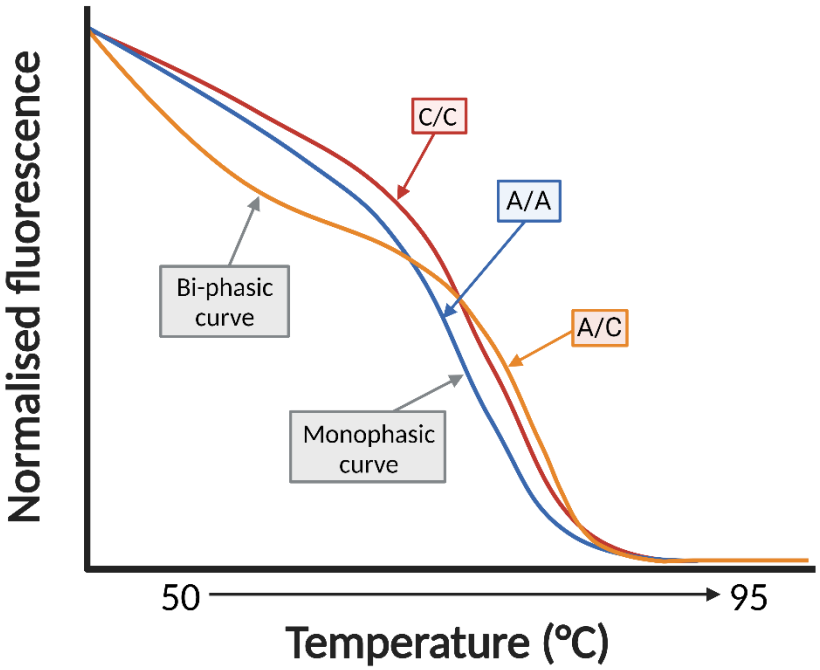


Figure 20: Example curves from HRM analysis. Three curves coloured red, blue and orange with. Each curve corresponds to a illustrates a G+C change present within that sample sequence. Red and blue curves represent a typical monophasic curve shape attributed to a homozygous sample, while orange curve represents a bi-phasic curve shape attributed to a heterozygous sample.

HRM is a low-cost technique that provides a quick and simple way to identify single base pair changes in sequences, and additionally the ability to distinguish between different genotypes. It has previously been used to reveal genotypic variations in both human and plant diseases; for example, in the rapid detection of antibiotic resistance in *Mycobacterium tuberculosis* (Chen *et al.,* 2011) and the genotyping of *Veneturia inaequalis*, the causative agent of apple scab disease (Chatzidimopoulos *et al*., 2019). It has also been used to successfully differentiate between species of the commonly found *T. cacao* pest and vector of CSSV, the mealybug (Wetten *et al.,* 2016).

4.1.6 Aims and objectives

This chapter focusses on work which contributed to Muller *et al*. (2021; Appendix II); a collaboration between colleagues in The French Agricultural Research Centre for International Development (CIRAD; Montpellier, France), the University of Reading (Reading, UK) and the University of the West of England (Bristol, UK). Described are some of the findings that lead to the identification of a novel

badnaviral sequence integrated in the genome of *T. cacao*, and additional research which aims to provide a rapid and cost-effective method to distinguish between insert types, as well as insights into the heritability of these insert types.

4.2 Methods

The work outlined in this chapter aimed to use transmission studies to provide evidence of the integrated characteristic of these Endogenous *T. cacao* Bacilliform Viruses (eTcBVs) and also identify heritability patterns associated with two of the 12 eTcBV insert types described in Muller *et al.* (2021).

4.2.1 Mealybug based transmission of CSSV viral particles and eTcBV elements from *T. cacao* plants

Transmission studies were utilised in order to determine whether the described eTcBV elements could be transmitted from *T. cacao* positive for these EVEs to eTcBV negative *T. cacao* by mealybug feeding; the way in which CSSV transmission occurs. Three different *T. cacao* plants with three different insert types were chosen for investigation (CP, EP and ICS), along with one eTcBV negative *T. cacao* plant (EN; Table 4).

Table 4: List of *T. cacao* used in transmission studies.

Sample code (plant number)	**Plant name[1]**	**Insert type[2]**	**Number of plants**
Plants used in CSSV viral particle transmission tests			
G20	G20	Negative, CSSV positive	1
G10	G10	Negative, CSSV positive	1
G44	G44	Negative, CSSV positive	1
G18	G18	Negative, CSSV positive	1
AM(1, 2, 5, 6 & 7)	RB1	Negative, CSSV-free	5
Plants used in eTcBV element transmission tests			
EN (1-6)	EET19 self	Negative, CSSV-free	6
EP (1-2)	EET183 self	VI, CSSV-free	2
ICS (1-2)	ICS10 x ICS40[3]	IV, CSSV-free	2

[1] Plant name given by the ICQC, Reading, from which the plants were obtained.
[2] Insert type refers to the endogenous viral element characterised by Muller *et al*., 2020. Of the 12 insert types described, two were chosen for analysis here.
[3] ICS10 X ICS40 is a cross of two plants both containing the type IV fragment.

4.2.1.1 *P. citri* population maintenance

A population of CSSV-free *P. citri* was established using adult females sourced from the University of Reading tropical plant glasshouse and characterised to species using HRM analysis (Wetten *et al.,* 2016). These were maintained on sprouting potatoes (cv Charlotte) in sealed boxes in the dark at 24 ± 2 °C. Boxes were sealed with a layer of Vaseline to ensure younger larvae were unable to escape and opened weekly to allow air flow. Egg sacks laid by female adult mealybugs were transferred to a clean box with fresh sprouting potatoes every month to sustain the population.

4.2.1.2 *P. citri* trapping, feeding and relocation

For CSSV transmission studies, 2nd – 3rd instar larvae (Figure 21) were picked from the sprouting potatoes using a dissecting microscope and damp paintbrush. Feeding mealybugs were tapped gently on their dorsal surface at least 30 s prior to removal to ensure stylet withdrawal, reducing the risk of stylet damage which could impede feeding ability. Feeding traps (Figure 22 and Figure 23) are designed so that they are secure enough that *P. citri* are unable to escape, while keeping leaf damage to a minimum. Ten, mobile *P. citri* larvae were placed in a single trap to feed. Once transferred to a trap, mealybugs were left to feed on the abaxial surface of fully-expanded *T. cacao* true leaf for three days before being transferred to a trap on a new *T. cacao* plant or destroyed.

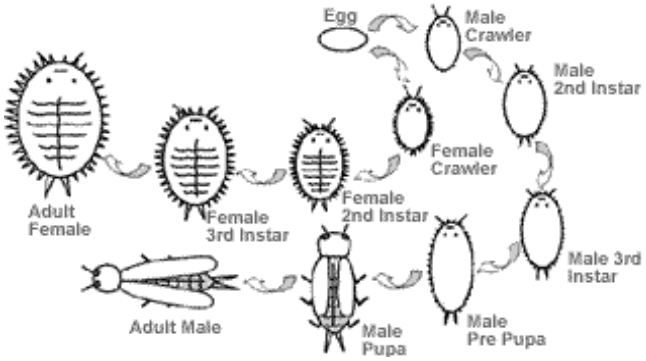


Figure 21: The lifecycle of Grapevine mealybug. Image sourced from Annecke and Moran, 1982.

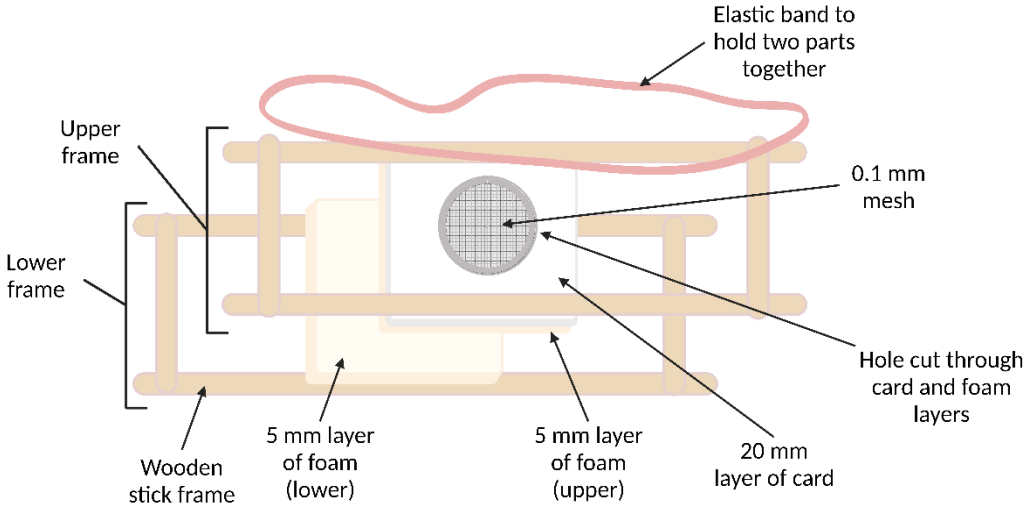


Figure 22: Mealybug feeding trap. Two frames with foam layers are attached to each side of the leaf surface using an elastic band. A 2 cm diameter hole in the layers of card and foam on the upper frame provides a space for mealybugs to feed. The hole is covered by a layer of 0.1 mm mesh, sufficiently fine to prevent stage one juvenile mealybugs from escaping.

Figure 23: An insect feeding trap placed on about the midrib of a *T. cacao* leaf lamina. *P. citri* juveniles are placed in the trap and allowed to feed on the underside of the leaf.

P. citri were first placed on either a CSSV-positive or eTcBV-positive seedling to feed, before being relocated and allowed to feed on a CSSV-negative, eTcBV negative seedling (Table 5). The final *P. citri* feed lasted three days, with the trap removed on day three.

Table 5: *P. citri* feeding details for transmission test for eTcBVs.

Transmission (plant code → plant code)	**P. citri first feed**			**P. citri second feed**		
	'Donor' Plant	Insert type (CSSV +/-)	Plant code	'Recipient' Plant	Insert type (CSSV +/-)	'Recipient' Plant code
Plants used in CSSV viral particle transmission						
G20 → AM7	G20	- (+)	G20(2)	RB1	- (-)	AM7
G10 → AM6	G10	- (+)	G10	RB1	- (-)	AM6
G44 → AM5	G44	- (+)	G44	RB1	- (-)	AM5
G20 → AM2	G20	- (+)	G20	RB1	- (-)	AM2
G18 → AM1	G18	- (+)	G18	RB1	- (-)	AM1
Plants used in eTcBV element transmission						
EP1 → EN1	EET183 self	VI (-)	EP1	EET19 self	- (-)	EN1
EP2 → EN2	EET183 self	VI (-)	EP2	EET19 self	- (-)	EN2
ICS1 → EN3	ICS10 x ICS40	IV (-)	ICS1	EET19 self	- (-)	EN3
ICS2 → EN4	ICS10 x ICS40	IV (-)	ICS2	EET19 self	- (-)	EN4

4.2.2 Cross pollination of ICQC *T. cacao* for generation of eTcBV positive/negative *T. cacao* crosses

Cacao pods from crosses of various clones of badnavirus-free *T. cacao*, both positive and negative for eTcBV, were produced at the ICQC for this experiment. A full list of the crosses obtained can be seen in Table 6. Hand pollination was performed by the ICQC staff, using a paintbrush to collect pollen from the stamen of one *T. cacao* flower and brush it onto the stigma of another previously emasculated. Blossoms were marked with a paper tag indicating the identity of the cross immediately after pollination.

4.2.2.1 Cacao pod processing and seed planting

For those crosses that were successful, resultant cocoa pods were mature five to seven months following pollination and these were harvested and transported to UWE Bristol. Each pod was labelled with the cross type and the date of collection. The number of seeds obtained varied with pod size but ranged from >15 seeds to <50 seeds per pod. An opening in the pod was created using a knife and pods were cracked open to reveal seeds covered in mucilage (Figure 24). The mucilage was removed, and seeds were planted in John Innes No. 2 compost in segmented seedling trays holding up to 40 seedlings per tray. Planted seeds were covered in a thin layer of vermiculite and kept in the UWE Envirotron in a contained compartment free from CSSV infected material. Seedlings were kept at ≥25 °C in 80% RH, watered for 10 minutes twice a day through an automated irrigation system. Seedlings and trays were given a code corresponding to the seeds which were planted in them (Table 6).

Figure 24: Opening *T. cacao* pods to remove seeds. White mucilage on the outside of the seeds is removed prior to planting.

Table 6: *T. cacao* crosses generated at the ICQC, Reading for this experiment.

Sample code[1]	Cross	Paternal donor	Paternal insert type[2]	Maternal donor	Maternal insert type
A, B, C, D, E, F, G, W, X, Y & Z	EET183 X IMC60	EET183	VI	IMC60	V
H, I, J	ICS40 X IMC60	ICS40	IV	IMC60	V
S, T, U, V	IMC60 X ICS40	IMC60	V	ICS40	IV
N, O, P, Q, AE, AF, AJ	EET183 X ICS40	EET183	VI	ICS40	IV
AI	IMC60 X EET183	IMC60	V	EET183	VI
AB	ICS40 X EET183	ICS40	IV	EET183	VI

[1] Sample code assigned at the time of seed planting.
[2] Insert type refers to the endogenous viral element characterised by Muller *et al*., 2021. Of the 12 insert types described, the crosses here involve three versions: IV, V and VI.

4.2.3 Sample collection, DNA extraction, EP-PCR and gel-electrophoresis

Once plants were > one-month post *P. citri* inoculation, or seedling true leaves had matured (> 3 months post planting), or samples were taken and processed by DNA extraction as described in 2.3.1. DNA extracts underwent either EP-PCR and gel electrophoresis (2.3.2 and 2.3.3) using Badnavirus specific primers (section 2.3, Table 2; Badna1Deg2 and Badna4Deg2) for identification of samples positive for eTcBV inserts and CSSV, or qPCR using RTase specific primers (section 2.3, Table 2, RTaseNf, RTaseNR and RTaseNP) for detection of CSSV.

4.2.4 PCR product clean-up, Sanger sequencing and sequencing analysis

Clean-up of PCR products and subsequent sequencing were performed as described in sections 2.3.5 and 2.3.6. Analysis, including trimming and multiple alignment of sequencing products was performed using Geneious Prime. A Maximum Likelihood phylogenetic tree using the Tamura-Nei model was built using MEGA11.

4.2.5 High Resolution Melt (HRM) for the identification of three different eTcBV inserts in *T. cacao* DNA

In order to find an appropriate primer combination to characterise and distinguish between different eTcBV insert types, three forward primers and five reverse primers (Table 7) were designed using Primer3. Combinations of the different forward and reverse primers were tested against a range of cocoa clones containing the putative insert types and resultant PCR products were sequenced in order to determine the ideal pair for detection of eTcBV insert types (Table 9).

A 2 µM primer working stock was made using 10 µL of forward primer (at 100 µM μL^{-1}), 10 µL of reverse primer (at 100 µM μL^{-1}) and 480 µL of MGW. Typical master mix for a 20 µL reaction included 10 µL of SensiFAST HRM (Meridian Bioscience; Scientific Laboratory Supplies Limited, UK), 2 µL of primer working stock, 6 µL MGW and 2 µL DNA. Following a series of temperature optimisation trials, reactions were run in a Rotor-Gene Q 2-plex (QIAGEN, UK), cycling as follows: 95 °C for 10 minutes; [95 °C for 10 seconds, 56 °C for 15 seconds, 72 °C for 20 seconds] x40; 94 °C for five minutes. HRM data was collected from 61 °C to 83 °C, rising by 0.1 °C each step. Synthetic reference sequences were designed using samples of seedlings from crossed *T. cacao* plants (table #) and used as positive controls in HRM experiments. A hemizygous melt profile was represented using 2 µL of oligos 2 and 4 at 10 µM, while a heterozygous melt profile was represented by using a 1:1 mix of the two oligos, with each at 5 µM. A no-template control of MGW was also used in every run.

Table 7: Primers designed to detect the three eTcBV insert types in *T. cacao* clones and seedlings.

Primer number	Primer name	Sequence (5'-3')	Tm (°C)
1	F8-28	AGTGTTCCAAGCTTTGGGAT	57.3
2	F54-73	CCAAGCCAAGATCGCAGTCC	63.5
3	F59-78	CCAAGATCGCAGTCCTTTCT	59.4
4	R64-83	ACAAAGAAAGGACTGCGATC	55.3
5	R94-113	GATTGCAGACTTCCCTGATG	57.3
6	R166-185	ATCAGAAATTGAATTCCTTG	53.2
7	R211-230	AGACAACGGCTTGATCTTAT	55.3
8	R235-254	CCAACAAATGTTGTCCATCT	53.2

Table 8: Synthetic oligos designed for use as reference sequences in HRM.

Synthetic reference oligo	Reference origin sample	eTcBV insertion type[1]	Oligo sequence (5' – 3')
2	U23	IV	GAGTGTTCCAAGCTTTGGGACGTACGCTCGAGAGTAATTGAGAATCCCGAGCCATGATCGCAGTCCT TTCTTTGT
4	B24	VI	AAGTGTTCCAAGCTTTGGGACAAAGGCTCGGGAATAATTGAGTACTCCAAGCCAAGATCGCAGTCCT TTCTTTGT

[1] Based on phylogenetic tree analysis (Figure 31)

Table 9: HRM primer combinations tested to detect the three eTcBV insert types in *T. cacao* clones and seedlings.

Primer combination name	Forward Primer	Reverse Primer	Expected fragment size (bp)
1/4	(1) F8-28	(4) R64-83	75
1/5	(1) F8-28	(5) R94-113	105
1/6	(1) F8-28	(6) R166-185	177
1/7	(1) F8-28	(7) R211-230	222
1/8	(1) F8-28	(8) R235-254	246
2/5	(2) F54-73	(5) R94-113	59
2/6	(2) F54-73	(6) R166-185	131
2/7	(2) F54-73	(7) R211-230	176
2/8	(2) F54-73	(8) R235-254	200
3/5	(3) F59-78	(5) R94-113	54
3/6	(3) F59-78	(6) R166-185	126
3/7	(3) F59-78	(7) R211-230	171
3/8	(3) F59-78	(8) R235-254	195

4.2.6 Data analysis

All data analysis was performed using Prism GraphPad 9 (GraphPad Software, San Diego, California USA, www.graphpad.com).

4.3 Results

4.3.1 Identification of ICQC *T. cacao* clones positive for badnaviral sequence

During routine screening of putatively badnavirus-free ICQC *T. cacao* clones in order to provide negative control material for the present study, analysis using badnavirus specific EP-PCR and gel electrophoresis revealed high proportions of the plant collection that generated a positive PCR (Figure 25). The locations of all *T. cacao* germplasm in ICQC (Reading, UK) polytunnels can be seen in Figure 26. Germplasm is dispersed between four separated polytunnels which are routinely checked for disease and pest symptoms. No symptoms of typical badnavirus infection were identified on any of these plants during six monthly inspections over the past eight years. Germplasm identified as PCR positive for a Badnaviral sequences are highlighted in orange in Figure 27. The physical location of these PCR positive plants in the ICQC tunnels was not consistent with a case of insect vectored disease movement or the 'crowd disease' nature of CSSVD. In some cases, the PCR positive germplasm can be found located next to PCR negative germplasm (Figure 27). This suggests that these positive PCR products could be the result of an integrated badnaviral sequence in the genome of some *T. cacao* clones.

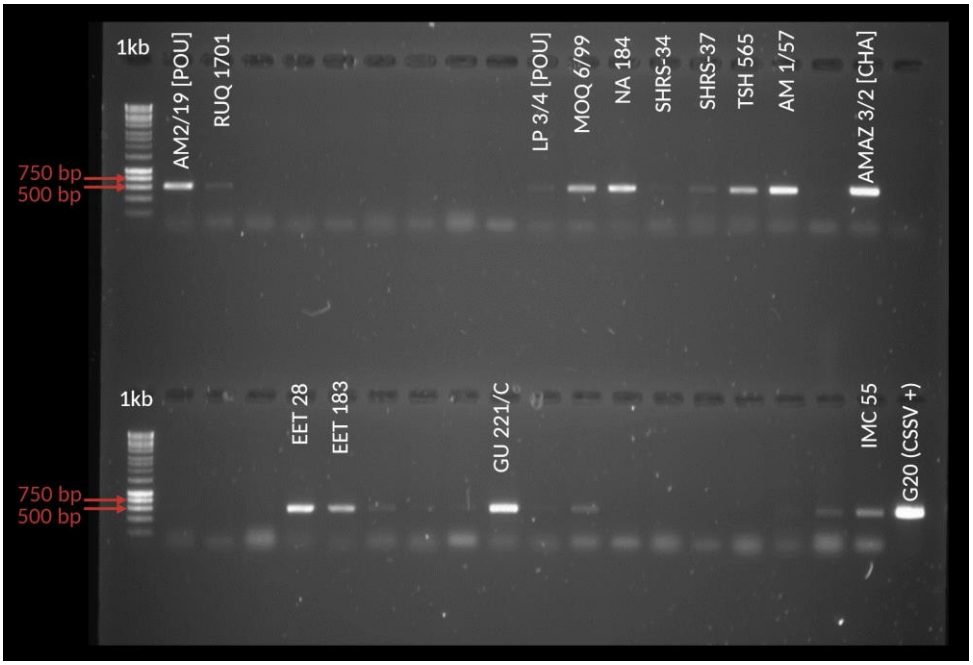


Figure 25: Samples of asymptomatic *T. cacao* clones from the ICQC (Reading, UK). Badnavirus specific primers were used in order to obtain positive PCR products. A product of 628 bp is expected in positive samples. Sample G20 represents a symptomatic, CSSV positive (strain New Juaben) *T. cacao* plant kept in the UWE Envirotron.

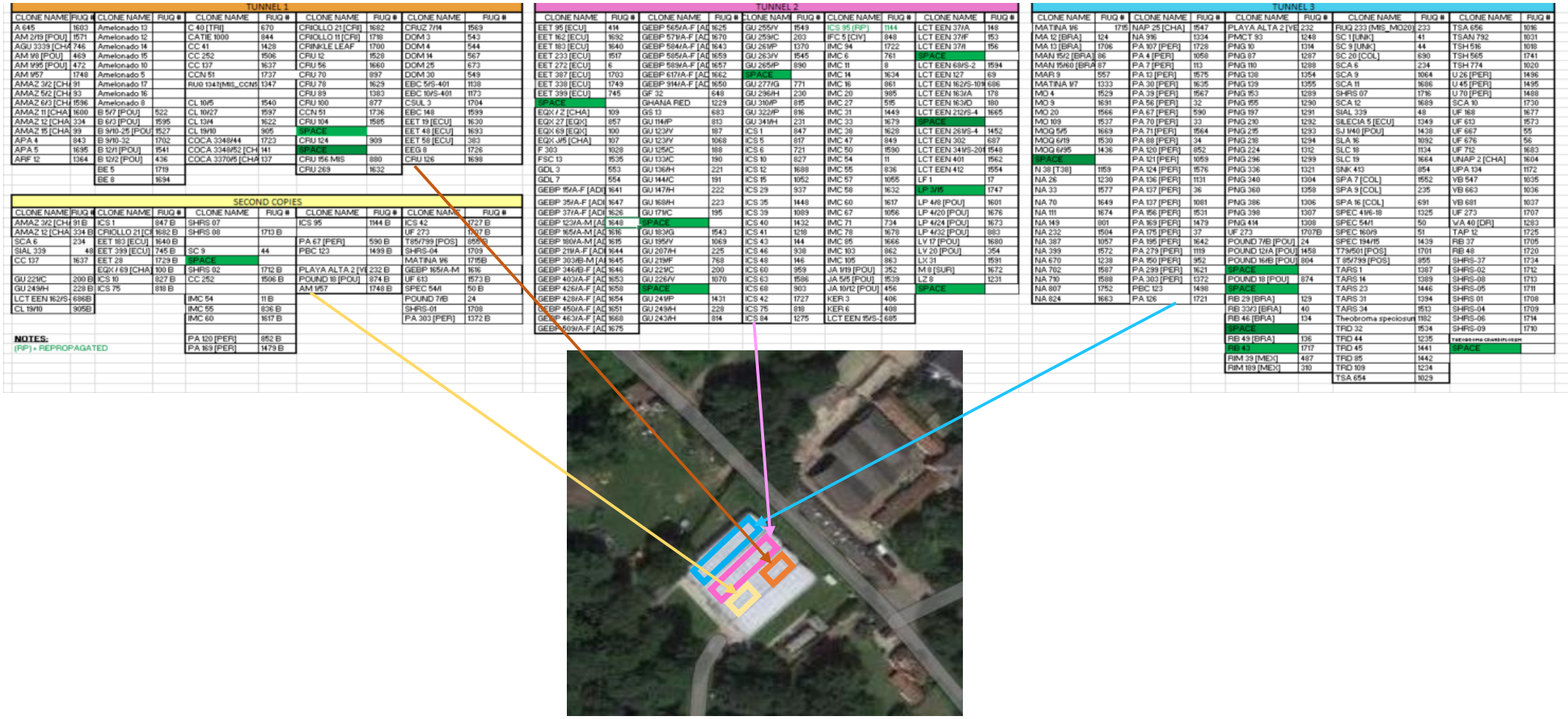


Figure 26: Aerial perspective satellite image of the ICQC (Reading, UK) polytunnels containing asymptomatic *T. cacao* germplasm. Full inventory of *T. cacao* arranged by location listed above, with polytunnel location highlighted by either blue, pink, yellow or orange.

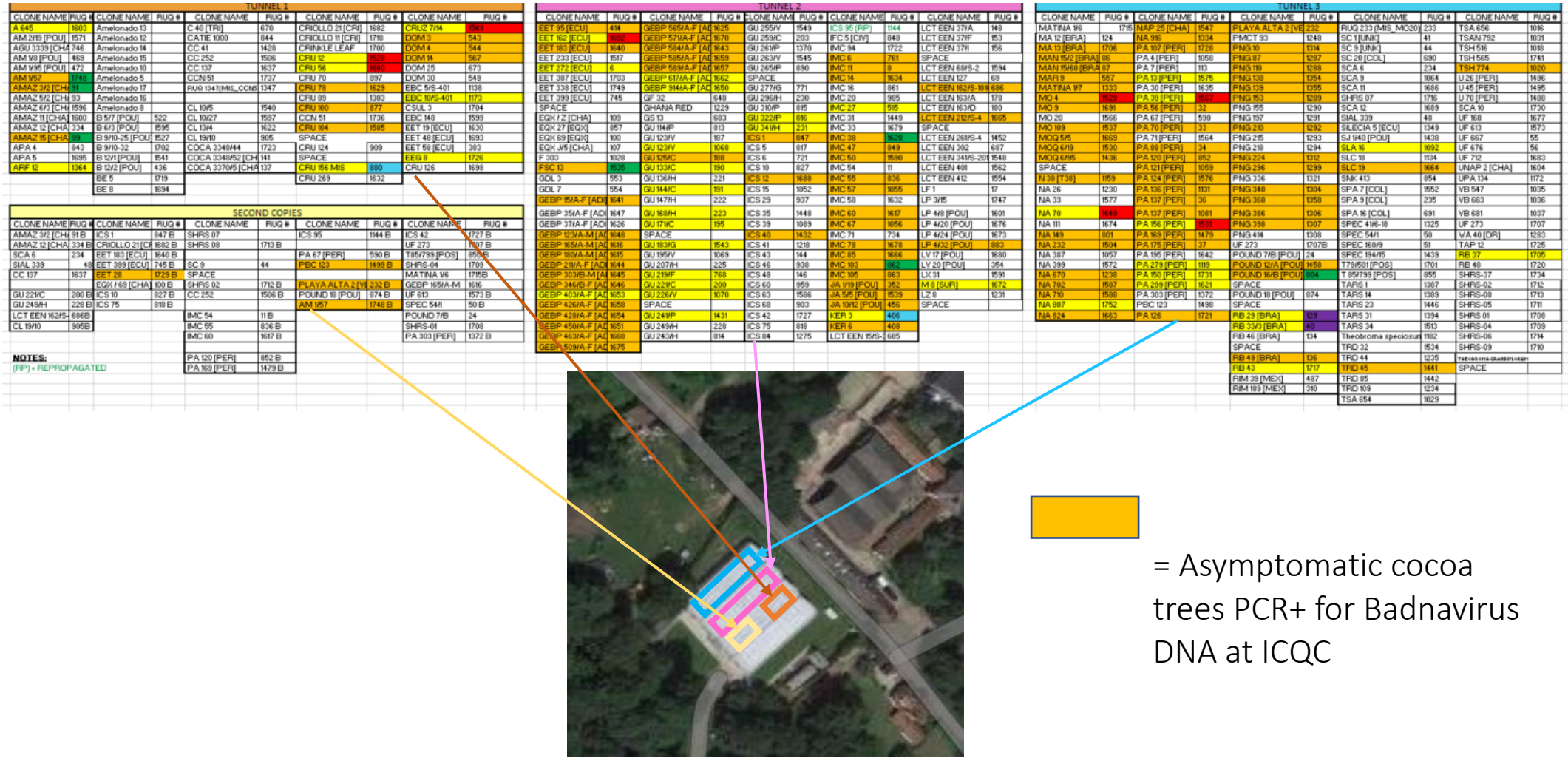


Figure 27: Aerial perspective satellite image of the ICQC (Reading, UK) polytunnels containing asymptomatic *T. cacao* germplasm with *T. cacao* which was identified as PCR positive for Badnavirus highlighted in orange. Full inventory of *T. cacao* arranged by location listed above, with polytunnel location highlighted by either blue, pink, yellow or orange.

4.3.2 Could the putative integrated sequences belong to viral particles, transmissible by *P. citri*?

In order to determine whether these putative integrated sequences could actually belong to viral particles, representative clones were investigated using 'inoculation' via *P. citri* feeding and trapping was performed.

4.3.2.1 Transmission of CSSV particles from high viral load *T. cacao* to CSSV-free *T. cacao* by *P. citri* trapping and feeding

Firstly, *P. citri* free from CSSV were fed on CSSV positive G plants (Table 1) and transferred to CSSV-free *T. cacao* seedlings. The concentration of CSSV particles in samples of these plants was determined by qPCR. While the concentrations of CSSV particles were significantly different between the group of plants inoculated > 9-years ago compared to those inoculated < 1-month ago (at the time of sampling), CSSV was still detected in all ten samples, indicating that this method of transmission of CSSV particles was successful.

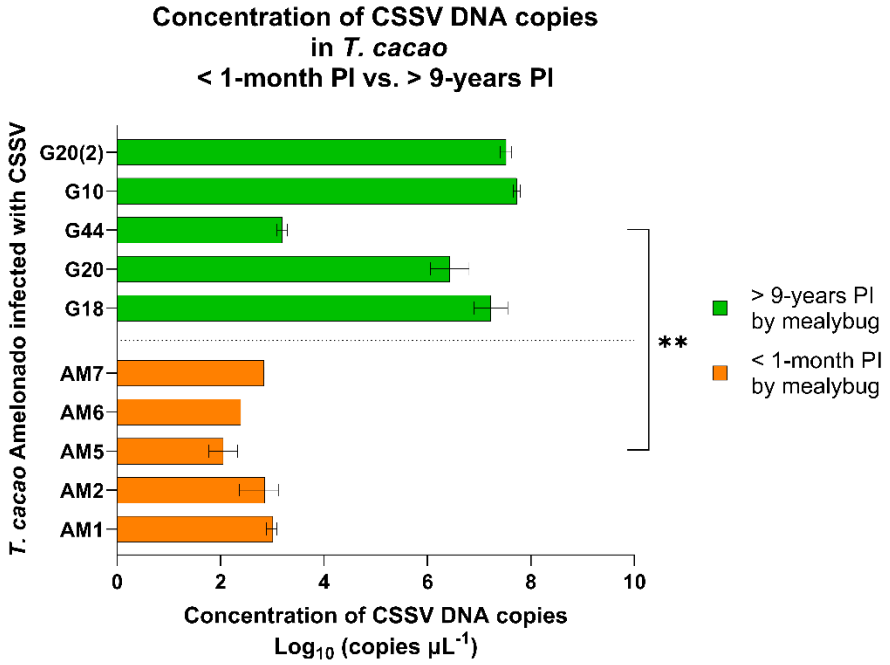


Figure 28: Concentration of CSSV DNA copies in *T. cacao* 2-weeks post inoculation vs. 9-years post inoculation, determined by qPCR. Plotted is the mean with error bars showing 95% confidence interval. Significance is illustrated in GraphPad style: ** p<.01 (nested t-test comparing > 9-years PI and < 1-month PI). n = 3.

Following this, the same procedure was used to determine whether the eTcBV could be transmitted by *P. citri*. *P. citri* were fed on eTcBV positive plants for three days before being moved onto an eTcBV negative plant to feed (Table 4). EP-PCR using Badnavirus specific primers (Table 2), followed by gel electrophoresis, was used to determine whether the eTcBV could be identified in eTcBV negative plants post mealybug feeding. Gel results (Figure 29) indicate that no eTcBV was detectable in samples EN 1, 2, 3 or 4, indicating that the eTcBV was not transmitted by *P. citri* feeding.

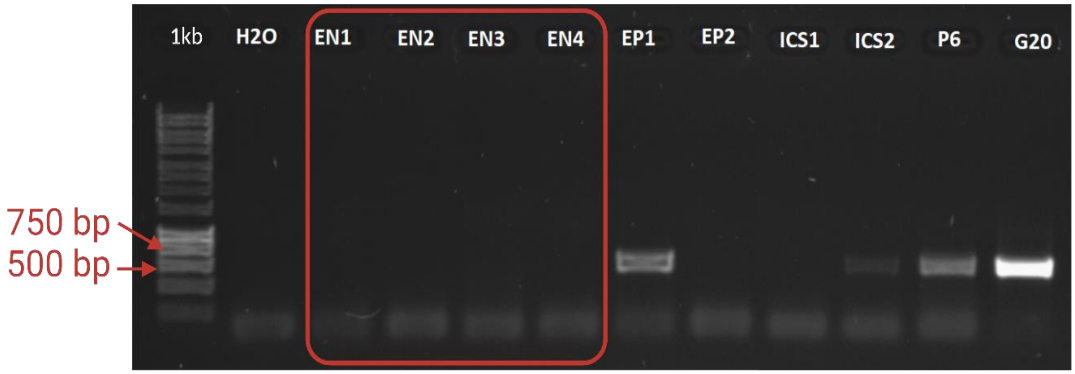


Figure 29: Samples of F1 generation *T. cacao* eTcBV positive and negative crosses post-mealybug transmission studies (1-month post mealybug feed). Bands represent presence of a 628 bp fragment amplified by badnavirus specific primers. Negative samples highlighted in red (EN1-4) are all eTcBV negative *T. cacao* plants challenged with mealybugs previously fed on eTcBV positive *T. cacao* plans (EP1 and ICS2). Samples P6 and G20 represent eTcBV and CSSV (strain New Juaben) positives respectively.

4.3.3 Presence of Badnaviral sequences in F1 generations of eTcBV hybrids

All F_1 generation seedlings underwent EP-PCR using badnavirus specific primers and gel electrophoresis to determine presence or absence of an eTcBV. A representative gel image is shown in Figure 30, where 44 PCR products from the F_1 generation EET183 x ICS60 were examined. Of the 44 samples run on this gel, bands in the 628 bp region can be seen in 16 samples, indicating a positive result. However, only three of these bands are high intensity. Only samples with high intensity bands were chosen for sequencing. Of all seedlings tested, intense bands in the 628 bp region were detected in 31 PCR products; these were cleaned and sent for Sanger sequencing. Sequences from 19 samples were successful and

underwent trimming, multiple alignment and phylogenetic analysis (Figure 31). Additional PCR products from the three parent plants (EET183, ICS40 and IMC60) were also sent for sequencing; successful sequences for ICS40 and IMC60 were obtained and used in multiple alignment and phylogenetic analysis.

Phylogenetic analysis revealed that samples fall into three distinct clades (Figure 31). The addition of two parent samples (ICS40 and IMC60, Table 10; highlighted in red and orange respectively, figure #) allowed for eTcBV types to be designated to each clade. Two IMC60 x ICS40 seedlings were included (U23 and V2, table #), which fell within the clade assigned to type IV inserts, a characteristic of the parent plant ICS40. Similarly, all four samples from EET183 x ICS40 seedlings (N15, O20, P23 and Q8) fell within the clade assigned to type IV inserts. Clean sequences were obtained for 13 seedlings of the cross EET183 x IMC60, which segregated into two clusters. One cluster aligned with the type V insert characteristic of IMC60, and the other diverging, with a similar relatedness to the type IV cluster, which was designated type VI, a characteristic of parent plant EET183.

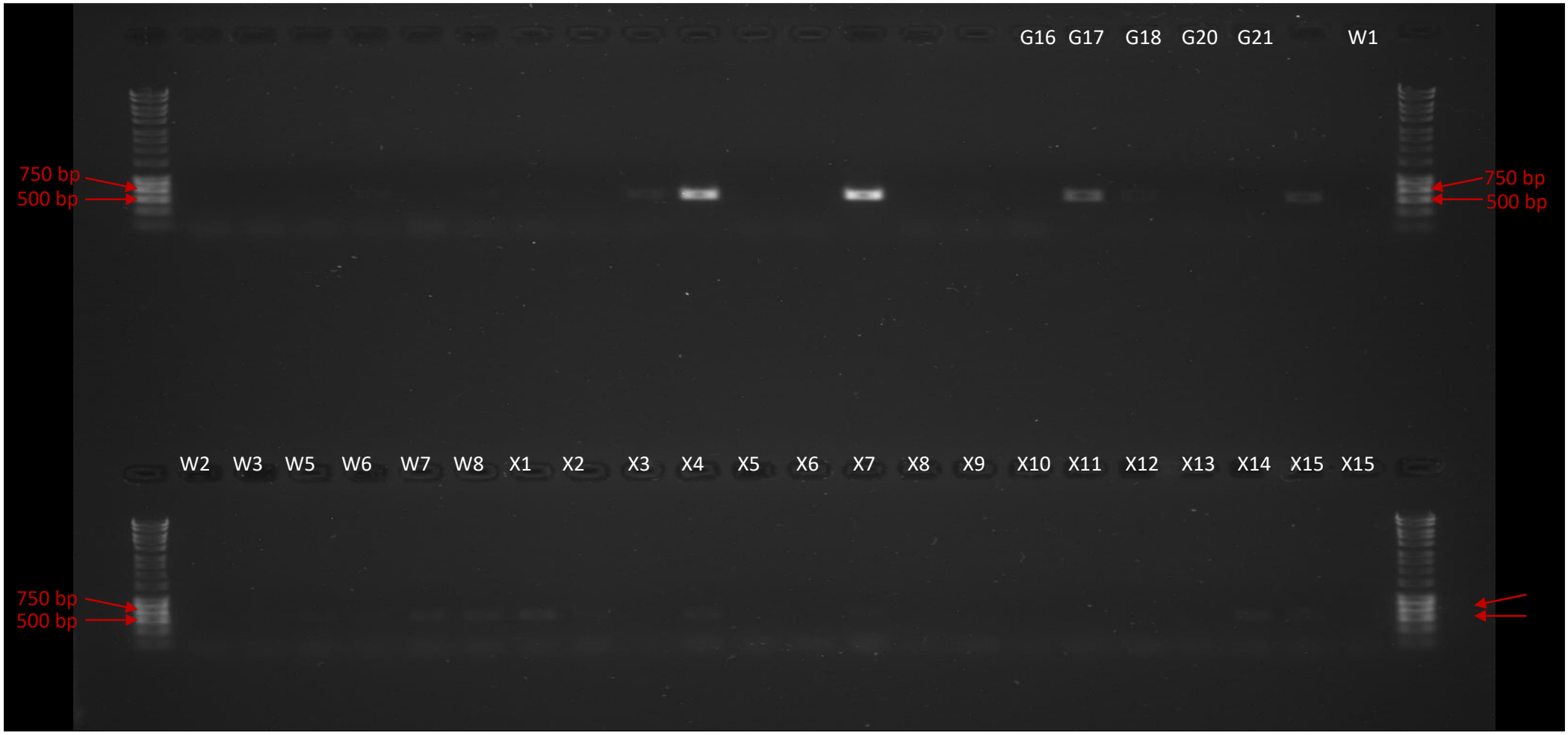


Figure 30: A representative gel image showing PCR products from EP-PCR with Badnavirus specific primers. Each sample represents one DNA extraction from an EET183 x IMC60 seedling. Wells on either end of the samples represent a 1kb ladder with 500 and 750 bp labelled in red.

Table 10: RT RNaseH (primer pair Badna1Deg2/Badna4Deg2) PCR-positive samples submitted for Sanger sequencing. eTcBV insertion type assigned through phylogenetic analysis (Figure 31).

PCR positive sample	Parent plants (paternal x maternal)	Parent insertion types[1] (paternal x maternal)	eTcBV insertion type (based on phylogenetic groupings)
F_1 generation			
A5	EET183 x IMC60	VI x V	VI
A6	EET183 x IMC60	VI x V	VI
A19	EET183 x IMC60	VI x V	V
AI3	IMC60 x EET183	V x VI	VI
B3	EET183 x IMC60	VI x V	VI
B24	EET183 x IMC60	VI x V	VI
D14	EET183 x IMC60	VI x V	V
E7	EET183 x IMC60	VI x V	VI
F8	EET183 x IMC60	VI x V	V
G9	EET183 x IMC60	VI x V	V
G13	EET183 x IMC60	VI x V	V
X1	EET183 x IMC60	VI x V	VI
Y9	EET183 x IMC60	VI x V	VI
N15	EET183 x ICS40	VI x IV	IV
O20	EET183 x ICS40	VI x IV	IV
P23	EET183 x ICS40	VI x IV	IV
Q8	EET183 x ICS40	VI x IV	IV
U23	IMC60 x ICS40	V x IV	IV
V2	IMC60 x ICS40	V x IV	IV
Parent plants			
IMC60	IMC60	V	V
ICS40	ICS40	IV	IV
EET183[2]	EET183	VI	VI

[1] Insertion types derived from results of Muller *et al*. (2021).

[2] EET183 sequence obtained from Muller *et al*. (2021; accession number MW009806).

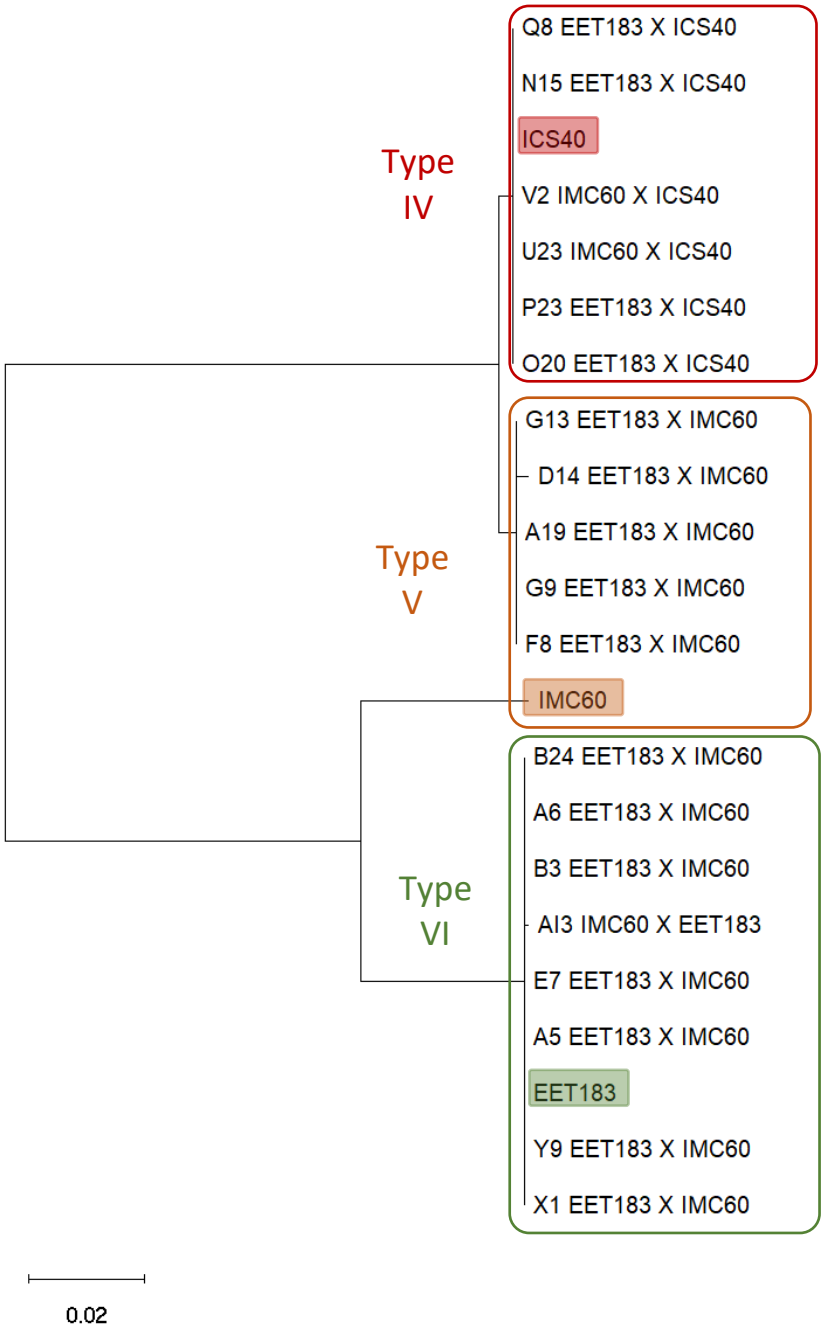


Figure 31: Maximum Likelihood phylogenetic tree using the Tamura-Nei model based on the alignment of the RT RNase H region of ORF3. Sequences obtained from the F_1 generations from crosses of three *T. cacao* clones, each containing a different eTcBV insert type as characterised in Muller *et al.* (2021). The three insert types observed are highlighted in red (type IV), green (type VI) and orange (type V), with parents highlighted in red, orange and green as per their insert type. EET183 sequence obtained from Muller *et al*. (2021; accession number MW009806).

4.3.4 HRM method to model inheritance patterns of eTcBV insert types across different F_1 generations of eTcBV crosses.

A HRM method was developed in order to distinguish between insert types and different zygosities of F_1 generations of eTcBV hybrids. Using this method, the mode of inheritance of these eTcBVs can be assessed.

4.3.4.1 Testing of different primer pair combinations for HRM analysis of F_1 generation eTcBV positive crosses and characterising different zygosities based on melt profile

Primers for use in HRM (Table 7) were designed following multiple alignment of RT RNase H eTcBV F_1 generation sequences obtained using Badna1Deg2 and Badna4Deg2 primers (Table 2). Three forward primers and five reverse primers were designed using conserved areas of the sequences of all of the F_1 generations (Figure 31).

Primer pairs outlined in Table 9 were tested for specificity using two samples, G9 and AI3 (Table 10). The gel-image results show that all primer pairs with the exception of 2/7 and 3/7 resulted in single bands aligning with the expected bp for that primer set when compared to the ladder (Figure 32).

Following gel analysis, each primer pair underwent HRM analysis using the same two samples (G9 and AI3), in order to determine which primer pairs produce unique melt profiles allowing for discrimination between different eTcBV insert types. HRM analysis showed that all primer pairs except 2/5 and 3/5 produced a positive and distinguishable HRM melt profile when comparing the two different samples. Based on results from Figure 33 & Figure 34, primer pair combination 1/4 was chosen for use in all further HRM experiments.

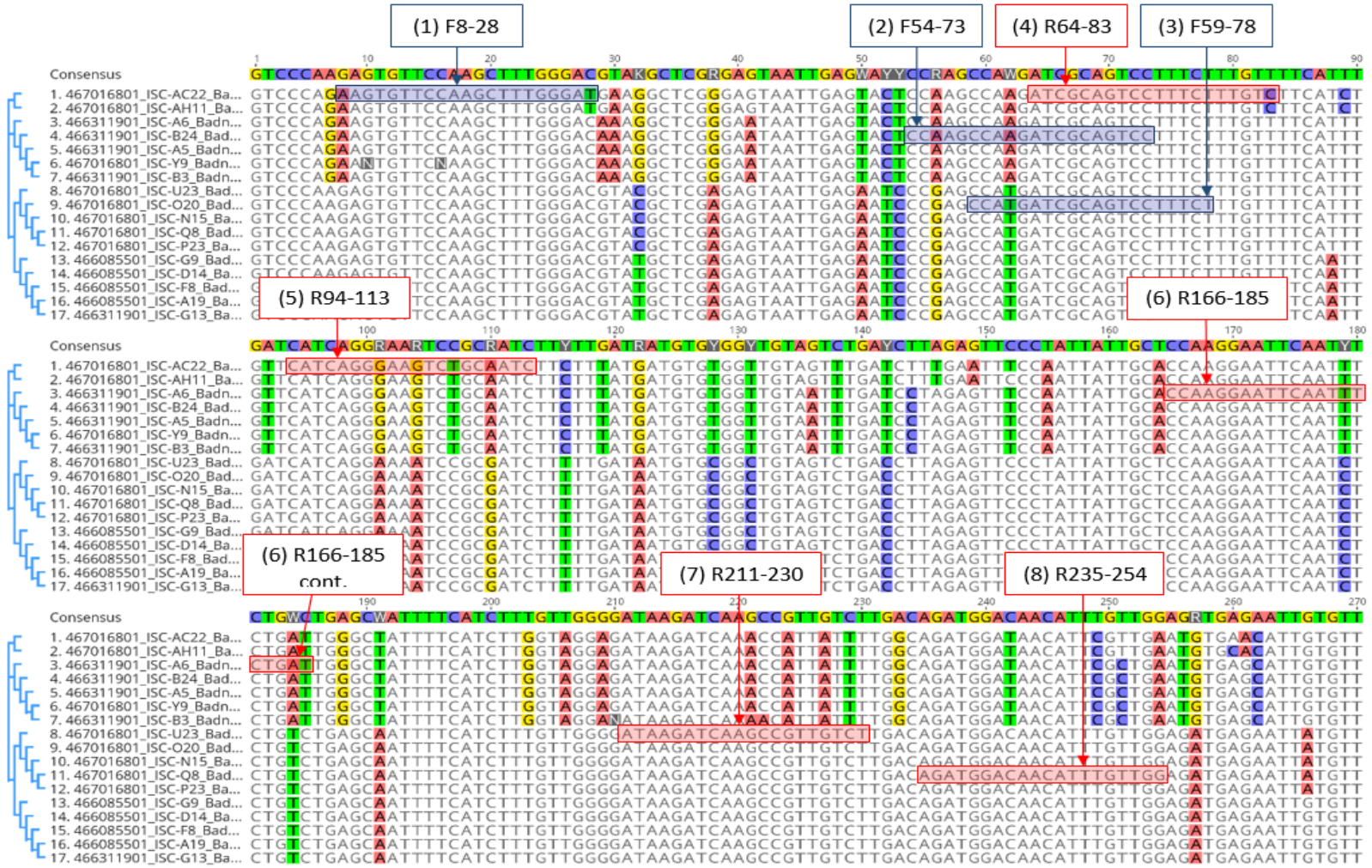


Figure 32: Multiple alignment of eTcBV sequences from 17 *T. cacao* progeny derived from F1 generation crosses showing primers designed for subsequent testing. Primers were designed using most conserved regions of sequences. Forward primers are highlighted in blue with reverse primers highlighted in red.

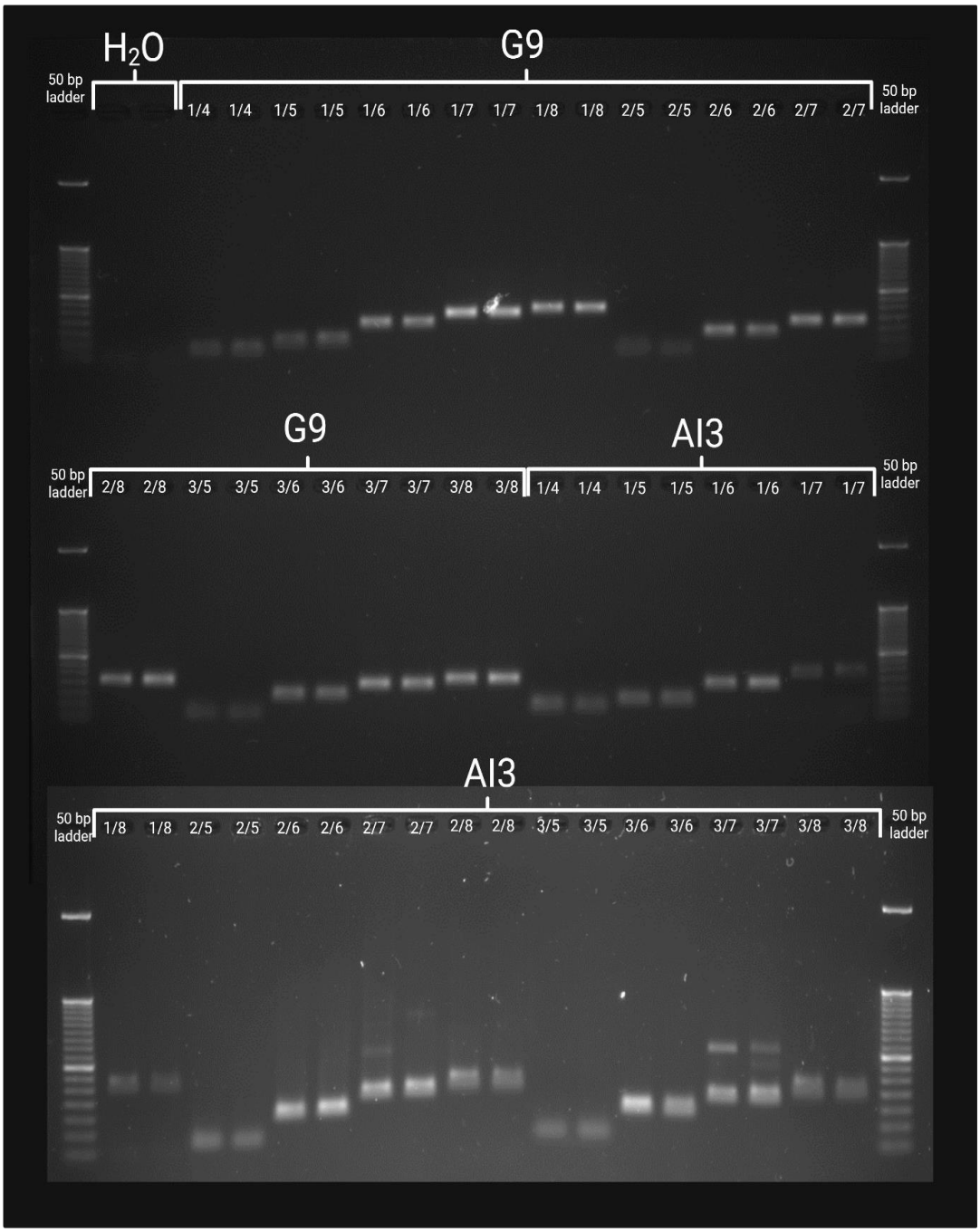


Figure 33: Samples from two F1 eTcBV positive *T. cacao* crosses, G9 (insert type V) and AI3 (insert type VI); 13 different primer pair combinations (Table 9) were used to amplify the eTcBV insert. Gel electrophoresis was used to determine the optimal primer combinations to use in future HRM experiments. A 50 kb ladder was used to confirm fragment size. Bands can be seen in expected regions for all primer pair combinations with the exception of 2/7 and 3/7, where multiple bands can be detected; these two combinations were excluded from future HRM experiments.

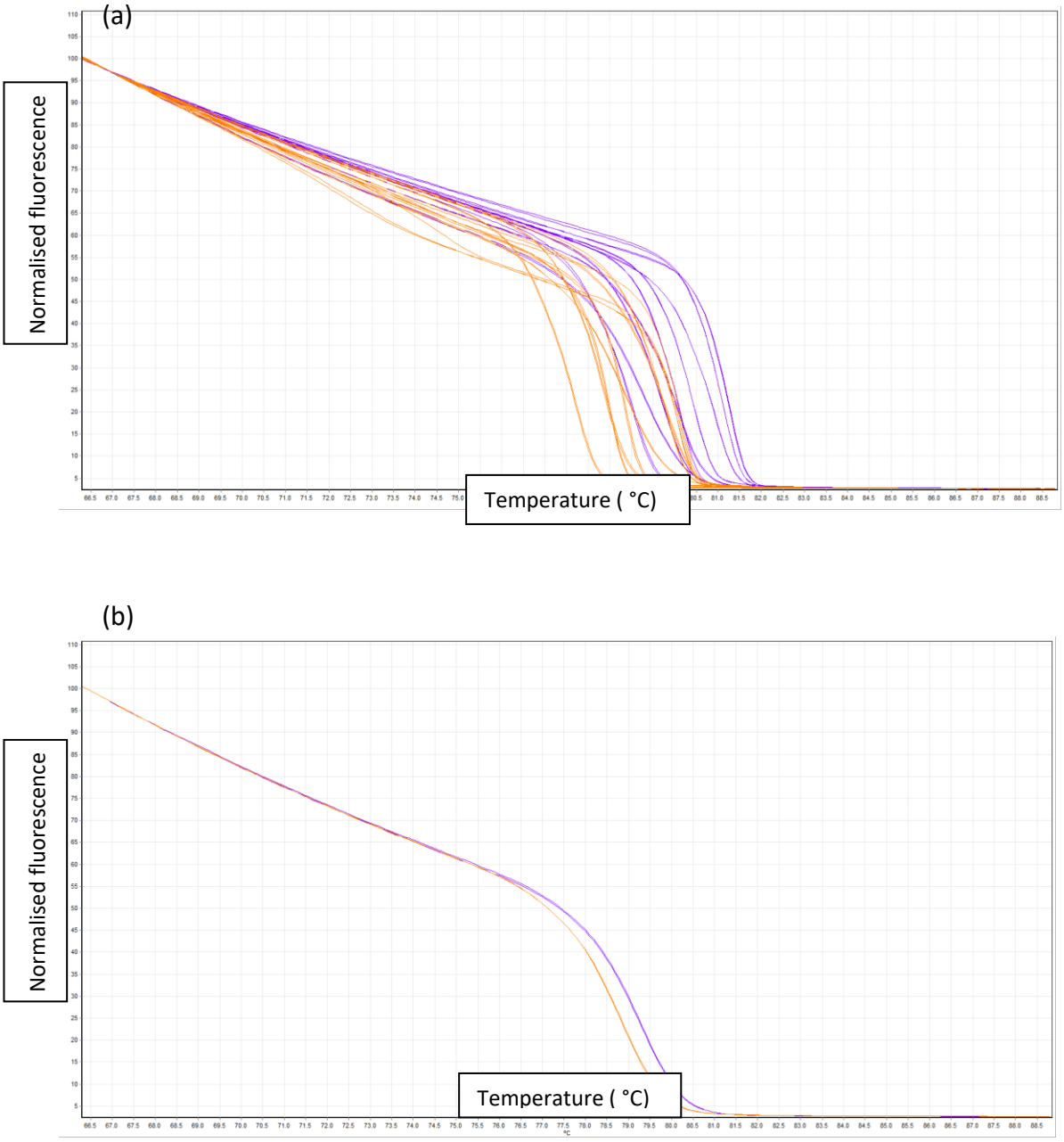


Figure 34: HRM analysis showing normalised fluorescence two samples of *T. cacao* positive for an eTcBV (G9, insert type V, purple; AI3, insert type VI, orange) using 11 different primer pair combinations (Table 9; primer combinations 2/7 and 3/7 were excluded following results from Figure 33) (a), and focussing on primer pair 1/4 (b) which was chosen for use in subsequent HRM experiments due to the distinct melt curves exhibited between the two different insert types.

4.3.4.2 Development of type IV and type VI reference sequences for use in HRM experiments

Synthetic oligos based on sequences obtained from two F_1 generation samples, U23 (insert type IV, synthetic oligo 2) and B24 (insert type VI, synthetic oligo 4; Table 8), were created to act as reference sequences in HRM analysis. HRM analysis showed that the two reference sequences exhibit two different monophasic melt profiles, labelled 2 and 4 on Figure 35. Mixing the two oligos in a ratio of 1:1 result in a distinctive bi-phasic melt profile, labelled 2+4. These different melt profiles can be used to distinguish between different zygosities; melt profiles 2 and 4 are representative of hemizygous samples, representing insert type IV and VI respectively, while 2+4 is representative of heterozygous sample containing both insert types IV and VI.

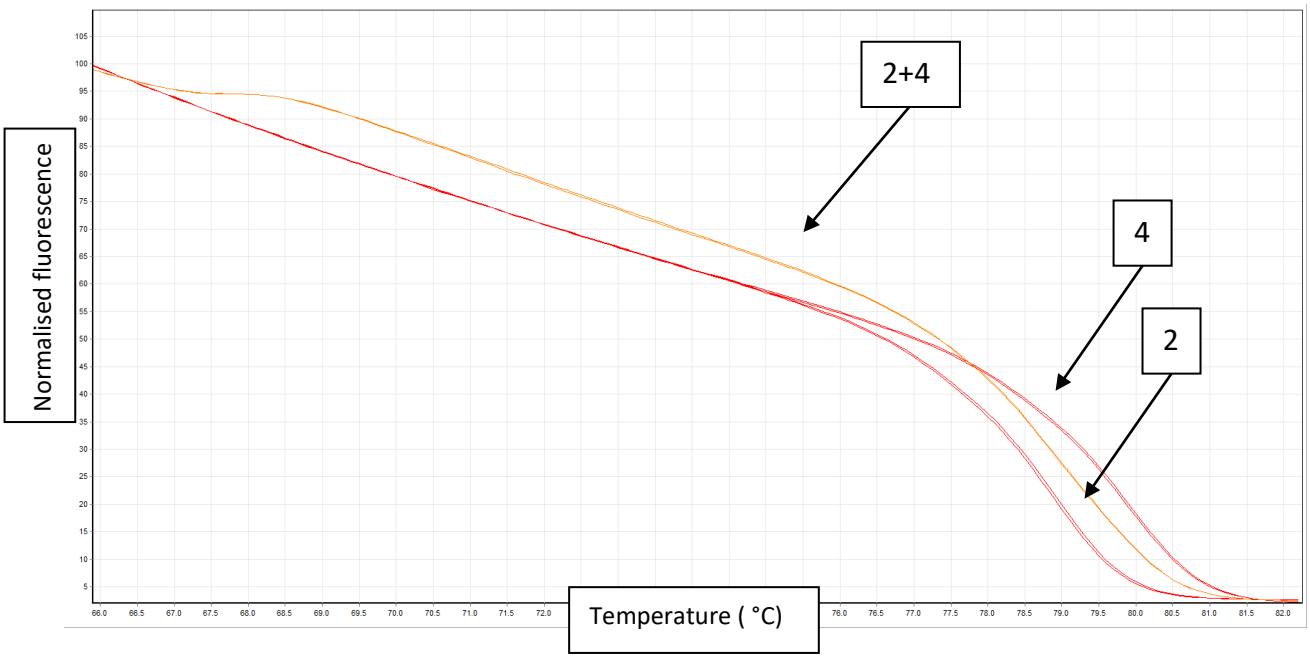


Figure 35: HRM analysis showing normalised fluorescence of two synthetic oligos designed as reference samples for HRM analysis of *T. cacao* F1 generation crosses from *T. cacao* parent plants containing different eTcBV inserts types (IV and VI). Two synthetic oligos representing curve profiles of hemizygous samples are labelled as 2 (U23, type IV) and 4 (B24, type VI), while a mix of the two oligos showing a heterozygous melt profile is labelled 2+4.

4.3.4.3 Determination of inheritance patterns of eTcBV type V and VI F_1 generation crosses using HRM

Through HRM using primers 1/4 and utilising the three reference sequences tested in 4.3.4.2 Development of type IV and type VI reference sequences for use in HRM experiments, the zygosity of 64 progenies from F_1 generation eTcBV hybrids were determined. These samples all originate from EET183 x IMC60 (female x male identities not factored; Table 11). A representative normalised fluorescence graph can be seen in Figure 36. All samples presented with different melt curves depending on which insert type they contain.

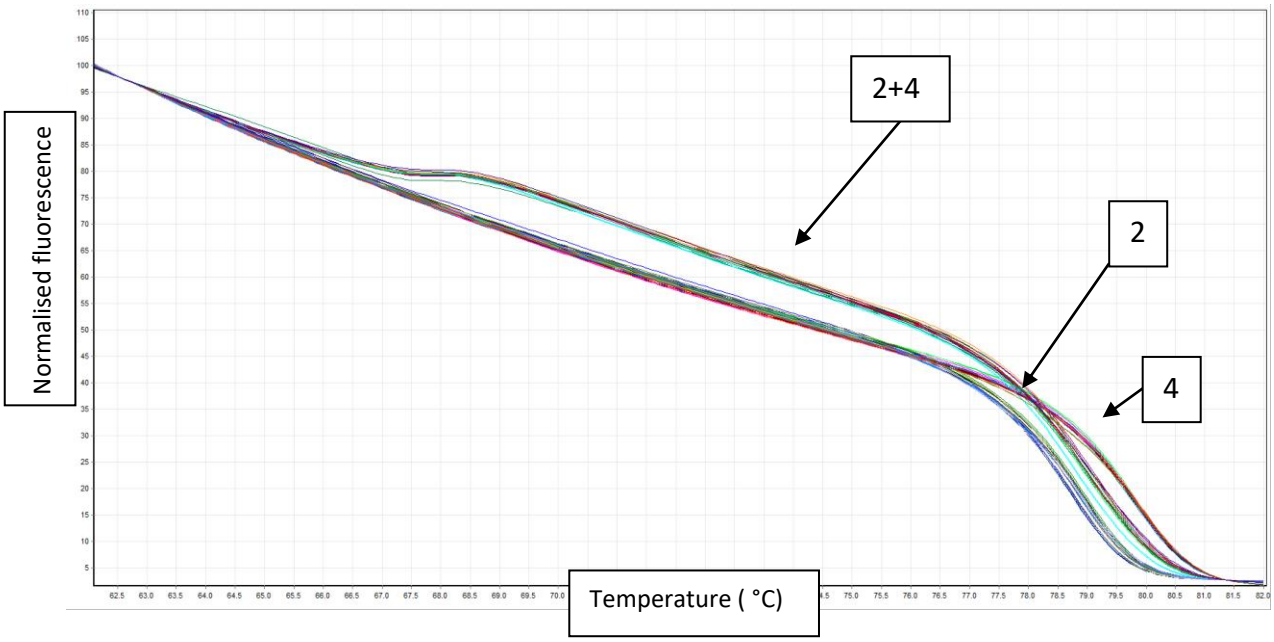


Figure 36: HRM analysis showing normalised fluorescence of 32 EET183 X ICS40 F1 generations using primer pair 1/4. Reference oligos 2 (type IV), 4 (type VI) and 2+4 (types IV and VI) included and labelled.

Table 11: Sample zygosities from eTcBV hybridisation EET183 x ICS40 as determined by HRM.

F_1 generation cross	**Total number of samples tested**	**Sum of zygosities based on sample HRM results**			
		Hemizygous		Heterozygous	Negative
		IV	VI		
EET183 x ICS40	64	27	10	20	7

Two Chi-Square Goodness of Fit Tests were performed to compare the observed genotype frequencies of F_1 seedlings with expected Mendelian ratios. The expected frequencies of type V positive and negative and type VI positive and negative generations in the sample of 64 progeny were determined separately using a 1:1 ratio; the zygosity ratios expected when crossing two diploid plants hemizygous for an eTcBV insert (Figure 37).

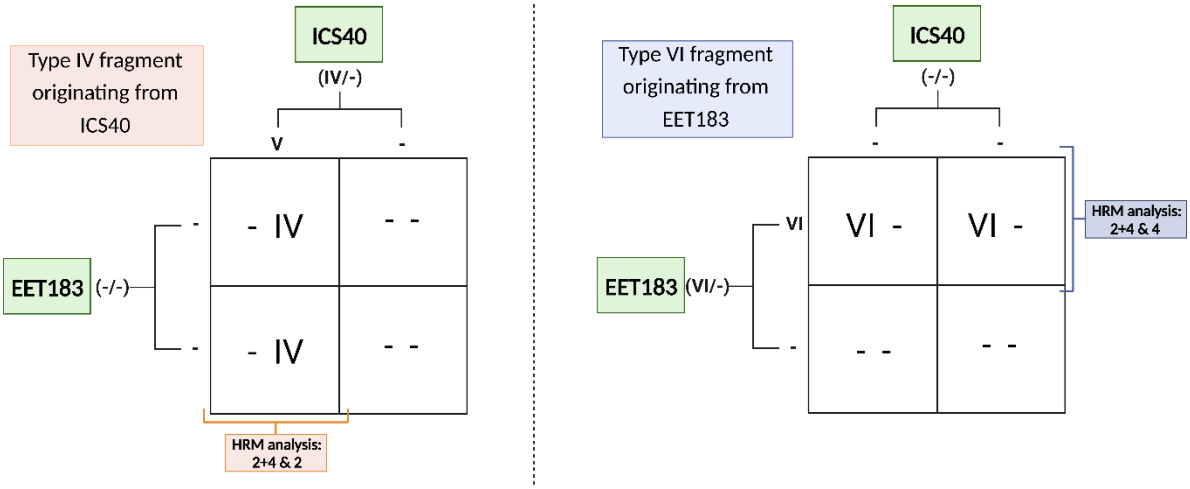


Figure 37: Expected zygosities of progeny resulting from crossing of ICS40 (hemizygous insert type V) and EET183 (hemizygous insert type VI), assuming that insertions are single in nature. Resulting progeny should segregate in a 1:1 ratio (insert type positive : negative).

4.3.4.4 Chi-square results: inheritance of the type IV insert

Chi-square results (Table 13) indicate there is a significant difference (*** p = .0002; < .001) between the observed frequencies of positive type IV inserts and negative samples in the F_1 generation hybrid EET183 x ICS40 samples tested, compared to expected percentages of zygosities based on mendelian inheritance patterns. This suggests that the type V insert is not inherited in a traditional Mendelian fashion, based on one hemizygous insert type genome (Figure 37).

Table 12: Expected vs. Observed numbers and percentages of progeny positive for a type IV fragment based on a sample size of 64 and results outlined in Table 10. Expected numbers based on typical Mendelian inheritance observed when crossing two single copy, hemizygous diploids (Figure 37).

Outcome	**Expected number**	**Observed number**	**Expected %**	**Observed %**
Type IV	32	37	50	73.44
Negative	32	17	50	26.56
Total	64	64	100	100

Table 13: Results of Chi-square Goodness of Fit test based on expected and observed percentages outlined in Table 11

Chi-square value	**14.06**
Degrees of Freedom	1
P value (two tailed)	0.0002
P value summary[1]	***
Is discrepancy significant (p < .05)?	Yes

[1] Significance illustrated using GraphPad style: *** p < .001.

4.3.4.5 Chi-square results: inheritance of the type VI insert

Chi-square results (Table 13) indicate there is no significant difference (ns, p = .617; ≥ .05) between the observed frequencies of positive type VI inserts and negative samples in the F_1 generation hybrid EET183 x ICS40 samples tested, compared to expected percentages of zygosities based on Mendelian inheritance patterns, suggesting that the type VI insert is inherited in a traditional Mendelian fashion, based on one hemizygous insert type (Figure 37).

Table 14: Expected vs. Observed numbers and percentages of progeny positive for a type VI fragment based on a sample size of 64 and results outlined in Table 10.

Outcome	**Expected number**	**Observed number**	**Expected %**	**Observed %**
Type VI	32	30	50	46.88
Negative	32	34	50	53.13
Total	64	64	100	100

Table 15: Results of Chi-square Goodness of Fit test based on expected and observed percentages outlined in Table 11.

Chi-square value	**0.25**
Degrees of Freedom	1
P value (two tailed)	0.617
P value summary[1]	ns
Is discrepancy significant (p < .05)?	No

[1] Significance illustrated using GraphPad style: ns, p ≥ .05.

4.3.4.6 Chi-square results: inheritance of two non-linked type IV inserts

Given the zygosity ratios and outcome of the Chi-square Goodness of Fit test for the type IV fragment, an amended Chi-square Goodness of Fit test was performed using different expected zygosity frequencies. This time, the expected frequencies of positive type IV fragments and negative generations in a sample of 64 was determined using a 3:1 ratio; the zygosity ratios expected of progeny resulting from an IMC40 cross, with two non-linked hemizygous type V inserts (Figure 38).

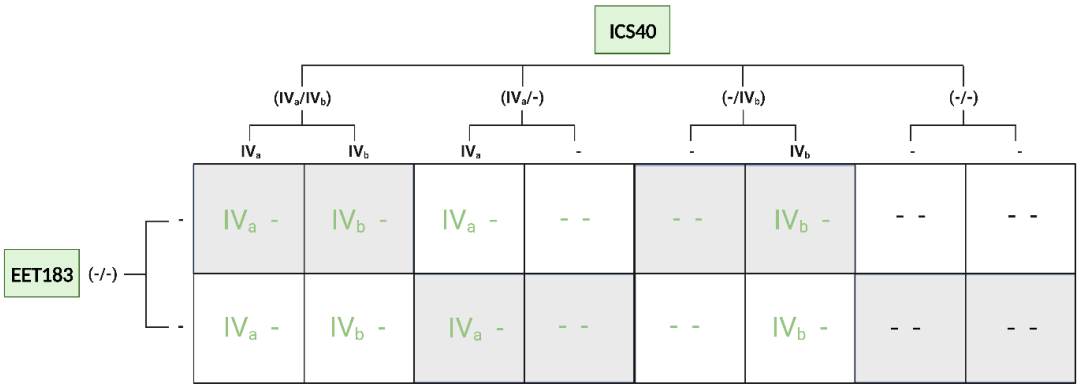


Figure 38: If two non-linked type IV fragments originate from *T. cacao* clone ICS40 (IV_a and IV_b), the expected resulting progeny split in a ratio of 3:1 (type IV positive : negative). Progenies comprise two adjacent squares highlighted in the same colour (either white or grey), with positives in green and negatives in black.

Chi-square results (Table 17) indicate no significant difference (ns, $p \geq .05$) between the observed frequencies of positive type VI inserts and negative samples in the F_1 generation hybrid EET183 x ICS40 samples tested, compared to expected percentages of zygosities based on Mendelian inheritance patterns assuming that parent plant ICS40 contains two non-linked hemizygous type IV inserts (Figure 38). The resulting progeny split in a ratio of 3:1, indicating that the type IV fragment is inherited on a Mendelian basis, but that two copies of the type IV fragment exist within the ICS40 parent plant.

Table 16: Expected vs. Observed numbers and percentages of each zygosity based on a sample size of 64 and results outlined in Table 11

Outcome	Expected number	Observed number	Expected %	Observed %
Type V positive	48	47	75	73.44
Negative	16	17	25	26.56
Total	64	64	100	100

Table 17: Results of Chi-square Goodness of Fit test based on expected and observed percentages outlined in Table 16.

Chi-square value	0.083
Degrees of Freedom	1
P value (two tailed)	0.773
P value summary[1]	Ns
Is discrepancy significant (P < 0.05)?	No

[1] Significance illustrated using GraphPad style: ns, p ≥ .05.

4.4 Discussion

4.4.1 Identification of ICQC *T. cacao* positive for badnaviral sequences

Putatively Badnavirus-free germplasm kept in the ICQC was screened for presence of Badnaviral DNA using degenerative Badnavirus specific primers developed by Muller *et al*. (2018) to incorporate sequence variability in a what was then, all known CSSV strains identified in West Africa. A high proportion of plants screened produced a positive PCR, an example of which is shown in Figure 25. Apart from occasional leaf morphology associated with nutrient deficiency, none of the ICQC clones have exhibited symptoms typical of badnavirus infection during six monthly inspections over the past eight years. Furthermore, the non-clustered positions of many of the PCR positive cocoa clones within the ICQC was not consistent with the 'crowd disease' nature of CSSVD, a disease that has not been observed to spread quickly or far (Thresh and Fargette, 2001). In 2018 during a study aimed to elucidate cocoa badnavirus diversity and describe cocoa badnaviral species, Muller *et al*. identified presence of an incomplete CSSV-related sequence, named species S, which clustered separately from other cocoa badnaviral species. This sequence was widely detected in asymptomatic cocoa trees, including samples that did not originate from West Africa. Results from Figure 25 and Figure 29, as well as taking into consideration the findings of Muller *et al*. (2018), brought into question whether integrated badnaviral sequences are prevalent in *T. cacao* populations worldwide.

4.4.2 Is transmission of these positive Badnaviral sequences by mealybug possible?

It is well established that CSSV, along with all other Badnaviruses, are transmitted from plant-to-plant through mealybug feeding (Strickland, 1951; Muller, 2008; Ameyaw *et al*., 2014; Bhat *et al*., 2016; Andres *et al*., 2017; Obok *et al*., 2018).

Transmission experiments were performed in order to determine whether these Badnaviral sequences identified in asymptomatic *T. cacao* germplasm belong to viral particles, or whether they are potentially integrated into the genome of *T. cacao.* An initial experiment verifying the use of *P. citri* mealybugs for the transmission of CSSV from a high viral load, CSSV positive *T. cacao* plant to a CSSV negative *T. cacao* plant was performed. Results indicate that mealybugs were able to successfully transmit CSSV viral particles from a CSSV positive plant to a CSSV negative plant, with CSSV viral particles at a detectable level using qPCR, in previously CSSV negative plants, just 1 month post viruliferous *P. citri* inoculation (Figure 28). Concentrations of CSSV were significantly lower in recently inoculated seedlings when compared to the *T. cacao* plants inoculated over nine years ago (Figure 28); the retained high concentration of CSSV within these 9-year old *T. cacao* plants is likely due to the optimal conditions in which they are kept The plants are not affected by any abiotic changes which, in the wild, are attributed to disease development and *T. cacao* mortality following CSSV infection. Despite this difference, CSSV was detected in all five biological replicates, indicating that this method of transmission of CSSV particles is successful. Following this, the same method was used on ICQC germplasm EP-PCR positive for a Badnaviral sequence. *P. citri* were allowed to feed on these positive *T. cacao* plants before being moved onto confirmed PCR negative plants. EP-PCR was once again performed on these negative plants one-month post-inoculation with *P. citri* previously fed on the ICQC Badnaviral positive plants, and results indicate that the *P. citri* were unable to transmit the Badnaviral sequence from these positive plants to the negative plants through feeding (Figure 29).

Following the initial identification of putative Badnaviral sequences found in asymptomatic *T. cacao* germplasm in the ICQC, in a collaborative effort between colleagues at CIRAD, the University of Reading and the University of the West of England, the presence of these positive Badnaviral sequences found in ICQC germplasm was explored (Appendix 2; Muller *et al.*, 2021). A variety of healthy leaf samples were obtained from *T. cacao* kept in the ICQC, the CIRAD collection, and from seedlings grown from germinated seeds. Included in these samples were

plants originating from the same clone, kept separately in the ICQC and CIRAD. A phylogeny was constructed from the alignment of positive sequences originating from these samples following DNA extraction, EP-PCR using Badnaviral specific primers and sanger sequencing. In total, 12 separate clusters of different insert types were identified and named type I to XII, with sequences obtained from two clones but kept separately clustering together. This result, along with the results of transmission studies outlined in above, would indicate that these sequences are unlikely to originate from a *P. citri* and CSSV outbreak within the separate germplasm collections, and are in fact EVEs scattered through the genomes of *T. cacao,* clustering according to *T. cacao* genetic groups, now known as endogenous *Theobroma cacao* bacilliform viruses (eTcBV). Following this discovery, the remaining results of this chapter focussed on three *T. cacao* clones originating from the ICQC containing the eTcBV insert types IV, V and VI.

4.4.3 Detection of eTcBVs in progeny resulting from insert type V and VI hybrids

Hybrid seedlings from ICQC *T. cacao* plants IMC60, ICS40 and EET183 (type IV, V and VI inserts respectively) were obtained and used to create a HRM method for identification of these two insert types, as well as to further understanding of the inheritance of these insert types. It is known that the parent plant EET183 is hemizygous for the type VI insertion locus (Muller *et al*., 2021), which provides a good basis on which to identify inheritance patterns. However, the zygosity status of the ICS40 and IMC60 insertion loci are unknown.

Seedlings underwent DNA extraction, EP-PCR using Badnaviral specific primers and gel-electrophoresis to identify those positive for an eTcBV (Figure 30). All positive PCR products were sent for sanger sequencing, which resulted in 19 clean Badnaviral sequences belonging to three different hybrid types: EET183 X IMC60 (13 samples), EET183 X ICS40 (four samples) and IMC60 X ICS40 (two samples; Table 10), with which a phylogenetic tree was created (Figure 19). In addition to these 19 samples, clean sequences were also obtained for two of the parent plants: IMC60 and ICS40. The sequence for parent plant EET183 was obtained from Muller *et al.*

(2021), accession number MW009806, which was isolated using the same Badnaviral specific primers. Phylogenetic analysis of these samples revealed three distinct clusters which were each attributed to the presence of a different insert type (Figure 19) as expected. Branch lengths of each outgroup are small (>0.02 substitutions per site), indicating very low genetic variability between samples within each outgroup. Samples from EET183 X ICS40 and ICS40 X IMC60 hybrids clustered together with the ICS40 parent sample, indicating presence of a type IV insert. Samples from EET183 X IMC60 hybrids segregated into two different clusters, each representing a type V and type VI fragment, associated with parent plants IMC60 and EET183 respectively. The vertical transmission of these sequences provides more evidence of their integrated origin. Using results from this multiple alignment and phylogenetic analysis, a method for HRM analysis of different insert types was developed.

4.4.5 Development of at HRM method for the detection of type IV and type VI eTcBV inserts in *T. cacao*

Eight primers were designed based on the multiple alignment of Badnaviral specific sequences obtained from 19 type IV, V and VI insert hybrid seedlings. Three forward primers and five reverse primers were combined in different combinations with different amplicon lengths ranging from just 54 bp to 246 bp, in order to determine the best combination for use in HRM experiments. The main consideration was that HRM analysis resulted in different melt curve profiles depending on the insert type of the sample, thus allowing the user to distinguish between different insert types.

An initial experiment using EP-PCR and the different primer combinations on two eTcBV positive samples (G9, type V and AI3, type VI) was performed. While bands were identified in the expected amplicon region for each primer pair, a few were not specific enough, resulting in the amplification of larger fragments, producing multiple bands (Figure 33). All successful primer pairs, except those producing multiple products (2/7 and 3/7) were then implemented in a HRM experiment using the two eTcBV positive samples G9 and AI3. HRM analysis revealed that all primer

pairs produced typical monophasic curve profiles, with results from G9 and AI3 easily distinguishable from one another (Figure 34). A smaller amplicon size is optimal for HRM experiments, as the ability to distinguish between single bp differences becomes more difficult in larger (>200 bp) fragments (Tong and Giffard, 2012). It is for this reason that primer pair 1/4 was chosen for all subsequent HRM experiments.

In the following experiments, samples from EET183 X ICS40 hybrids (type IV X type VI) were used as this has the largest number of samples available to test. Using the results from the phylogenetic analysis of F_1 generation hybrid samples, one type IV sample (U23) and one type VI sample (B24) were chosen as a template from which to design synthetic oligos for use as reference sequences in HRM experiments. The two synthetic oligos (named 2; type VI and 4; type IV) underwent HRM analysis individually and in a 1:1 mix, to represent a heterozygous melt pattern. The different G-C content in each sample resulted in distinct monophasic melt curve profiles for sample 2 and 4 individually, with a biphasic melt curve profile attributed to a heterozygotic sample (both type IV and type VI; Tong and Giffard, 2012) obtained with the 1:1 reference sequence mix (Figure 35).

4.4.6 Screening F1 generation eTcBV insert type IV and VI hybrids for to determine progeny insert type and heritability patterns using HRM

In total, 64 F_1 generation progeny of the cross between eTcBV positive *T. cacao* plants EET183 and ICS40 (type VI and IV respectively) were screened using the HRM method developed above in order to determine the percentages of samples containing either a type IV, type VI, type IV and VI, or no eTcBV insert. HRM analysis was successful in segregating the 64 samples into either type IV, type VI, type IV + VI or negative based on the synthetic reference oligo melt profiles (Figure 35). In total, 27 samples were positive for a type IV eTcBV, 10 were positive for a type VI eTcBV, 20 were positive for both type IV and VI eTcBVs, and seven were eTcBV negative (Appendix III).

Two Chi-square Goodness of Fit tests were performed using these observed outcomes (Table 12 and Table 13) in order to determine whether or not the patterns are analogous to those of typical Mendelian inheritance; the suggested mode of inheritance of these eTcBV elements (Figure 37; Muller *et al.,* 2021). Chi-square results indicate that the mode of inheritance for the type VI insert is not significantly different to that of typical mendelian inheritance (Table 13), whereas results for the type IV insert differ significantly (Table 15). It was theorised that this could be due to the presence of an additional copy of the type IV insert on a different locus in the genome of IMC60, resulting in the progeny inheriting this insert type at a ratio of 3:1, opposed to a 1:1 ratio as expected with a single copy (Figure 38). A revised Chi-square Goodness of Fit test using this updated ratio yielded a non-significant result, suggesting that two copies of the type IV fragment are present within the IMC60 genome, resulting in a larger percentage of progeny acquiring this insert type (Table 11), and therefore confirming a Mendelian mode of inheritance.

4.4.7 Type IV and VI eTcBV insert heritability

In this chapter, a successful method for the detection of type IV and VI by HRM analysis has been developed, which will aid in *T. cacao* quarantine practices to distinguish between CSSV infection and presence of an eTcBV insert. This method has also proved successful in identifying patterns of inheritance associated with these insert types. Results from this chapter have indicated that both type IV and VI inserts are inherited in a typical Mendelian fashion. The presence of an additional copy of the type IV insert has been identified in IMC60, resulting in a higher percentage of progeny inheriting this type IV insert. Multiple EVE inserts have previously been identified in the genomes of host plants, such as *N. tabacum* and *Musa spp.* (Bejarano *et al.*, 1996; Staginnus and Richert-Pöggeler, 2006).

A question remains over the reason for these integrated sequences to be maintained within the genome throughout generations; it is likely that the insertion is beneficial to the host, providing some positive impact on competitiveness, otherwise one would expect degeneration of the insert over time and eventually

elimination. The benefits of EVEs have been shown by Mette *et al.* (2002) and Umber *et al*. (2016), where two endogenous viral forms have been shown to play possible roles in anti-viral defence through homology-dependent gene silencing. Mette *et al*. (2002) observed features of TEPRVs that might be conferring resistance to the exogenous form for the virus, while Umber *et al*. (2016) discovered that the *Dioscorea bacilliform TR virus* (DBTRV; genus Badnavirus)- like endogenous sequence in the genome of *Diocorea cayenesis-rotundata* (yam) plays a possible role in anti-viral defence against DBTRV.

A HRM method for the detection of these insert types was chosen for development due to the speed at which the method can be performed, as well as the low cost associated with reagents and its specificity, allowing it to distinguish between single base pair changes in different DNA fragments (Reed *et al.*, 2007; Tong and Giffard, 2012) when compared to an alternative multiplex molecular screening method such as qPCR, which requires expensive probes and fluorescent dyes, or EP-PCR which is time consuming. However, HRM is limited in its ability to reveal specific sequence variations within DNA fragments, especially when the base pair change has no effect on %G-C content (Tong and Giffard, 2012). Despite this limitation, the results outlined in this chapter have shown that HRM is able to successfully distinguish between eTcBV insert types. It has the potential to enhance knowledge of the remaining 10 eTcBV insert types discovered so far and will provide a quick, easy and cheap screening technique for use in quarantine and research situations where distinguishing between CSSV infection and eTcBV presence is vital.

CHAPTER FIVE

In situ localisation of CSSV within *Theobroma cacao* leaf tissues

5.1 Introduction

5.1.1 Plant virus transmission

Elucidating the translocation patterns of viruses within organs and between plant cells is necessary for understanding plant-virus interactions. Identifying the localization of viruses *in situ* provides valuable information about viral transmission and is necessary for the development of infection prevention methods. While studying the feeding behaviour of hemipteran vectors of *Cucumber Mosaic Virus* (CMV), for example, Garzo *et al.,* (2002) determined that the CMV-resistant status of particular cucurbit varieties was due to those plants' capacity to prevent aphids from initiating ingestion from their phloem, thereby interrupting transmission of this phloem-limited virus. A crucial step to virus transport is through systemic transport through the vascular system, however there is also potential for virus transmission to occur from mother plant to seed (Shagril *et al*., 2015). In order to design an appropriate disease control programme, it is fundamental to first understand how the viral infection is spread (Chen *et al*., 2006).

Generally, the transmission of plant viruses can occur through two pathways. Horizontal transmission, whereby a virus from an infected host is transmitted to any other host independent of lineage is the most common method of virus transmission. Indirect horizontal transmission involves an intermediate host which acts as a vector (Chen *et al*., 2006). For example, the horizontal transmission of badnaviral species often occurs through vectoring by various mealybug species (Bhat *et al*., 2016), although some other viruses such as *Tomato yellow leaf curl virus* (TYLCV) is vectored by whitefly, specifically *Bemisia tabaci* (Andret-Link and Fuchs, 2005). Direct routes of horizontal transmission include through air and venereal, while indirect routes involve an intermediate biological host such as an insect vector (Chen *et al.*, 2006). The vertical transmission of viruses from an infected parent to its progeny can also occur through invasion of gametes by viral elements prior to fertilisation, invasion of the embryo post-fertilisation, and in

some cases infection of a seedling after germination through contact with infected maternal seed tissues. Vertical transmission of viruses is considered inefficient, due to meristematic exclusion of viruses in stem cells that rise to organs including flowers, gametophytes, and gametes (Bradamante *et al*., 2021).

5.1.2 Vector mediated transmission of plant viruses

Vectors are an essential component in the horizontal transmission of viruses from plant to plant. Plant virus vectors include arthropods, nematodes, and protozoa (Bragard *et al*., 2013; Bradamante *et al*., 2021). Of the 32 orders of the class Insecta, seven include vectors of plant viruses. Hemipterans are particularly well suited to their role as viral vectors, as their stylets are able to pierce the cuticle coat of the plant epidermis and take up and deposit virus directly into cell cytoplasm (Raccah, 2001).

Major modes of insect-vector mediated transmission of plant viruses include non-persistent; non-circulative, semi-persistent; circulative, non-propagative; and circulative, propagative (Andret-Link and Fuchs, 2005; Ng and Falk., 2006; Raccah and Fereres, 2009; Blanc *et al*., 2014; Ng and Zhou, 2015; Bhat *et al*., 2016; Dietzgen *et al*., 2016). Non-persistent transmission occurs when virus acquisition and inoculation into a new host must occur within seconds; the virus is restricted to the cuticle mouthparts or foreguts of its vector and an exemplar is the movement of cucumber mosaic virus (CMV) by the aphid, *Aphis gossypii,* to melon (Martin *et al*., 1997). Transmission of these viruses occurs through one of two mechanisms: capsid-only, whereby efficient transmission depends on the presence of viral coat protein (CP) and conserved capsid surface domains; or helper-dependent, where viral proteins are necessary for aiding uptake and acquisition of the virus by the vector. Non-circulative, semi-persistent viruses enter insect guts but do not enter tissues. They are thought to be retained on the chitin lining of the gut and survive for a period of hours to days before losing infectivity. Circulative, non-propagative viruses are capable of traversing the insect gut, haemolymph and salivary tissues, but do not replicate inside the vector tissues. They accumulate in the salivary glands where they can be secreted and inoculated into new host plants with insect saliva

(Gutiérrez *et al*., 2013). Circulative, propagative virus acquisition takes place over a longer period of time; hours to days, and retention of the virus is indefinite. These viruses not only traverse the insect gut, haemolymph and salivary tissues, but are capable of replication and systemic invasion of insect tissues prior to transmission. (Raccah and Fereres, 2009; Blanc *et al*., 2014; Dietzgen *et al*., 2016).

5.1.3 Badnavirus location in hosts and vectors

The location of Badnavirus virions within their host and vectors is subject to speculation. There is a lack of research pertaining to the localisation of episomal virions in host and vector tissues, and therefore transmission methods are often looked at to provide insight into the location of virions in host tissues and vectors. Some Badnaviruses are capable of integrating into the host genome, surviving as endogenous elements (Bhat *et al.,* 2023), making the search for episomal viral particles more difficult. Conventional molecular methods such as end-point PCR and RT-PCR can provide information about the presence and/or concentration badnaviral genetic material present in different tissues, but do not distinguish necessarily between episomal particles and endogenous elements.

Badnavirus acquisition by hemipteran, sap-feeding vectors could be an incidental part of the feeding process (Blanc *et al*., 2014). Mealybugs and aphids, common Badnavirus vectors, feed primarily on phloem sap and spend a lot of their feeding time probing around different plant tissues attempting to identify the location of the phloem. A study by Medberry *et al*. (1992) used transgenic tobacco to identify the location of expression of the Commelina Yellow Mottle Virus (CoYMV) promoter in transgenic tobacco. Promoters are involved in the direct production of major viral transcripts and can therefore act as an indicator for viral presence. In that study, the CoYMV promoter drove expression primarily in the phloem and phloem related cells. A similar study identified that the transgenic expression driven by the CoYMV promoter in oats, was detected in the vascular tissues of plant cells (Torbert *et al*., 1998). CoYMV is the type member for the badnavirus genus, and it is because of evidence from studies such as this that Badnaviruses are considered to be phloem-limited.

Although no longer classified as Badnavirus, *Rice Tungro Bacilliform Virus* (RTBV) is a confirmed phloem-limited virus in the genus *Tungrovirus*, *Caulimoviridae*. In 1986, using electron microscopy, RTBV was identified accumulating in the cytoplasm of companion cells and the xylem parenchyma of RTBV infected plant. Further validation of this study came in 1993, when Bhattacharyya-Pakrasi *et al.* identified the RTBV promoter accumulating in the vascular bundles of rice leaves, and furthermore that expression was phloem-specific.

5.1.4 Cacao Swollen Shoot Virus location in *T. cacao*

The location of CSSV within *T. cacao* plant tissues has not yet been demonstrated *in vitro*. Theories based on feeding patterns of the CSSV vector, the mealybug tend to suggest that CSSV is localised in the vascular system, specifically the phloem tissues (Figure 39; Jacquot *et al.,* 1999).

In 1999, Jacquot *et al*. set out to identify the *in-situ* localisation of cocoa swollen shoot virus in agroinfected *T. cacao*. Agrobacterium-mediated inoculation was used to artificially inoculate *T. cacao* seedlings with CSSV as an alternative to mealybug inoculation. Two full-length DNA clones of CSSV isolate Agou 1 were used as a source of viral sequence and inserted into the Agrobacterium binary vector Pbin19, before being introduced into *Agrobacterium tumefaciens* through electroporation. Three-week-old *T. cacao* cv. Amelonado leaf midribs, petioles and stems were inoculated with the transformed *A. tumefaciens* using micro-injection. Following successful inoculation of these seedling with CSSV, the intracellular location of CSSV virions was identified using electron microscopy. With the search focussing on the leaves and petioles of the agroinfected *T. cacao*, CSSV particles were seen in the vascular tissues, mainly the phloem companion cells and xylem parenchyma cells (Figure 39).

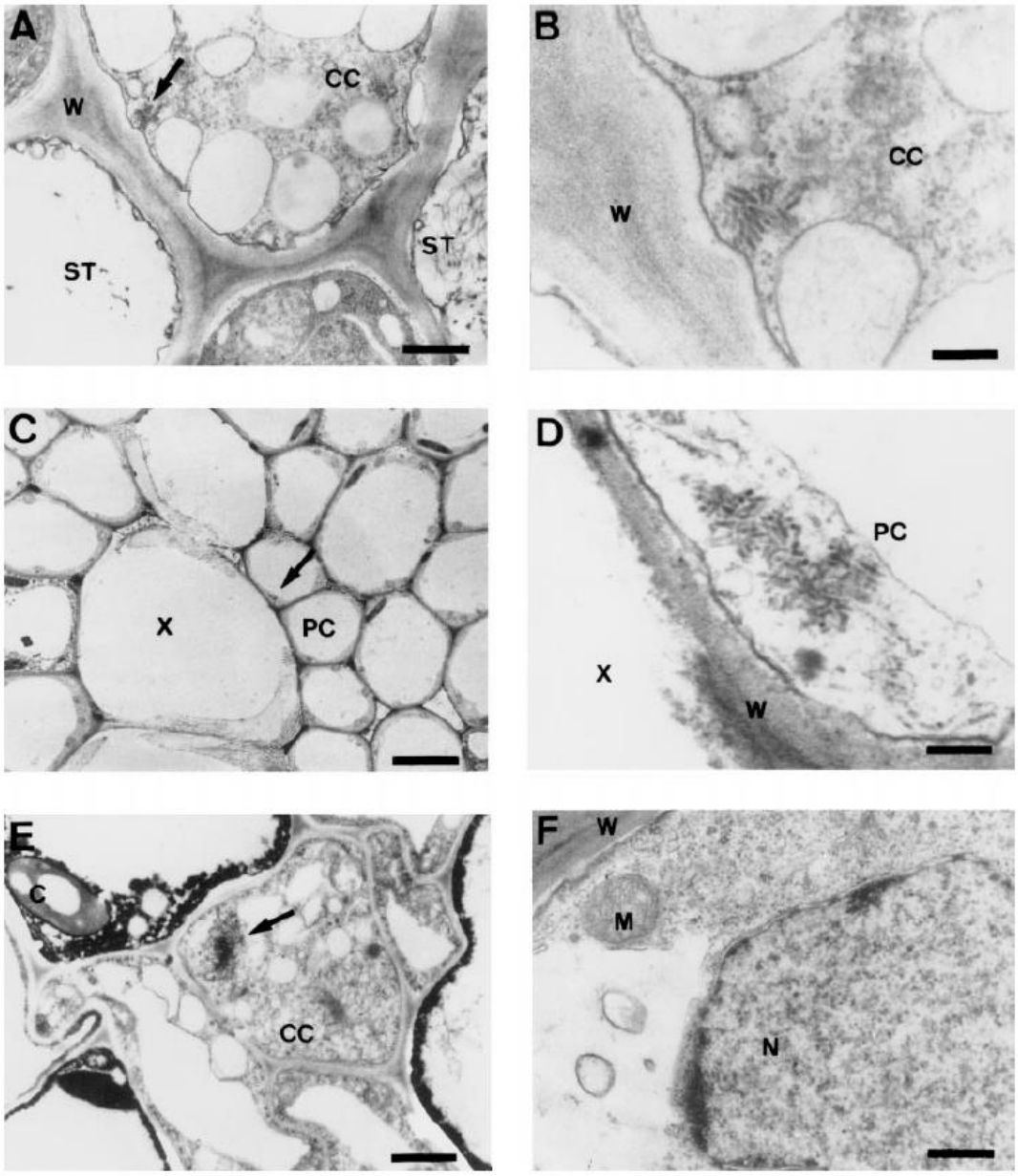


Figure 39: Figure taken from Jacquot *et al.* (1999). A selection of electron micrographs of vascular bundles of healthy and agro-infected *T. cacao* leaves and petioles. CSSV particles can be seen in phloem companion cells (images A, B, E) and in xylem parenchyma cells (images C, D). Image F depicts healthy phloem companion cells. B is a magnification of the region indicated by the arrow in A. D is a magnification of C. CC Companion cell; PC parenchyma cell; X xylem; ST sieve tube; W cell wall; M mitochondria; N nucleus; C chloroplast. Black bars correspond to: 625 nm (A), 200 nm (B, D), 6.6 µm(C), 1 µm (E) and 250 nm (F).

In 2008, a study by Quainoo *et al.* used PCR and agarose electrophoresis/capillary electrophoresis to identify the presence of CSSV DNA in cocoa seeds. Pollen grains, testae, cotyledons, embryos, and offspring of CSSV positive *T. cacao* were all tested. CSSV DNA was detected in every component part of the cocoa pod, and in the resulting seedlings. It was determined by the authors that CSSV can be transmitted through cocoa seedlings. This report contrasted with a series of studies showing an absence of vertical transmission in other Badnavirus species with *Sugarcane bacilliform virus* (SCBV) being one of the are exceptions (Balan *et al.*, 2021). More recent studies by Muller *et al.* in 2020 outline the presence of novel badnaviral sequences integrated into the genome of *T. cacao,* which may provide an alternative explanation for the detection of CSSV DNA in samples tested.

Electrical penetration graph (EPG) analysis is an effective method for the characterisation of hemipteran feeding patterns associated with transmission of plant viruses, for viruses such as Cauliflower mosaic virus, Turnip mosaic virus, Beet yellow virus and Cacao Swollen Shoot virus (Moreno *et al*., 2012; Jiménez *et al*., 2021; Obok *et al*., 2018). A study by Obok *et al*. in 2018 used EPG analysis and molecular based virus detection to monitor the stylet penetration behaviours of two West African mealybug species (*Planococcus citri* and *Pseudococcus longispinus*) and one non-West African mealybug species (*Ps. viburni*) on *T. cacao* leaves. Before EPG analysis began, mealybug species were identified using both DNA barcoding using cytochrome c oxidase (CO1) and high-resolution melt (HRM) analysis (Wetten *et al*., 2016). Following starvation, the feeding behaviour of the three mealybug species were recorded in real-time for 24 hours by EPG analysis. The different stylet probing and feeding behaviours were recorded and identified through the generation of eight different EPG waveforms, each with different, specific features: C – intracellular activities during penetration; E1e – extracellular salivation; E1 – salivation in sieve element; E2 – phloem ingestion; F – derailed stylet mechanics; G – xylem ingestion; Np – non-probing; and pd – intracellular stylet tip puncture. The time each mealybug species spent in each of these waveforms, and therefore feeding and stylet probing behaviours, was recorded and compared. Obok *et al*. identified that after the initial non-probing phase, waveform

C, intracellular activities during penetration, was the most frequently generated EPG pattern observed in all three mealybug species. This would indicate that the mealybugs spend the majority of their time probing for and establishing stylet contact with vascular tissues, raising the question: can CSSV be transmitted from mealybug vector to plant host through probing only? Or is the direct stylet contact and feeding within vascular tissues necessary?

5.1.5 Microscopy methods for the detection of plant viruses in host tissues

Knowledge of virus infection can be enhanced significantly through the imaging of viruses *in planta* using methods such as light microscopy and electron microscopy (EM). Initially, viruses were imaged predominantly by EM. However, with recently advances in light microscopy techniques such as improvements to sensitivity, functionality, resolution and live-imaging, methods utilising light microscopes and fluorescent approaches to imaging viruses has become more popular (Witte *et al.,* 2018). In 1999, Jacquot *et al*. published the first EM image of CSSV virions detected *in planta*. While this was a significant breakthrough in the understanding of CSSV pathogenicity, the method used artificial inoculation of *T. cacao* seedling (as described in 5.1.4). The localisation of CSSV virions *in planta* following inoculation by mealybug vector has not been identified.

Confocal microscopes use optical sectioning to generate high-contrast images, which make them particularly useful when imaging samples labelled with fluorescent probes. Samples do not need to be physically sectioned to very small depths, as the high-resolution objective lens can generate optical sections thinner than 1 µm, resulting in a single confocal image known as a 'slice'. A series of slices can be taken to produce a 3D dataset known as a z-stack, which is useful for reconstruction of the entire sample. Using confocal microscopy, the intensity and spatial arrangement of fluorescently tagged molecules or organisms can be calculated with high precision (Jonkman *et al*., 2020); however, it must be used in

conjunction with conventional biochemical and molecular assays to support the results (McClelland *et al.*, 2021).

5.1.6 Sample preparation for confocal microscopy using fluorescent molecules

5.1.6.1 Fluorescence *in situ* hybridisation (FISH)

FISH is a technique used to visualise gene transcripts in eukaryotic cells. Fluorescently labelled short DNA probes hybridise to complimentary sequences within the cell and can be detected by confocal microscopy (Figure 40). This method can be easily modified to allow targeted visualisation of specific viral or bacterial gene sequences.

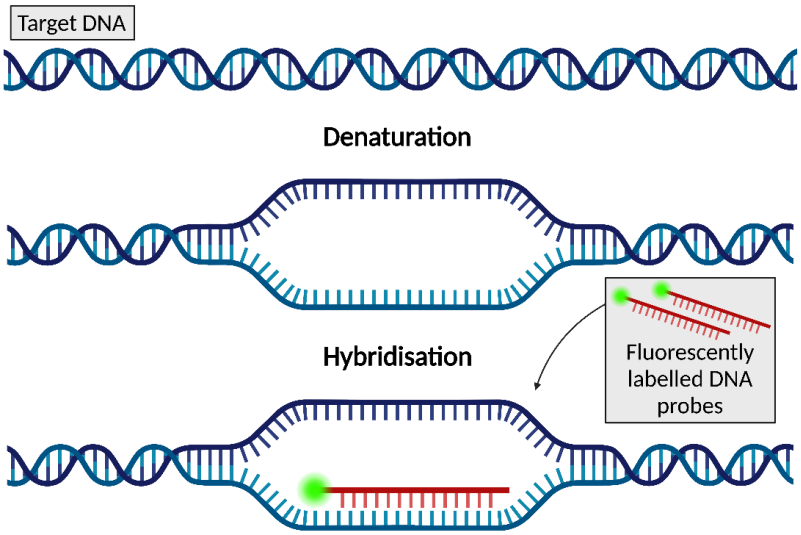


Figure 40: FISH detection method. DNA is denatured and targeted, complimentary fluorescently labelled DNA probes hybridise with the target DNA. Fluorescence emitted by these probes can be detected using confocal microscopy.

5.1.6.2 Immunofluorescence: fluorescently tagged antibodies

Antibodies are commonly implemented as ways of tagging proteins of interest for visualising under confocal microscopy. Two different fluorescent antibody binding techniques can be applied. Direct tagging involves the binding of a fluorescently tagged primary antibody to its complimentary target antigen, while indirect tagging

involves an un-tagged primary antibody and complimentary tagged secondary antibodies. The primary antibody binds to its target antigen, and fluorescently tagged secondary antibodies bind to the primary antibody (Figure 41). While the direct method of tagging is quicker, the indirect method is often employed due to its signal amplification, high sensitivity and ability to detect multiple targets in the same sample (Im *et al*., 2019).

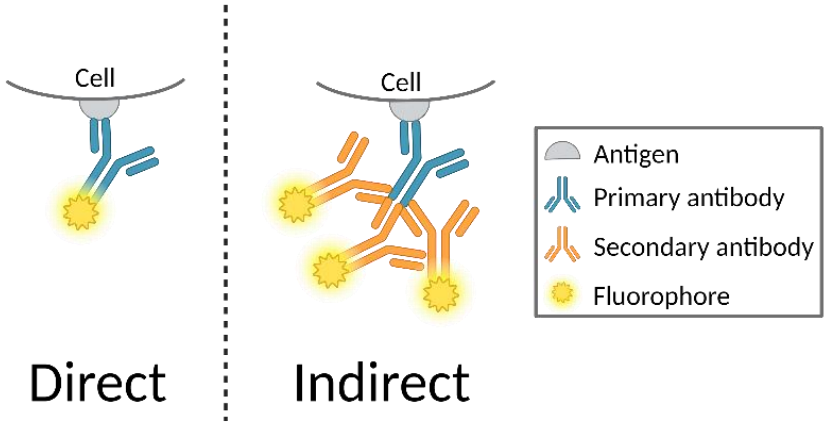


Figure 41: Direct and indirect modes immunofluorescence. Direct immunofluorescence involves a fluorescently tagged antibody binding directly to the antigen, whereas indirect immunofluorescence uses an un-tagged primary antibody with complimentary tagged secondary antibodies to provide a stronger fluorescent signal.

5.1.7 Use of confocal microscopy for *in situ* localisation of plant viruses

A study by Ghanim *et al*. in 2009 used FISH for the localisation of TYLCV and *Potato leaf roll virus* (PLRV) in both plant and insect vectors. The coat protein sequences of TYLCV and PLRV were used to synthesise two short oligonucleotide probes, conjugated with the fluorescent dye Cy3 at the 5' end. The ability of these probes to penetrate plant and insect tissues was expected to be high due to their short sequences. Hand-cut sections of TYLCV infected *Lycopersicon esculentum* (cv. Beefsteak) stems and leaves and PLRV infected *Physalis floridana* plants were prepared, along with control, uninfected samples from both plant species. Plant sections were first fixed before being hybridised overnight, mounted and viewed under a confocal microscope. Detection specificity was confirmed through the use

of no-probe, RNA-digested and healthy plant controls; no fluorescent signal was detected in any of these samples. TYLCV transcripts were present in the sieve elements of infected *L. esculentum* sections, while other vascular and surrounding plant tissues remained transcript-free. The same results were observed in the PLRV infected *P. floridana* sections. The short processing time, rapid generation of results and specificity of this method was praised by the authors.

In 2015, Shargil *et al*. developed a FISH technique, using confocal microscopy, for the visualisation of *Cucumber green mottle mosaic virus* (CGMMV) in plant tissues. Two DNA oligonucleotide probes specific to the viral genome of CGMMV were designed, labelled with either cyanine (Cy3) or fluorescein isothiocyanate (FITC). Conserved CGMMV ORF4-CP and replicase regions were targeted. *Cucumis sativus* (cv. Derby) and *C. melo* (cv. Raanan) leaf and flower samples from CGMMV infected plants and uninfected control plants were collected. A total of 60 samples, 30 CGMMV infected and 30 uninfected controls, were hand sectioned and fixed before undergoing an overnight hybridization using the CGMMV specific, fluorescently labelled DNA probes. Samples were examined using a confocal microscope. All 30 infected samples showed a positive fluorescent signal, while no signal was detected in the 30 uninfected controls. Confocal images showed that both CGMMV ORF4-CP and replicase transcripts were localised to the palisade cells, spongy mesophyll cells, trichomes and upper surface of *C. sativus* leaves, and trichomes of *C. melo* leaves. CGMMV ORF4-CP transcripts were also detected in the male flower of infected *C. sativus* and *C. melo,* but replicase was not present. Neither CGMMV ORF4-CP or replicase transcripts were detected in pollen grains. This study concluded that FISH is reliable method for localisation studies of CGMMV transcripts within a wide range of cell types and has the potential to be used for the study of other plant virus types.

5.1.8 Aims and objectives

In this chapter, immunofluorescence confocal microscopy using an indirect staining technique together with a CSSV specific antibody was employed to identify the location of CSSV in *T. cacao*. This work aims to build on the initial localisation study

performed by Jacquot *et al*. in 1999, by using CSSV positive *T. cacao* inoculated by mealybug, opposed to an agroinfected plant.

5.2 Methods

5.2.1 Sample selection

All *T. cacao* plants used in this experiment are maintained in the UWE Envirotron, as described in 2.2.2 *T. cacao* maintenance. In this chapter, leaf tissues taken from a selection of symptomatic, CSSV-infected G plants and CSSV-free Amelonado seedlings (Table 1) were utilised. The CSSV-free control plants were raised from seed supplied by the ICQC and were maintained in a compartment of the Envirotron separate from the CSSV-infected plants. Before sample collection, *T. cacao* plants were selected and tested for CSSV presence using qPCR.

5.2.2 Sample preparation

Mature leaves from *T. cacao* plants confirmed to be either CSSV positive or CSSV negative by qPCR were sampled. In the case of CSSV positive plants, leaves with visible CSSV symptoms (Figure 42 and Figure 43) were specifically chosen. Samples positive for CSSV were processed separately from negative samples to avoid any cross contamination.

Leaves were sectioned perpendicular to the veins (Figure 43) by hand using a razor blade with the leaf sandwiched between carrot blocks to minimise tissue crushing while slicing. Once sectioned, samples were kept in Phosphate Buffered Saline (PBS; Sigma) to avoid sample dehydration.

Figure 42: Symptomatic leaves from CSSV-infected cocoa plants chosen for sectioning, staining and visualisation using confocal microscopy. The code of G plant from which the leaf was selected is labelled in white on the top left of each image. Fern-like chlorosis, a typical symptom of CSSV, can be seen on all leaves. Sectioning locations are indicated in red.

Figure 43: Cocoa leaf sectioning by hand using a razor blade. Leaves were sectioned perpendicular to the secondary vein.

5.2.3 Sample staining

The staining procedure involved use of a primary antibody and secondary antibody. The primary antibody named Anti-CSSV AbD 31998.8 (patent number PCT/GB2019/053056; BioRad AbD Serotech, GmbH, Germany), designed to target and bind to surface proteins of CSSV, was provided for this experiment by Dr Jackie Barnett, Sue Taylor and Dr Joel Allainguilaume (University of the West of England, Bristol). The secondary antibody comprised a Goat anti-Human IgG tagged with a FITC fluorophore (Figure 44; Thermo Fisher Scientific, UK).

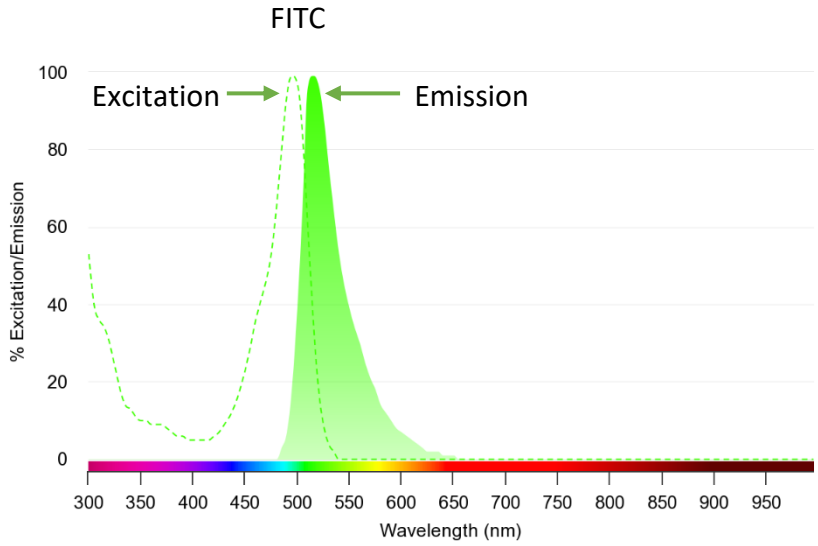


Figure 44: Excitation/emission spectrum for the FITC fluorophore. Max excitation and emission occur at 490 nm and 525 nm respectively. Spectrum downloaded from: https://app.fluorofinder.com/dyes/26-fitc-ex-max-490-nm-em-max-525-nm.

Antigen CSSV-CP-01, used to create AbD 31998.8, was designed by aligning five sequences of CSSV ORF3 (Hagen *et al.,* 1993; Muller and Sackey, 2005) and identifying the highly conserved consensus sequence encoding the CSSV coat protein. This consensus sequence was inserted into a plasmid and expressed in *Escherichia coli*, as described by Frisch *et al.* (2003); the coat protein unit was generated by BioRad AbD Serotech, GmbH, Germany. These recombinant coat proteins are capable of binding to multiple strains of CSSV. A selection of candidate antibodies on a human IgG with antigen CSSV-CP-01 were used to create a phage

display library using HuCal technology (BioRad, AbD Serotech, GmbH, Germany). AbD 31998.8 was identified following candidate screening on CSSV extract.

Staining was performed in two stages. The first stage involved sample incubation in the primary antibody (200 µg mL^{-1}), which was diluted to 1 µg mL^{-1} using sterile water (Thermo Fisher Scientific, UK). Sectioned tissues were carefully grasped using forceps, placed into a six-well plate, and submerged in 1 mL of diluted primary antibody. Samples were left to incubate, rocking at room temperature, for one hour. Following incubation, the primary antibody was pipetted away from the wells and discarded. Samples were washed in 1 mL Phosphate Buffered Saline (PBS; Thermo Fisher Scientific, UK), rocking for 10 minutes, three times in total.

The second stage of the staining procedure involved the addition of the secondary antibody with FITC fluorophore. While working with this antibody, aluminium foil was used to cover the samples and any vessels containing the antibody, to minimise exposure to light which would lead to degradation of the fluorophore. Samples were carefully transferred to a new, sterile six-well plate using forceps. The secondary antibody was diluted using sterile water to 1 µg mL^{-1}, before 1 mL was added to each of the samples. The six-well plate was covered in aluminium foil and incubated, rocking at room temperature for 30 minutes. Following incubation, the secondary antibody was removed from the wells using a pipette and discarded. Samples were washed in 1 mL PBS, rocking for 10 minutes, three times in total.

Stained samples were carefully placed on glass slides using forceps and mounted in 100 µL of 50% v/v PBS/glycerol (Thermo Fisher Scientific, UK). A coverslip was added, slides were covered in aluminium foil, and left to equilibrate at 4 °C for 1-12 hours.

5.2.4 Confocal microscopy

All confocal microscopy was performed using a Leica DMi8 Inverted Microscope (Leica Microsystems (UK) Ltd, UK). Fluorescence emission spectra were acquired using the 440 nm for excitation of chlorophyll and 488 nm for excitation of FITC

fluorophore. Emission fluorescence was recorded from 675 nm to 685 nm and 520 nm to 545 nm.

5.3 Results

Confocal microscope settings remained consistent throughout imaging, with fluorescence emission spectra acquired at 440 nm for chlorophyll and 488 nm for the FITC fluorophore, which acts as a marker for cells of interest. Emission fluorescence was recorded from 675 nm to 685 nm and 520 nm to 545 nm respectively. Emission at 675 nm to 685 nm is highlighted in red, while emission at 520 nm to 545 nm is highlighted in green. Master gain was adjusted and recorded with each sample in order to obtain the best possible visualisation of cell types and fluorescence.

Samples of CSSV negative and CSSV positive *T. cacao* underwent a range of preparation methods in order to determine whether background autofluorescence and/or non-specific binding of antibodies impact on image acquisition and subsequent putative virion visualisation. Antibody concentration remained the same, while the combination of antibodies was changed.

5.3.1 Does sample autofluorescence have an impact on potential visualisation of virions?

It was first necessary to determine whether sample autofluorescence and/or non-specific binding of the secondary antibody with fluorophore have a confounding effect on virion visualisation. Samples were visualised when prepared without any antibody, with secondary antibody only and with both primary and secondary antibody. Images were compared to determine whether or not the FITC fluorophore would be suitable for visualisation.

5.3.1.1 CSSV negative and CSSV positive *T. cacao* leaf tissues prepared in absence of primary and secondary antibodies

CSSV negative and CSSV positive samples underwent preparation as described in section 5.2.5, but without the addition of antibodies. Where samples would be incubated in antibody, they were instead incubated in PBS only.

Figure 45, images A and B, are representative confocal images showing a section of a *T. cacao* leaf sample negative for CSSV and a sample positive for CSSV respectively. Images were taken at a magnification of 63x, with master gain set to 520. The cuticle (C), upper epidermis (UE), palisade cells (P), spongy mesophyll (SM) and lower epidermis (LE) are labelled.

Images C and D show two confocal images focussing on the same *T. cacao* leaf sections as in Figure 45 when viewed using both dye detectors highlighted in green (460 nm) and red (650 nm). Image C depicts a CSSV negative sample of *T. cacao* leaf tissue, where some background fluorescence highlighted in green can be seen located around the upper and lower epidermis. Similarly, in image D, autofluorescence highlighted in green can be seen in a positive sample of CSSV positive *T. cacao* leaf tissue, again located around the upper and lower epidermis.

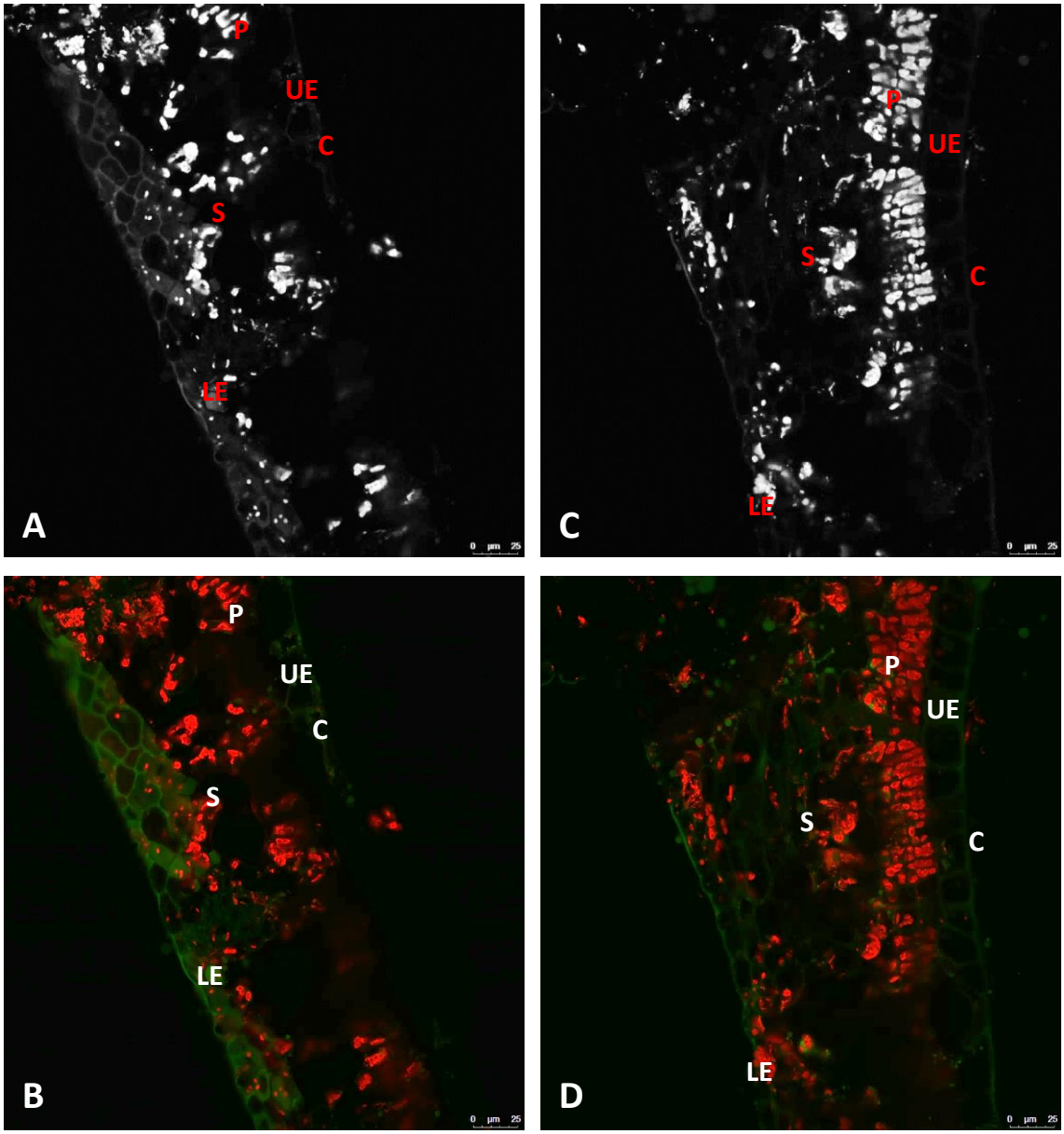


Figure 45: Confocal images of CSSV negative (A, B) and CSSV positive (C, D) *T. cacao* leaf cross sections with no antibody staining, at 63x magnification with master gain set to 520. Plant tissues are labelled as follows: C = cuticle, UE = upper epidermis, SM = spongy mesophyll and LE = lower epidermis. In images B and D, green highlights show autofluorescence of plant tissues emitting at the same wavelength as FITC (460 nm), with red highlighting chloroplasts emitting at wavelength 650 nm. Background fluorescence in both CSSV-negative (B) and -positive (D) samples is localised to the upper and lower epidermis.

5.3.1.2 CSSV negative and CSSV positive *T. cacao* leaf tissues incubated in only secondary antibody

A CSSV negative and CSSV positive sample underwent sample preparation as described in section 5.2.5, but without the addition of the primary antibody. Where samples would have been incubated in the primary antibody, there were instead incubated in PBS only.

Image B (Figure 46) depicts a CSSV negative sample of *T. cacao* leaf tissue, where some background autofluorescence highlighted in green can be seen. Similarly, in image D, autofluorescence highlighted in green can be seen in a positive sample of CSSV positive *T. cacao* leaf tissue. Autofluorescence appears to be limited to the upper epidermis (Figure 46, D).

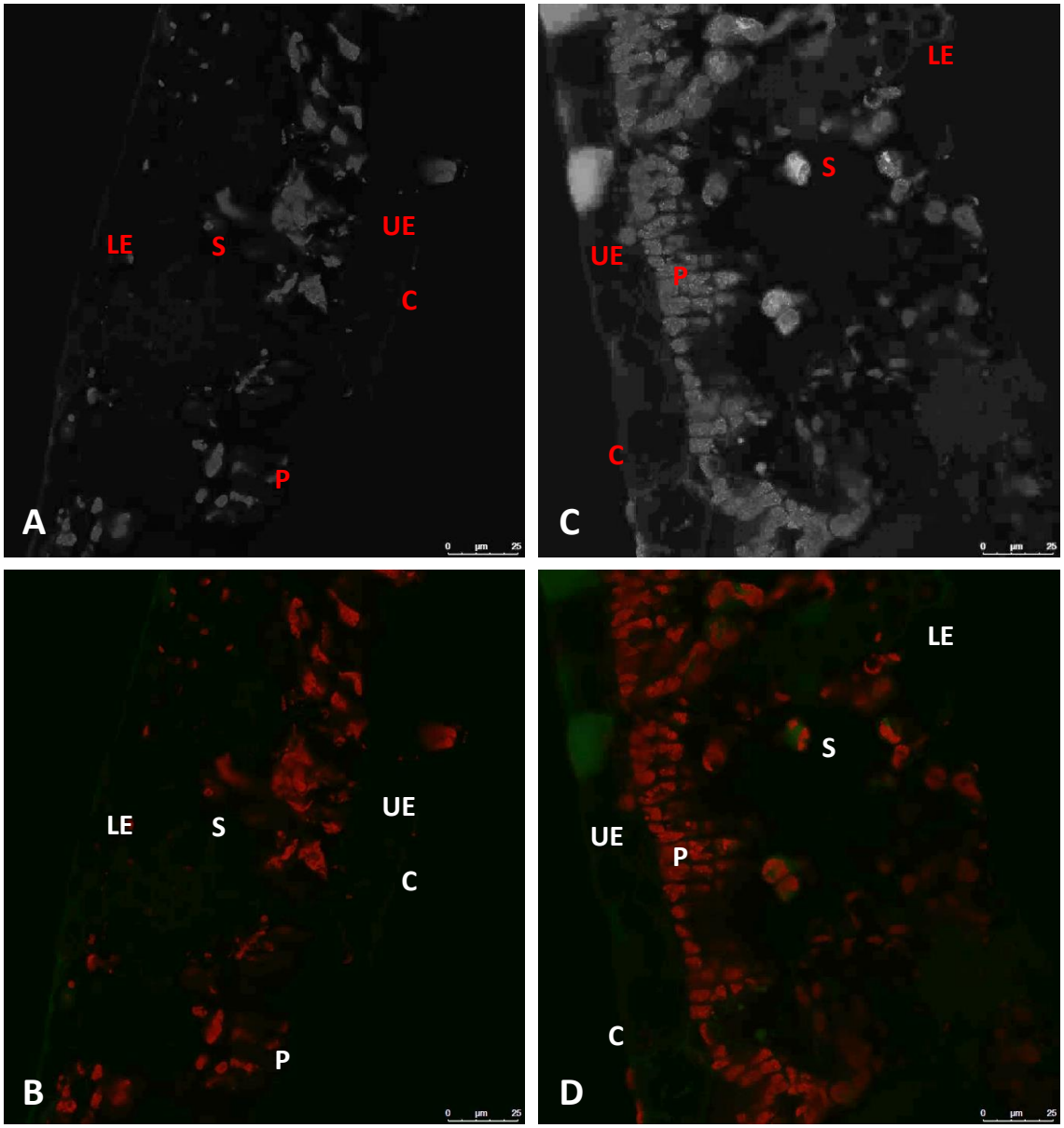


Figure 46: Confocal images of CSSV negative (A, B) and CSSV positive (C, D) *T. cacao* leaf cross sections with only secondary antibody staining, at 63x magnification with master gain set to 520. Plant tissues are labelled as follows: C = cuticle, UE = upper epidermis, SM = spongy mesophyll and LE = lower epidermis. In images B and D, green highlights show autofluorescence of plant tissues emitting at the same wavelength as FITC (460 nm), with red highlighting chloroplasts emitting at wavelength 650 nm. Background fluorescence is limited to the upper epidermis in CSSV-positive sample (D).

5.3.1.3 CSSV negative and CSSV positive *T. cacao* leaf tissues incubated in both primary and secondary antibody

A CSSV negative and CSSV positive sample underwent sample preparation as described in section 5.2.5, with no amendments.

Image B (Figure 47) depicts a CSSV negative sample of *T. cacao* leaf tissue, where some background autofluorescence around the upper epidermis, palisade layer, spongy mesophyll and lower epidermis can be seen. Similarly, in image D, autofluorescence around the lower epidermis can be seen in a positive sample of CSSV positive *T. cacao* leaf tissue. Bright spots of increased fluorescence in the FITC wavelength (460 nm) can also be seen in image D, limited to the spongy mesophyll.

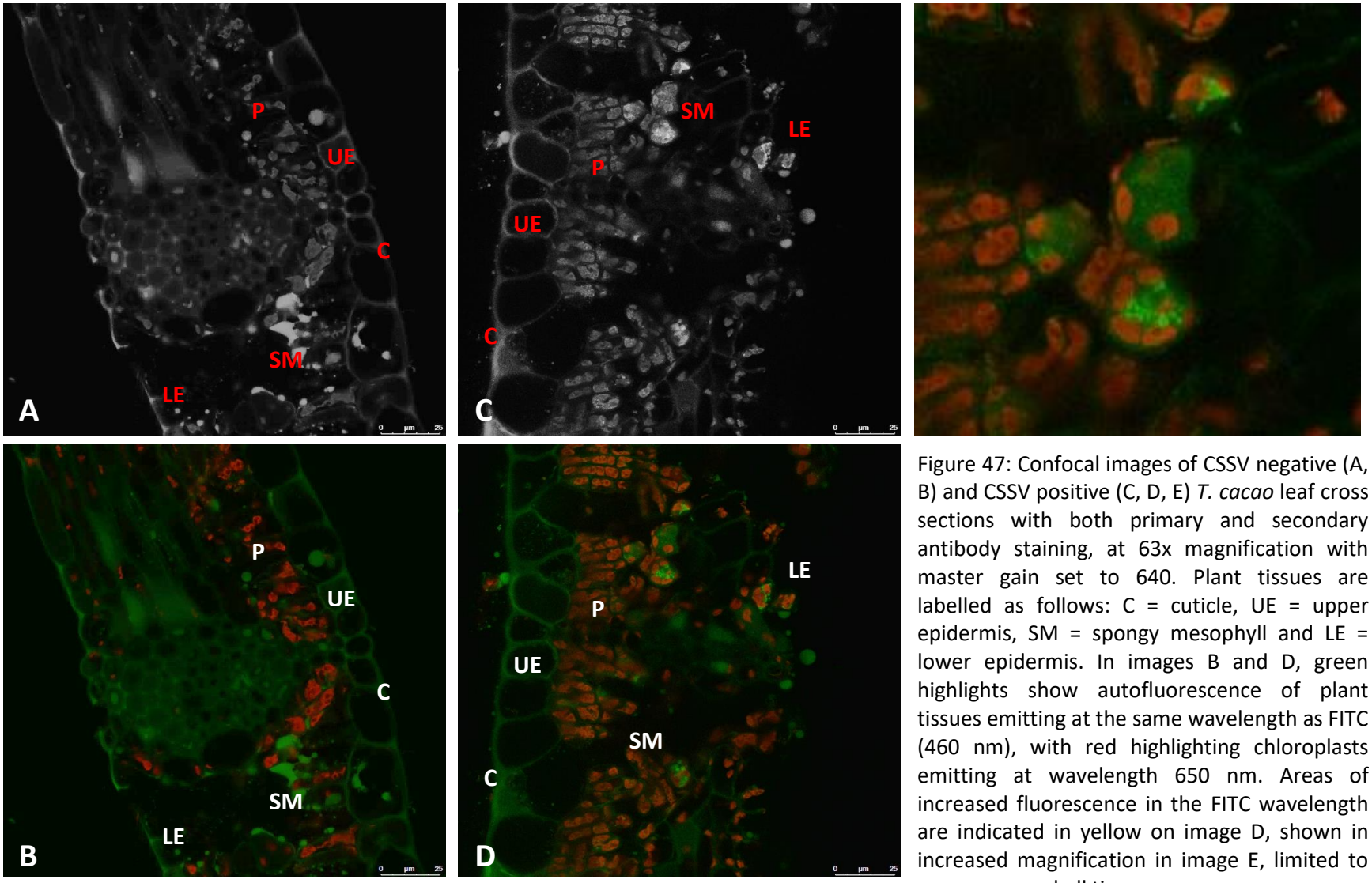


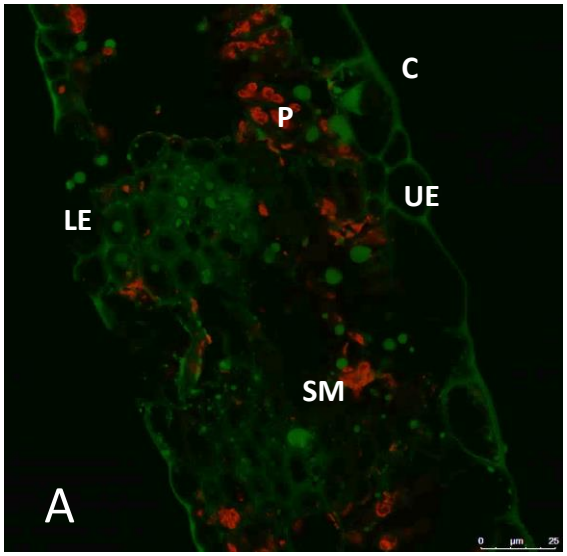
Figure 47: Confocal images of CSSV negative (A, B) and CSSV positive (C, D, E) *T. cacao* leaf cross sections with both primary and secondary antibody staining, at 63x magnification with master gain set to 640. Plant tissues are labelled as follows: C = cuticle, UE = upper epidermis, SM = spongy mesophyll and LE = lower epidermis. In images B and D, green highlights show autofluorescence of plant tissues emitting at the same wavelength as FITC (460 nm), with red highlighting chloroplasts emitting at wavelength 650 nm. Areas of increased fluorescence in the FITC wavelength are indicated in yellow on image D, shown in increased magnification in image E, limited to spongy mesophyll tissues.

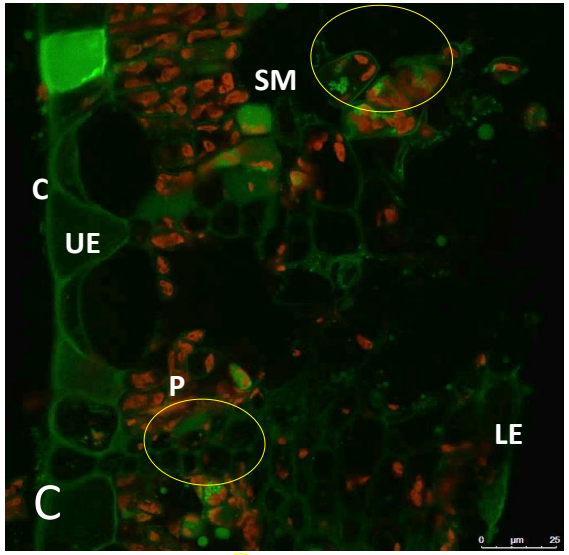
5.3.2 Does post-mounting incubation time have an effect on sample quality, signal fluorescence and/or background autofluorescence?

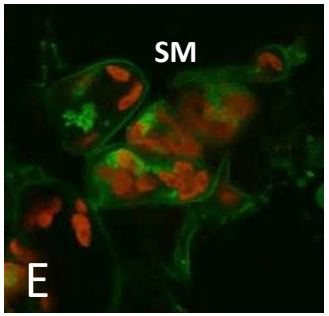
A one-hour post-mounting incubation time and 12-hour post-mounting incubation time were tested and compared visually to determine whether the amount of time the sample had to equilibrate post-mounting has an effect on the quality of the sample and/or presence of autofluorescence. A CSSV negative and CSSV positive sample underwent sample preparation as described in section 5.2.5, but with amended mounting incubation times as stated above.

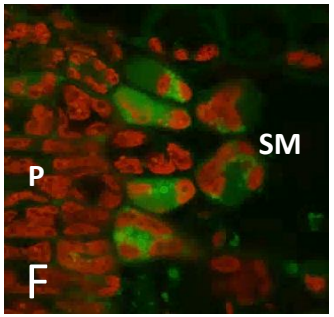
Samples shown in Figure 48, images A and C, have been prepared using a post-mounting incubation period of one-hour. Image A depicts a CSSV negative sample of *T. cacao* leaf tissue, where some autofluorescence can be seen in green around the upper epidermis, palisade layer, spongy mesophyll, and lower epidermis. A sample positive for CSSV depicted in image C also appears to emit some autofluorescence, limited to the upper epidermis and lower epidermis only. In image D, bright spots of increased fluorescence in the FITC wavelength (460 nm) can be seen within the spongy mesophyll.

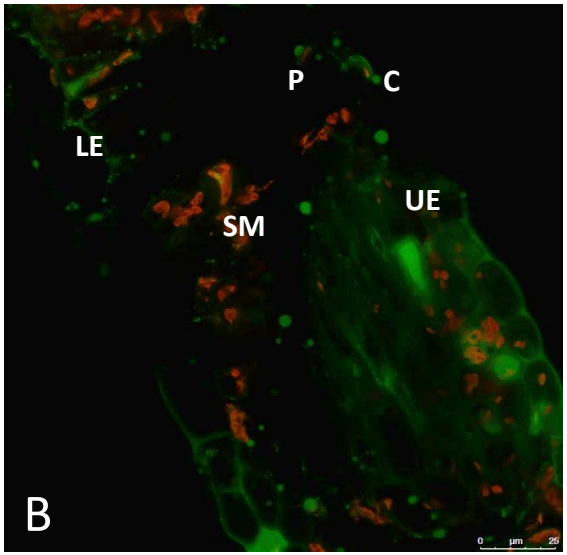
In Figure 48, images B and D, two samples prepared using a post-mounting incubation period of 12-hours are shown. Image B depicts a CSSV negative sample of *T. cacao* leaf tissue, where some autofluorescence can be seen in green around the upper epidermis, palisade layer, spongy mesophyll, and lower epidermis. A sample positive for CSSV depicted in image D also appears to emit some autofluorescence, limited to the upper epidermis and lower epidermis only. In image D, bright spots of increased fluorescence in the FITC wavelength (460 nm) can be seen within the spongy mesophyll.











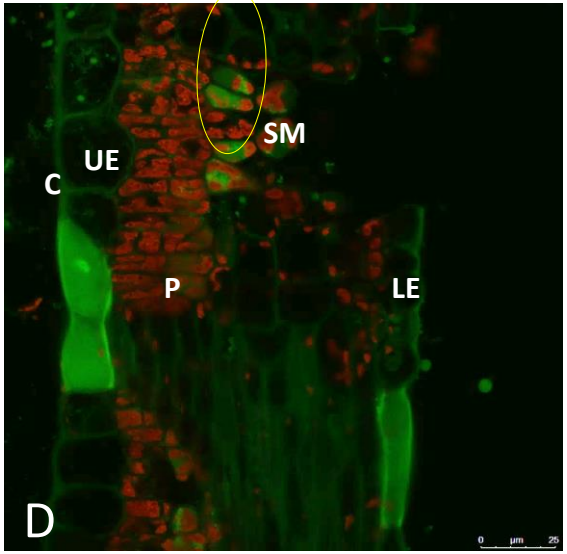


Figure 48: Confocal images of CSSV negative (A, B) and CSSV positive (C, D) *T. cacao* leaf cross sections with both primary and secondary antibody staining, and a 1-hour (A, C, E) or 12-hour (B, D, F) post-mount incubation, at 63x magnification with master gain set to 640. Plant tissues are labelled as follows: C = cuticle, UE = upper epidermis, SM = spongy mesophyll and LE = lower epidermis. Green highlights show fluorescence at 460 nm, with red highlighting chloroplasts emitting at wavelength 650 nm. Areas of increased fluorescence in the FITC wavelength (460 nm) are indicated in yellow on images C and D, limited to the spongy mesophyll tissues. Magnified areas of increased fluorescence within spongy mesophyll can be seen in images E (from image C).and image F (from image D)

5.3.3 Does the concentration of the primary antibody have an effect on signal brightness and/or background autofluorescence from plant tissues?

Figure 49, Figure 50 and Figure 51 depict both CSSV positive and CSSV negative *T. cacao* leaf samples when prepared as described in section 5.2.5, but with different concentrations of primary antibody used in the staining process. Concentrations of 0.1 µg mL^{-1}, 1 µg mL^{-1} and 4 µg mL^{-1} were chosen for comparison.

In Figure 49, two samples prepared using a primary antibody concentration of 0.1 µg mL^{-1} are shown. Image B (Figure 49) depicts a CSSV negative sample of *T. cacao* leaf tissue, where some autofluorescence can be seen in green, around the upper and lower epidermis. A sample positive for CSSV depicted in image D (Figure 49) also appears to emit some autofluorescence, limited to the cuticle. In image D (Figure 49), distinctive bright spots of increased fluorescence in the FITC wavelength (460 nm) can be seen within the spongy mesophyll. These signals were peculiar to the CSSV positive leaf tissues analysed.

Figure 50 depicts two samples prepared using a primary antibody concentration of 1 µg mL^{-1} . Image B (Figure 50) depicts a CSSV negative sample of *T. cacao* leaf tissue, where some autofluorescence can be seen in green, limited to the lower epidermis. A sample positive for CSSV depicted in image D (Figure 50) also appears to emit some autofluorescence, around the lower and upper epidermis. In image D (Figure 50), bright spots of increased fluorescence at 460 nm can be seen within the spongy mesophyll.

In Figure 51, two samples prepared using a primary antibody concentration of 4 µg mL^{-1} are shown. Image B (Figure 51) depicts a CSSV negative sample of *T. cacao* leaf tissue, where some autofluorescence can be seen in green, limited to the upper and lower epidermis. A sample positive for CSSV depicted in image D (Figure 51) also appears to emit some autofluorescence, around the lower and upper epidermis. In

image D (Figure 51), bright spots of increased fluorescence at 460 nm can be seen within the spongy mesophyll.

There are no significant differences in the levels of fluorescence detected at 460 nm in negative samples (Figure 49, Figure 50 and Figure 51; B) when comparing between the three different concentrations of primary antibody. A higher frequency of areas of increased fluorescence at 460 nm can be seen (Figure 50 and Figure 51; D) when samples are prepared using 1 µg mL^{-1} and 4 µg mL^{-1} concentrations of primary antibody.

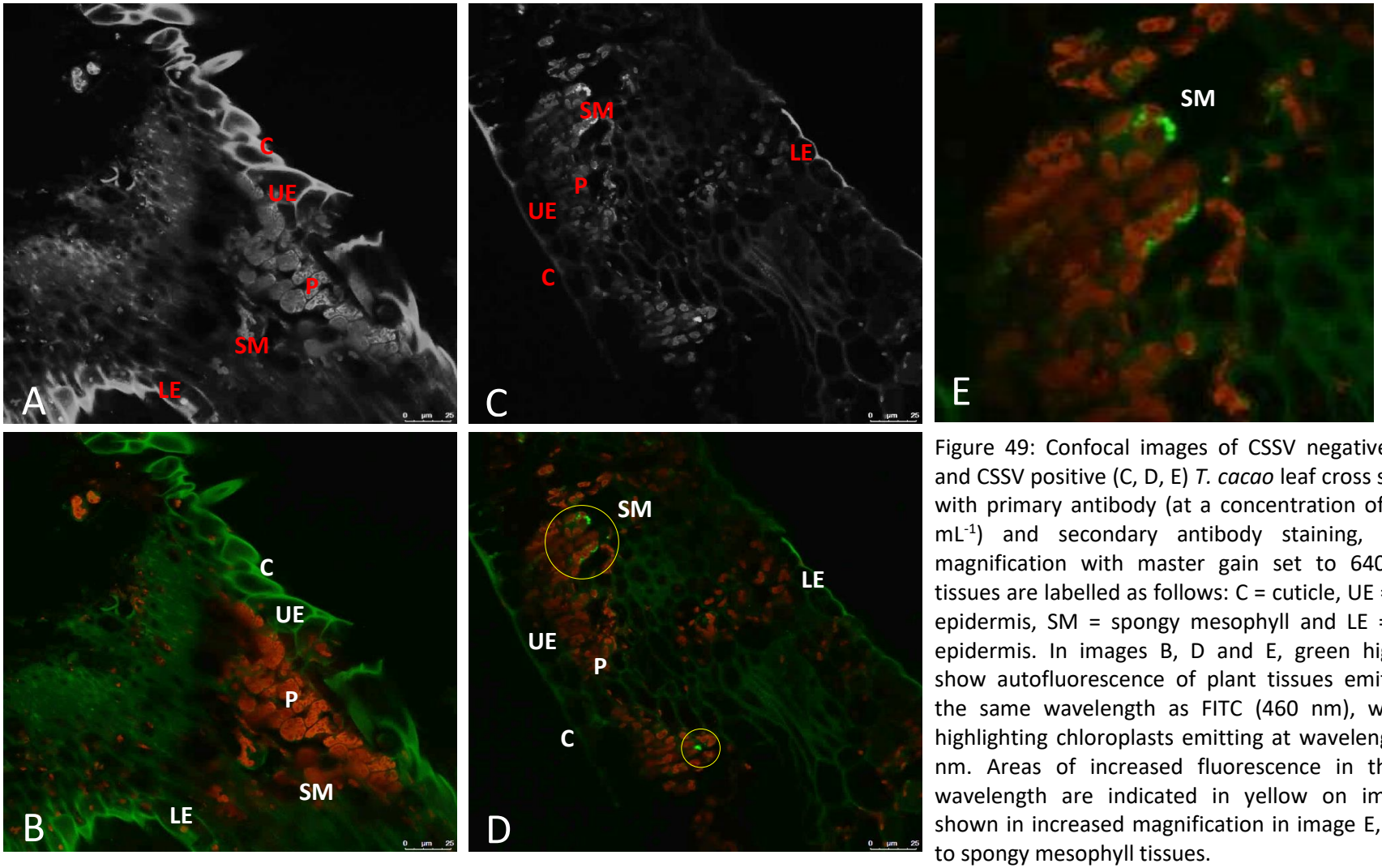


Figure 49: Confocal images of CSSV negative (A, B) and CSSV positive (C, D, E) *T. cacao* leaf cross sections with primary antibody (at a concentration of 0.1 µg mL^{-1}) and secondary antibody staining, at 63x magnification with master gain set to 640. Plant tissues are labelled as follows: C = cuticle, UE = upper epidermis, SM = spongy mesophyll and LE = lower epidermis. In images B, D and E, green highlights show autofluorescence of plant tissues emitting at the same wavelength as FITC (460 nm), with red highlighting chloroplasts emitting at wavelength 650 nm. Areas of increased fluorescence in the FITC wavelength are indicated in yellow on image D, shown in increased magnification in image E, limited to spongy mesophyll tissues.

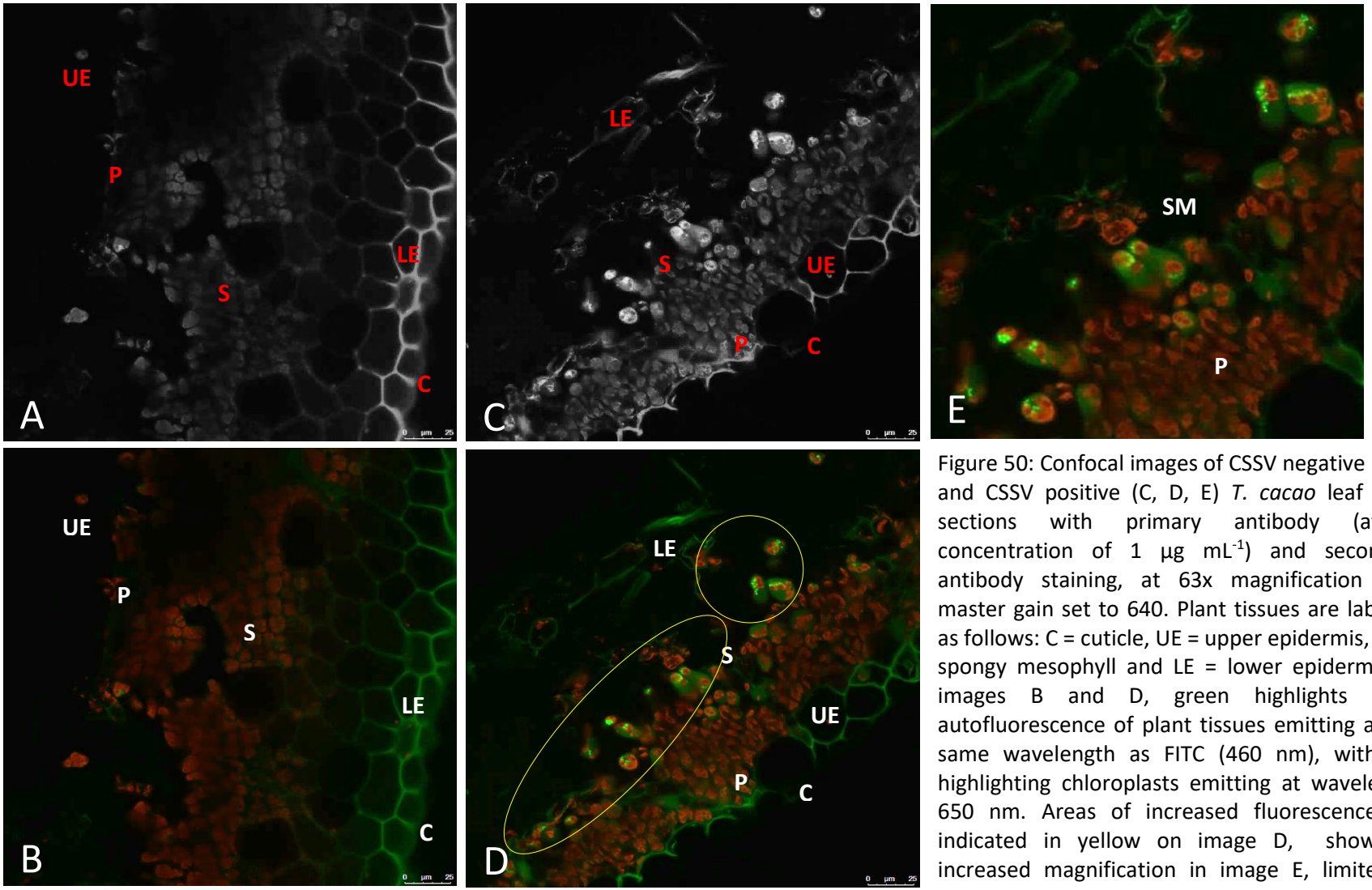


Figure 50: Confocal images of CSSV negative (A, B) and CSSV positive (C, D, E) *T. cacao* leaf cross sections with primary antibody (at a concentration of 1 µg mL^{-1}) and secondary antibody staining, at 63x magnification with master gain set to 640. Plant tissues are labelled as follows: C = cuticle, UE = upper epidermis, SM = spongy mesophyll and LE = lower epidermis. In images B and D, green highlights show autofluorescence of plant tissues emitting at the same wavelength as FITC (460 nm), with red highlighting chloroplasts emitting at wavelength 650 nm. Areas of increased fluorescence are indicated in yellow on image D, shown in increased magnification in image E, limited to

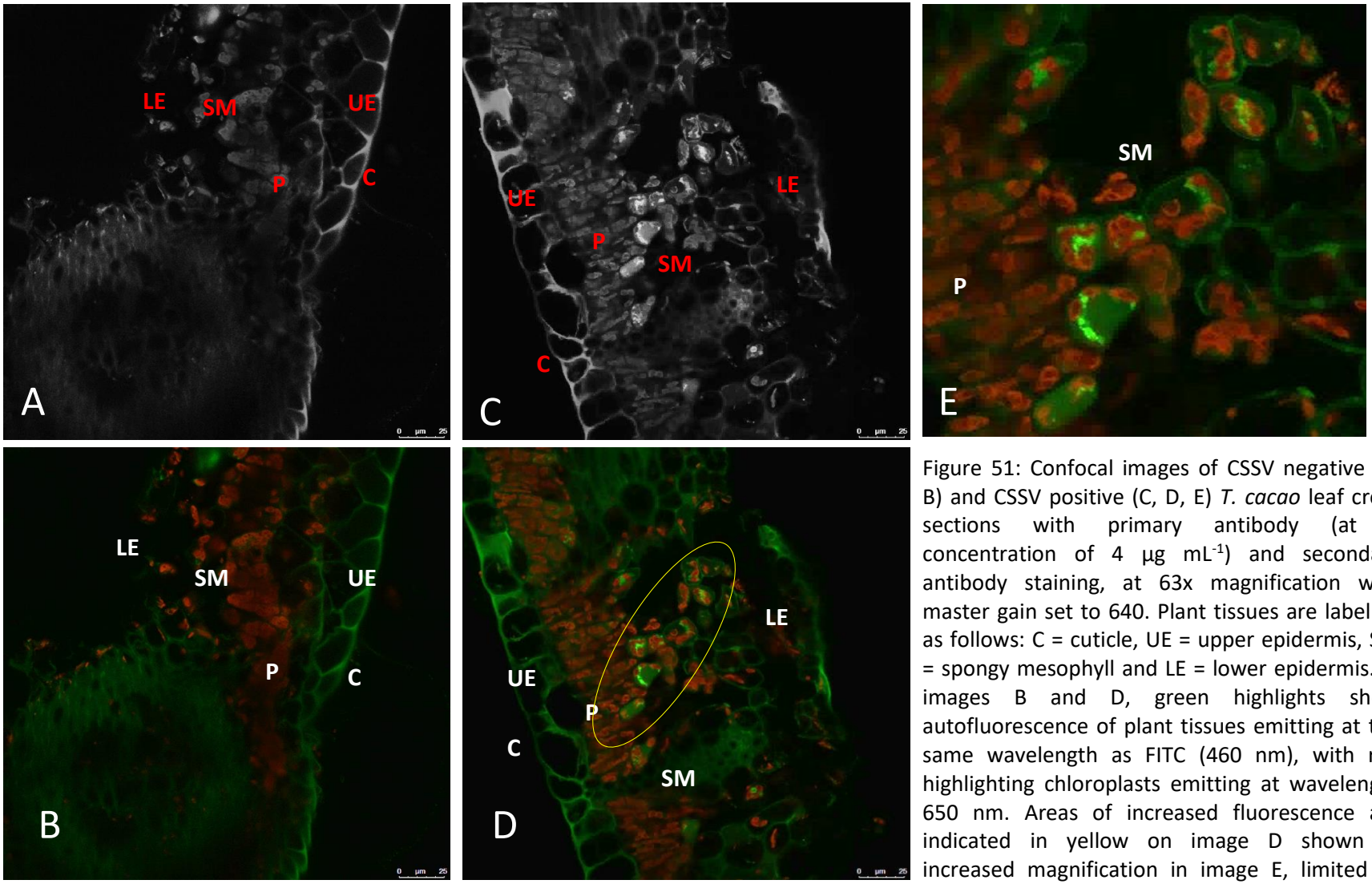


Figure 51: Confocal images of CSSV negative (A, B) and CSSV positive (C, D, E) *T. cacao* leaf cross sections with primary antibody (at a concentration of 4 µg mL^{-1}) and secondary antibody staining, at 63x magnification with master gain set to 640. Plant tissues are labelled as follows: C = cuticle, UE = upper epidermis, SM = spongy mesophyll and LE = lower epidermis. In images B and D, green highlights show autofluorescence of plant tissues emitting at the same wavelength as FITC (460 nm), with red highlighting chloroplasts emitting at wavelength 650 nm. Areas of increased fluorescence are indicated in yellow on image D shown in increased magnification in image E, limited to

5.3.4 Does the concentration of the secondary antibody have an effect on signal brightness and/or background autofluorescence from plant tissues?

Figure 52, Figure 53 and Figure 54 depict both CSSV positive and CSSV negative *T. cacao* leaf samples when prepared as described in section 5.2.5, but with different concentrations of secondary antibody used in the staining process. Concentrations of 0.01 µg mL^{-1}, 0.1 µg mL^{-1} and 1 µg mL^{-1} were chosen for comparison.

In Figure 52, two samples prepared using a secondary antibody concentration of 0.01 µg mL^{-1} are shown. Image B depicts a CSSV negative sample of *T. cacao* leaf tissue, where some autofluorescence can be seen limited to the upper and lower epidermis. A sample positive for CSSV depicted in image D (Figure 52) also appears to emit some autofluorescence, around the lower and upper epidermis. In image D (Figure 52), bright spots of increased fluorescence in the FITC wavelength (460 nm) can be seen within the spongy mesophyll.

Figure 53 depicts two samples prepared using a secondary antibody concentration of 0.1 µg mL^{-1} are shown. Image B depicts a CSSV negative sample of *T. cacao* leaf tissue, where very little autofluorescence can be seen limited to the cuticle. A sample positive for CSSV depicted in image D (Figure 53) also appears to emit some autofluorescence around the cuticle. In image D (Figure 53), bright spots of increased fluorescence at 460 nm can be seen within the spongy mesophyll.

Figure 54 depicts two samples prepared using a secondary antibody concentration of 1 µg mL^{-1} are shown. Image B (Figure 54) depicts a CSSV negative sample of *T. cacao* leaf tissue, where very little autofluorescence can be seen limited to the cuticle. A sample positive for CSSV depicted in image D (Figure 54) also appears to emit some autofluorescence around the cuticle and lower epidermis. In image D (Figure 54), bright spots of increased fluorescence in the FITC wavelength (460 nm) can be seen within the spongy mesophyll.

There are no discernible differences in the levels of fluorescence detected at 460 nm in negative samples (Figure 52, Figure 53 and Figure 54; B) when comparing between the three different concentrations of secondary antibody. The images taken of CSSV positive samples prepared with a secondary antibody concentration of 0.1 µg mL^{-1} (Figure 53 B and Figure 54 D) appear darker when compared to images of CSSV positive samples prepared using 0.01 µg mL^{-1} and 1 µg mL^{-1} (Figure 52 and Figure 53; B and D). The areas of increased fluorescence in the FITC wavelength all appear in the same areas across all Figure 52, Figure 53 and Figure 54, in the spongy mesophyll.

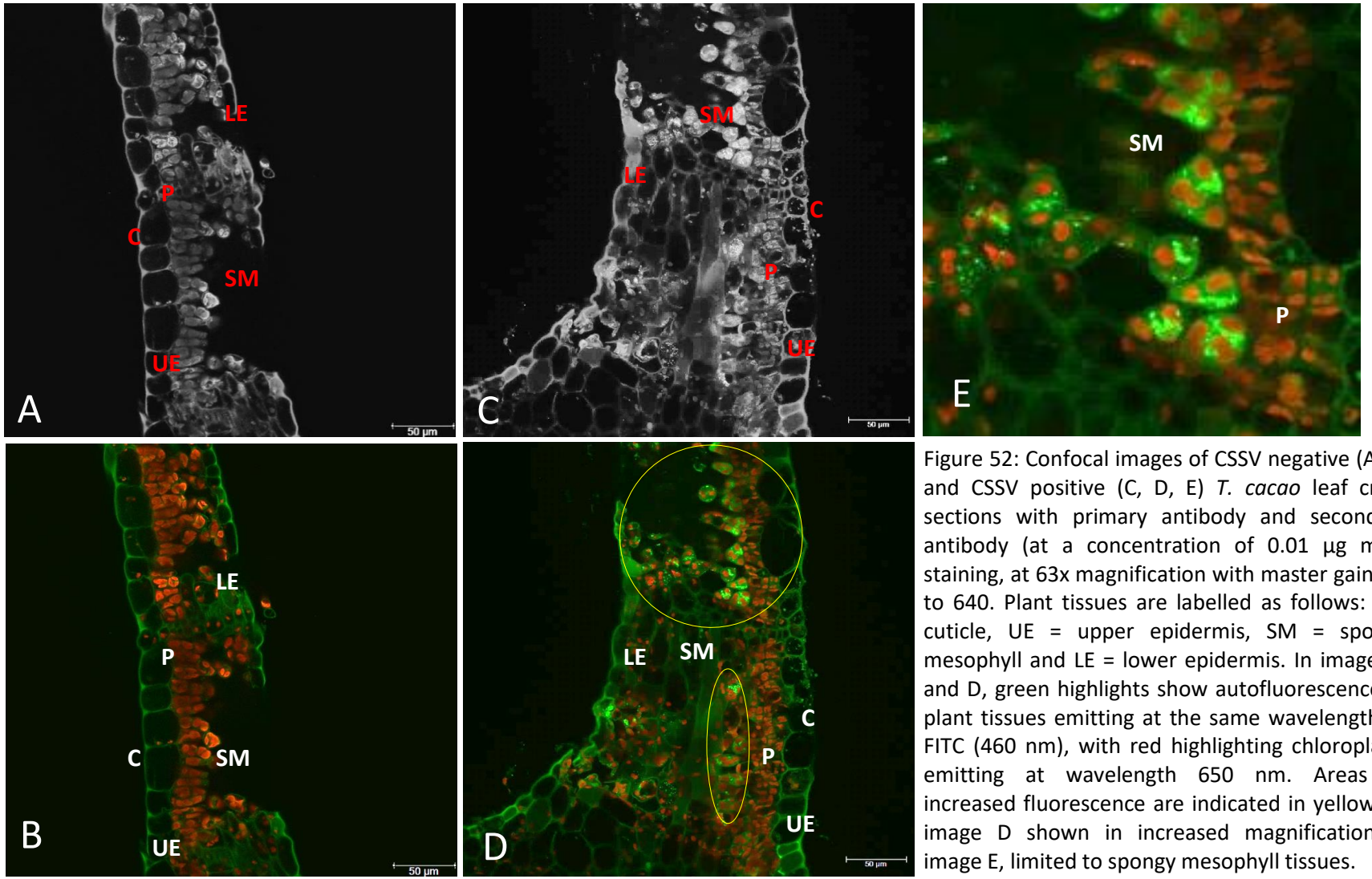


Figure 52: Confocal images of CSSV negative (A, B) and CSSV positive (C, D, E) *T. cacao* leaf cross sections with primary antibody and secondary antibody (at a concentration of 0.01 µg mL^{-1}) staining, at 63x magnification with master gain set to 640. Plant tissues are labelled as follows: C = cuticle, UE = upper epidermis, SM = spongy mesophyll and LE = lower epidermis. In images B and D, green highlights show autofluorescence of plant tissues emitting at the same wavelength as FITC (460 nm), with red highlighting chloroplasts emitting at wavelength 650 nm. Areas of increased fluorescence are indicated in yellow on image D shown in increased magnification in image E, limited to spongy mesophyll tissues.

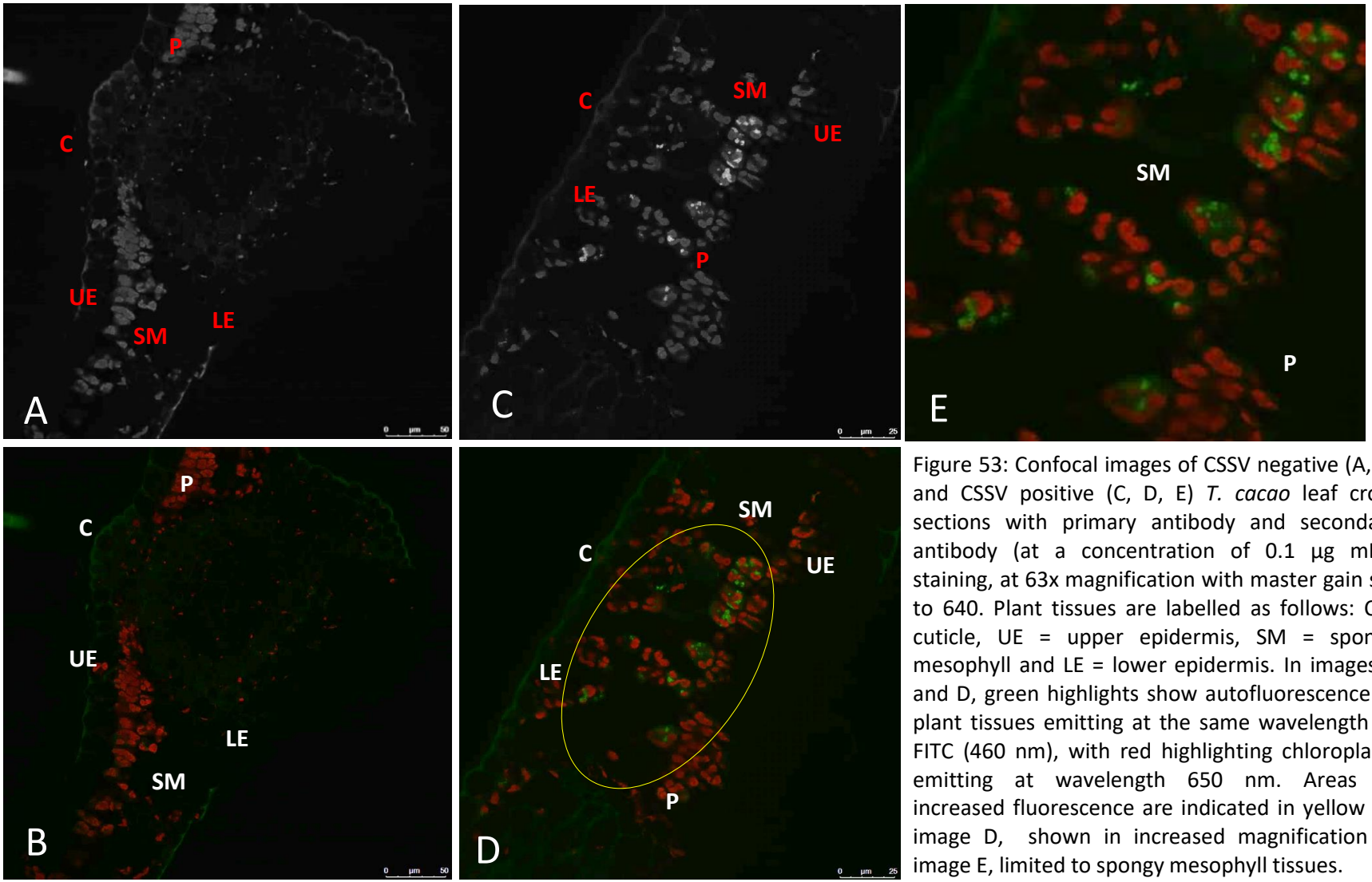


Figure 53: Confocal images of CSSV negative (A, B) and CSSV positive (C, D, E) *T. cacao* leaf cross sections with primary antibody and secondary antibody (at a concentration of 0.1 µg mL^{-1}) staining, at 63x magnification with master gain set to 640. Plant tissues are labelled as follows: C = cuticle, UE = upper epidermis, SM = spongy mesophyll and LE = lower epidermis. In images B and D, green highlights show autofluorescence of plant tissues emitting at the same wavelength as FITC (460 nm), with red highlighting chloroplasts emitting at wavelength 650 nm. Areas of increased fluorescence are indicated in yellow on image D, shown in increased magnification in image E, limited to spongy mesophyll tissues.

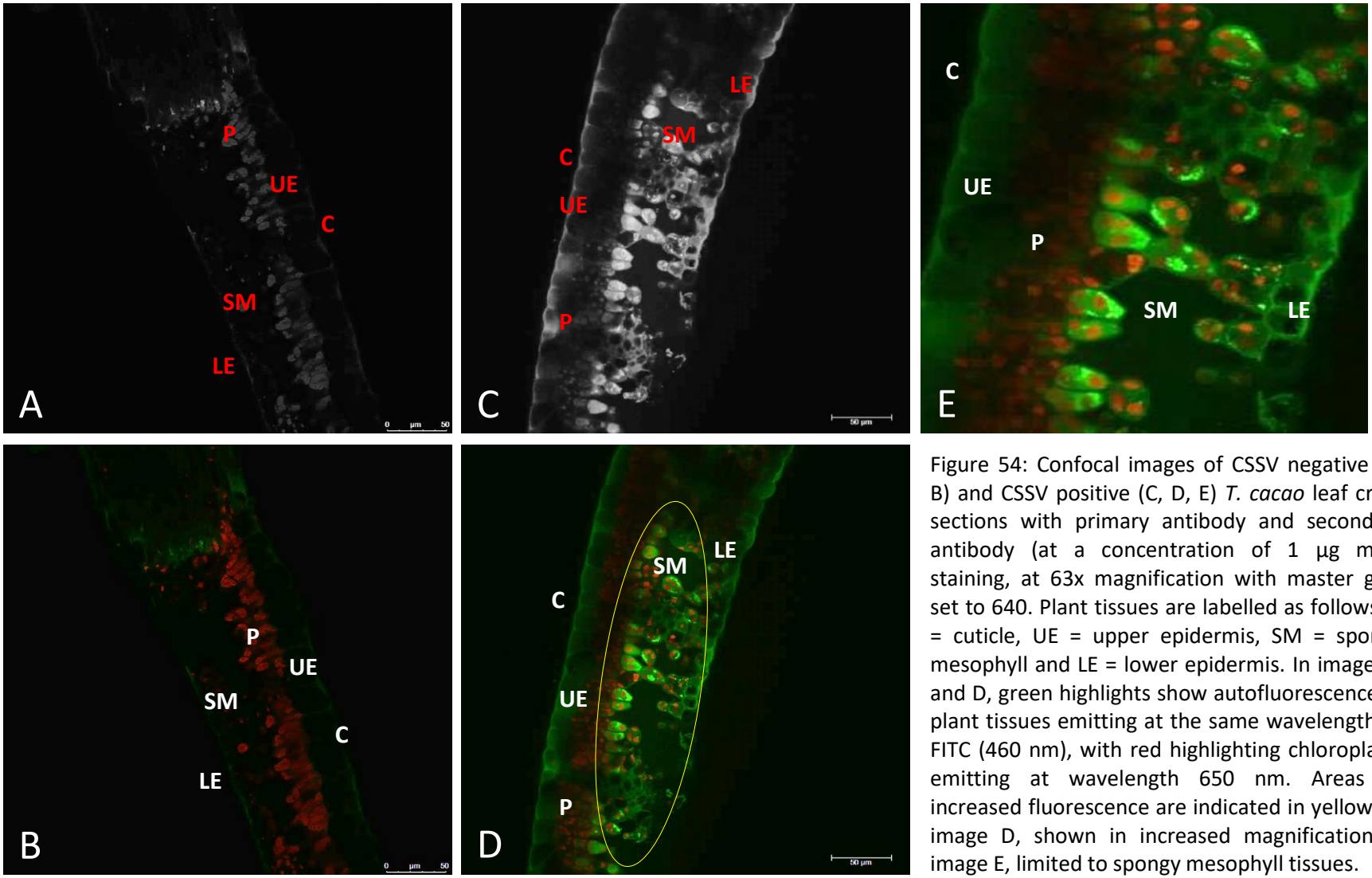


Figure 54: Confocal images of CSSV negative (A, B) and CSSV positive (C, D, E) *T. cacao* leaf cross sections with primary antibody and secondary antibody (at a concentration of 1 µg mL^{-1}) staining, at 63x magnification with master gain set to 640. Plant tissues are labelled as follows: C = cuticle, UE = upper epidermis, SM = spongy mesophyll and LE = lower epidermis. In images B and D, green highlights show autofluorescence of plant tissues emitting at the same wavelength as FITC (460 nm), with red highlighting chloroplasts emitting at wavelength 650 nm. Areas of increased fluorescence are indicated in yellow on image D, shown in increased magnification in image E, limited to spongy mesophyll tissues.

5.4 Discussion

5.4.1 Sample background fluorescence and non-specific binding of fluorescently tagged secondary antibodies

Plant tissues are abundant in fluorescent molecules, such as chlorophyll, alkaloids, tannins and polyethylenes to name just a few (Figure 55, García-Plazaola *et al*., 2015). The presence of these fluorescent molecules can be both helpful and problematic when imaging plant tissues under laser scanning confocal microscopes. Molecules such as chlorophyll and lignin can act as markers for plant cells (Donaldson, 2020), and have been utilised in this study to aid in the identification of cell types. However, when using a fluorescent probe with a similar excitation spectrum to these molecules, it can be difficult to distinguish between increased fluorescent probe and autofluorescence from these molecules. The fluorescence emitted by these molecules can decrease the apparent fluorescent yield of the fluorophore. It is important to choose a fluorophore carefully to ensure that autofluorescence is minimised (García-Plazaola *et al*., 2015).

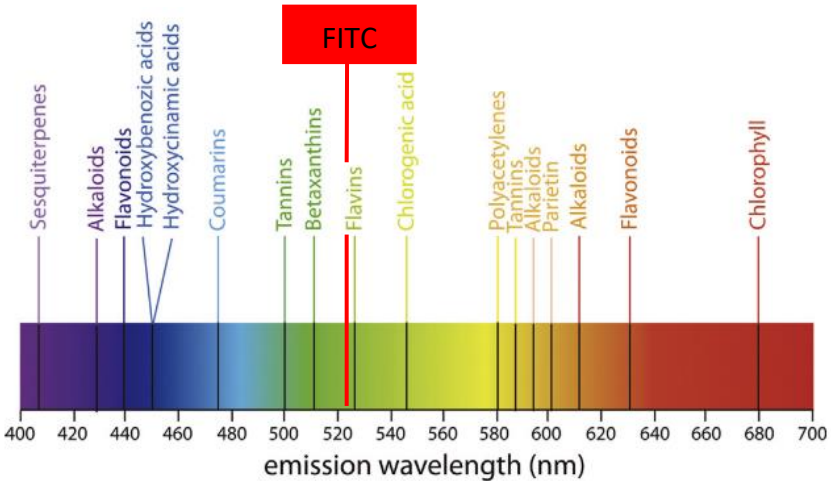


Figure 55: Emission spectrum of naturally occurring fluorochromes in plants. Taken from García-Plazaola *et al*., 2015. The emission wavelength of the FITC fluorophore is highlighted in red, at 525 nm.

In order to assess photosynthesis in a non-invasive way, researchers utilise the fluorescent properties of chlorophyll. In this specific case, the chlorophyll fluorescence was used as an indication of chloroplasts within cells. The presence of

chlorophyll in samples can be seen in all Figures in section 5.3 Results (B and D), highlighted in red. Utilising the fluorescence properties of chlorophyll provided valuable information about sample cell structure and aided in the identification of cell types with increased fluorescence emitting between 520 nm and 545 nm for FITC.

To determine the effects of plant auto-fluorescence on the detectability of the FITC fluorophore, untreated samples were imaged under the confocal microscope using both the 440 nm and 488 excitation rays for both chlorophyll and FITC excitation. An overwhelming fluorescence in the sample would present issues for subsequent imaging on treated samples. Figure 45 depicts four confocal images of fresh, unstained *T. cacao* leaf tissues when viewed using the above excitation rays. Chlorophyll can be clearly detected in red, emitting between 675 nm and 685 nm. Some emission between the wavelengths of 520 nm and 545 nm, for the detection of the FITC fluorophore, can also be seen highlighted in green. This is likely to be the auto-fluorescent properties of the flavin molecules, a component of cell walls (Donaldson, 2020), based on the location of the fluorescence surrounding cells (Figure 45, B and D; fluorescence limited to the upper and lower epidermis). Specifically, the cuticles associated with the epidermal layers have been found to exhibit intrinsic autofluorescence due to component flavonoids such as chalconaringenin (Buda *et al.,* 2009). The presence of this auto-fluorescence had no particular impact on the quality of the image or interpretation of the overall results, due to its low intensity; Badnaviruses are considered to be limited to the vascular system and surrounding tissues and have not been detected in the upper or lower epidermis. While the mealybug vector stylets penetrate the epidermis, the semi-persistent nature of CSSV, whereby the virions are present in the gut and passage into the plant through saliva, suggests that the virus would not enter *T. cacao* tissues until feeding begins (Torbert *et al*., 1998; Jacquot *et al*., 1999; Gutiérrez *et al*., 2013; Bhat *et al*., 2023).

The potential for non-specific binding of the fluorescently tagged secondary antibody was identified by staining a fresh *T. cacao* leaf sample with secondary antibody only. it is imperative that measures are taken to avoid non-specific binding

of fluorescently tagged secondary antibodies, as the increase in fluorescence seen on confocal images are indistinguishable from the primary-secondary antibody conjugate and can result in false positives (Jonkman *et al*., 2020). In some cases, a blocking buffer is applied; this helps to prevent antibodies from binding to non-target epitopes. Blocking reagents include bovine serum albumin, non-fat dry milk and gelatin (Im *et al*., 2019). In addition, the incubation time and concentration of antibodies can be amended to reduce background signals and produce the clearest possible images (Jonkman *et al*., 2020). Images were taken, and a representative image showing the results of the selection is depicted in Figure 46. In this case no non-specific binding of the secondary antibody could be identified, illustrated by the lack of fluorescence in the FITC emission wavelength (520 - 545 nm).

Finally, the protocol was performed as described in 5.2.5 with no amendments; both primary and secondary antibodies were utilised in the staining procedure. As in Figure 45, where the sample was prepared in absence of any antibodies, some low-level background fluorescence can be seen around the upper epidermis and in patches around the spongy mesophyll of the CSSV negative sample (Figure 47, image B). This suggests that some excess antibody is present, and that an adjustment to the concentrations of the antibodies used might be necessary (Jonkman *et al*., 2020). In comparison, small amounts of background fluorescence can be identified in the CSSV positive sample (Figure 47, image D) with some areas of brighter fluorescence in the FITC emission wavelength present in the spongy mesophyll (highlighted in yellow). This could be remedied by an enhanced washing step, which would lead to less noise present in images.

5.4.2 Does the preparation of the sample have an effect on the sample and image quality?

Sample preparation is one of the most important things to consider when undertaking any form of microscopy. In this case, the use of a high-performance microscope increases the likelihood of revealing inadequate sample preparation

(Jonkman *et al*., 2020). Important sample processing steps, such as staining and mounting were optimised in order to attain high quality images.

5.4.2.1 Does mounting procedure have an effect on sample and image quality?

The process of moving a sample onto its microscope slide ready for imaging is known as mounting. Inappropriate mounting of a sample can lead to loss of 3D information and loss of signal, which are imperative for morphology and colocalization analysis (Jonkman *et al*., 2020). The equilibration of tissue in mounting medium prior to imaging helps to reduce light-scattering air pockets; the longer a sample has to incubate in mounting media the longer the mounting media has to penetrate the tissues. In some cases, a long incubation time can stress the tissues, leading to an increase in secondary compounds that may interfere with microscopy (Dumur *et al*., 2019).

In order to explore the effect of post-mounting incubation time for fresh, stained *T. cacao* leaf tissues, two incubation periods of one-hour and 12-hours were tested. Tissue integrity and fluorescence intensity were compared using confocal microscopy. Background fluorescence present in Figure 48 images B and D, the CSSV-negative and CSSV-positive 12-hour incubated samples, appears to be stronger in intensity than that of the background fluorescence in images A and B, the CSSV-negative and CSSV-positive one-hour incubated samples. This may indicate that the increased incubation period has stressed the leaf tissues resulting in secondary compounds (Dumur *et al*., 2019) which fluoresce at the same wavelength as the FITC fluorophore. While there are more signals of increased fluorescence indicating CSSV presence detected in the spongy mesophyll of the one-hour incubation samples (Figure 48, image C) compared to the 12-hour incubation samples (Figure 49, image D), the difference in concentration of these bright spots may be due to the sampling location rather than an effect of the mounting time.

Based on these results it was determined that a shorter post-mounting incubation time of one-hour is preferable for imaging these fresh *T. cacao* leaf samples. The shorter incubation time reduces the likelihood of secondary compounds accumulating and resulting in an increase in background fluorescence, which has a detrimental effect on the quality of the confocal images.

5.4.2.2 Reducing non-specific binding by altering primary and secondary antibody concentrations

Non-specific binding of antibodies can be reduced in different ways, including altering their concentrations prior to staining, amending incubation times and also using a blocking agent. Blocking agents such as Bovine Serum Albumin (BSA) or gelatin act by blocking non-target reactive sites in the tissues (Jonkman *et al.,* 2020). This leads to competition between the blocking agent and antibodies when binding to reactive sites, and thus reduces the chances of antibodies binding to non-target reactive sites (Im *et al*., 2019). In this case, antibody concentrations were changed in order to achieve the best possible image in terms of fluorescence intensity from target sites with low to no non-specific binding from either primary or secondary antibody. In order to establish the appropriate antibody concentrations, three different concentrations of both primary and secondary antibody were applied to CSSV positive and CSSV negative *T. cacao* samples and imaged using confocal microscopy.

In section 5.3.3, the primary antibody concentration was changed while the secondary antibody concentration was kept the same. Concentrations of 4 µg mL^{-1}, 1 µg mL^{-1} and 0.1 µg mL^{-1} were chosen for testing. A survey of available plant tissue immunostaining protocols showed that the most common starting primary antibody concentration is 1 µg mL^{-1} , therefore an upper and lower concentration were also tested for comparison. A similar level of background fluorescence was detected in all three samples (Figure 49, Figure 50 and Figure 51) with differing concentrations of primary antibody. An increase the abundance of intense areas of fluorescence compared to the 0.1 µg mL^{-1} (Figure 49) sample was observed in the 1 µg mL^{-1} (Figure 50) and 4 µg mL^{-1} (Figure 51) samples, indicating CSSV presence.

This suggests that a primary antibody concentration of 0.1 µg mL^{-1} is not sufficient for the concentration of CSSV particles present, and that a higher concentration of primary antibody allows for more targeted binding and therefore fluorescence. An alternative argument could be that the lack of intense fluorescence in the 0.1 µg mL^{-1} sample is attributed to the quality and/or area of the sample that was imaged. While all of the samples were taken from the same leaf sections, the concentration of CSSV particles in those sections could be heterogeneous, resulting in some samples with an abundance of high intensity fluorescence due to binding to large concentrations of CSSV particles, and some with fewer areas of fluorescence due to lower concentrations of CSSV particles. It has been previously demonstrated that viruses such as *Cauliflower mosaic virus* (CaMV) and *Melon necrotic spot virus* are capable of travelling long distances within their host plants, from cotyledons to roots to the shoot apex (Roberts *et al*., 2007; Gosalves-Bernal *et al*., 2008; Hipper *et al*., 2013).

The effect of secondary antibody concentration on the signal brightness and/or background fluorescence of *T. cacao* samples was investigated in a similar fashion to the concentration of the primary antibody. The supplier suggested concentration of secondary antibody for use in microscopy is 0.1 µg mL^{-1} (Invitrogen, UK), and thus confocal images of samples stained using concentrations of 1 µg mL^{-1}, 0.1 µg mL^{-1} and 0.01 µg mL^{-1} were compared. Background fluorescence is present in all three samples (Figure 52, Figure 53 and Figure 54) stained using different concentrations of antibody. There is a stark difference in the brightness of the images taken of the sample stained using 0.1 µg mL^{-1} secondary antibody (Figure 53), possibly caused by physiological differences in the area of *T. cacao* leaf from which the sample was acquired. As these samples were prepared and imaged fresh, the quality of the samples varied despite being sourced from the same leaf location. It is unlikely that this variation can be attributed to the change in secondary antibody concentration, as if that were the case it would be expected that the result from the sample prepared with the lowest concentration of secondary antibody (Figure 52, secondary antibody concentration of 0.01 µg mL^{-1}) would result in a similar dark image, although this is not the case.

Sample preparation could have been improved upon through fixation steps and blocking steps to aid with antibody binding (Jonkman *et al*., 2020). Although, the increase in fluorescence due to antibody binding to viral particles was enough to deem it unnecessary to implement these sample preparation steps. Time restraints were also considered when deciding on whether to fix samples, as this would require lengthy optimisation procedures and the use of hazardous chemicals. While the presence of background fluorescence has been attributed to the sample preparation, the potential for CSSV spread during the quarantine process cannot be ruled out as a possibility, which could provide an alternative explanation for the 'background' fluorescence present in the negative samples. This however seems unlikely, as there have been no recorded incidences of insect pests present in the ICQC, and the spread of CSSV can only be achieved through mealybug feeding (Ameyaw *et al*., 2014; Blanc *et al*., 2014).

5.4.3 Location of CSSV viral particles in *T. cacao* leaf tissues

Using a novel antibody which binds to the CSSV coat protein antigen, an indirect staining method using a secondary goat anti-human IgG antibody with conjugated FITC fluorophore was successfully implemented to identify the location of CSSV particles in *T. cacao* leaf tissues. *T. cacao* leaves with visible symptoms of CSSV were chosen for these experiments, with sectioning focussing on areas of chlorosis; previous studies by Sagemann *et al*. (1985) using ELISA identified a higher concentration of CSSV particles present in areas where leaf chlorosis was present. CSSV particles were found localised to spongy mesophyll tissues, a contrasting result to previous studies which used TEM to identify presence of CSSV particles in phloem companion cells (Jacquot *et al*., 1999). This study by Jacquot *et al*. (1999) did use agroinfected *T. cacao* leaf tissues, opposed to the study presented above which used mealybug inoculated *T. cacao*, giving a more representative comparison to a wild-type CSSV infection. Obok *et al*. (2018) describe that *P. citri* spend the majority of their feeding time probing for and establishing stylet connection with vascular tissues. Considering the results of Obok *et al*. (2018), and this confocal study, there is evidence to suggest that uptake of CSSV particles by mealybugs can

take place in the spongy mesophyll tissues of *T. cacao*. This method of viral plant pathogen by insect vector has been previously described by Moreno *et al*. (2005), who used EPG analysis to study the aphid stylet activities and behaviours that lead to the inoculation of CaMV in turnip plants. CaMV has been described as preferentially acquired from the phloem, similar to CSSV. However, Moreno *et al*. discovered that the virus can be acquired after brief intracellular stylet punctures in non-vascular leaf tissues, such as spongy mesophyll, before the stylets reach the phloem. Considering previous results described in this thesis, for example in chapter three, where CSSV DNA was successfully detected in and/or around the vasculature of *T. cacao* stems, it is possible that CSSV particles can pass cell-to-cell from the spongy mesophyll tissues until they reach the vasculature, where they spread and infect the entire plant (Hipper *et al*., 2013).

CHAPTER SIX

General discussion

6.1 Chapter 3: Development of a method for the extraction and detection of CSSV DNA from *T. cacao* stem exudate

Cocoa quarantine processes undertaken by the ICQC often involve lengthy procedures such as grafting and symptom observation with which to identify presence of viral pathogens, such as CSSV. The process for screening germplasm for CSSV involves grafting onto a *T. cacao* Amelonado root stock and observation over the course of two years for identification of symptoms typical of CSSV infection. The two-year observation is a necessity due to the long latency period of the virus. While there are continuing advancements in molecular biology, an isolation-independent CSSV detection technique has not yet been produced for use in quarantine centres (Daymond, 2018; End *et al*., 2021). CSSV has long been characterised as a phloem limited pathogen (Jacquot *et al*, 1999). The method developed in chapter 3 has been used to successfully extract and detect CSSV DNA from phloem-disrupted stem samples; a method that could feasibly be used to isolate CSSV DNA from incoming quarantine budwood samples.

The suitability of this method for use in quarantine centres comes from the ability to detect CSSV DNA from stem exudate without the need for DNA extraction, which can be time consuming and expensive. The incubation of stem samples in water provides a labour-free method for extraction, the result of which can be used in a diluted form to detect CSSV-positive samples using qPCR. The dilution of the resulting stem exudate is an imperative step, as it aids in the dilution of qPCR inhibitors which result in false-positive results. Once diluted, CSSV DNA can be detected by qPCR in samples across six technical replicates. An important consideration for this work was that the CSSV DNA was still detectable in stem samples following a week in storage at 20 °C ± 2 °C. This is to account for time that samples spend in transport from country or origin to the ICQC. Results indicate that although CSSV DNA concentrations are lower in samples that had been stored, they were still well within detectable levels for qPCR. This method was also used to study the presence of CSSV DNA in each component of the stem. SEM images revealed

that the outer peel contained the phloem tissue, where CSSV has previously been detected. In one experiment, the outer peel was removed completely from the inner heartwood core and incubated separately. Result showed that the majority of CSSV DNA was detected in the outer peel, further adding to the evidence that CSSV is associated with phloem tissues.

The reliability of this method negates the need for lengthy grafting and monitoring procedures currently in place for detection of CSSV in *T. cacao* germplasm samples sent through the ICQC. An additional benefit comes from the detection method which can be amended and improved on to provide more specific identification of CSSV species, using data generated by Muller *et al*. (2018). A different method of detection such as HRM could also be implemented, following successful optimisation, which would speed up the detection process even further.

6.2 Chapter 4: Identification of a novel badnaviral sequence integrated in the genome of *T. cacao*: a HRM method to distinguish between insert types and modes of inheritance

During the experimental process, a collaboration between colleagues in The French Agricultural Research Centre for International Development (CIRAD; Montpellier, France), the University of Reading (Reading, UK) and the University of the West of England (Bristol, UK) resulted in the identification of novel badnaviral sequences integrated into the genome of *T. cacao*. Following the routine screening of the ICQC *T. cacao* clones collection for Badnaviral species using specific primers designed for identification of different CSSV species (Muller *et al*., 2018), a number of positive PCR results were obtained. None of the ICQC clones have exhibited symptoms typical of Badnavirus infection during six monthly inspections over the past eight years. Furthermore, the non-clustered positions of many of the PCR positive cocoa clones within the ICQC was not consistent with the 'crowd disease' nature of CSSVD, a disease that has not been observed to spread quickly or far (Thresh and Fargette, 2001). The theory that the positive PCR results may be due to an integrated Badnaviral sequence in the genome of cocoa was considered, first

suggested when 'species S' was discovered by Muller *et al*. in 2018. Subsequently, an article was published outlining the presence of 12 different integrated Badnaviral sequences, named eTcBVs in genomes of *T. cacao*. During this time, colleagues from the ICQC were involved in the deliberate crossing of *T. cacao* clones harbouring these different insert types for use in inheritance studies outlined in chapter 4. Transmission studies provided further evidence for the integration of these sequences in the genome of *T. cacao*, whereby transmission of sequences was not possible through mealybug feeding, the method by which CSSV is transmitted (Strickland, 1951)

Two of the 12 eTcBV insert types were used in this project: type IV, belonging to parent plant ICS40 and type VI, belonging to parent plant EET183. A HRM method for the detection of these different insert types was successfully developed, which provides a quick and straightforward method to distinguish between eTcBVs and CSSV infections. This HRM method was used to determine the patterns of heritability of these two different insert types. It was determined that they are inherited in a typical Mendelian fashion, with plant ICS40 containing two inserts of the type IV eTcBV. It is not uncommon for plants to contain multiple of the same insert type; multiple EVE inserts have previously been identified in the genomes of other virus host plants, such as *N. tabacum* and *Musa spp.* (Bejarano *et al.*, 1996; Staginnus and Richert-Pöggeler, 2006). While most EVEs remain dormant in plant genomes (Vassilieff *et al*., 2023), there are a few examples where EVEs containing near full-length viral genomes can activate and cause disease like symptoms (Staginnus and Richert-Pöggeler, 2006). Studies have also suggested that in some cases, EVEs can confer resistance to their exogenous counterparts (Mette *et al.,* 2002; Umber *et al*., 2016).

The results outlined in chapter 4 have shown that the HRM method developed is able to successfully distinguish between eTcBV insert types IV and VI. It has the potential to enhance knowledge of the remaining 10 eTcBV insert types discovered so far and will provide a highly specific and rapid screening technique (Reed *et al.*, 2007; Tong and Giffard, 2012) for use in quarantine and research situations, where distinguishing between CSSV infection and eTcBV presence is vital.

A question remains over the reason for these integrated sequences to be maintained within the genome throughout generations; it is likely that the insertion is beneficial to the host, providing some positive impact on competitiveness, otherwise one would expect degeneration of the insert over time and eventually, elimination (Mette *et al.,* 2002; Feschotte and Gilbert, 2012; Umber *et al*., 2016). The identification of these EVEs within plant genomes could provide valuable insights into the arms race between plants and their pathogens.

6.3 *In situ* localisation of CSSV within *T. cacao* leaf tissues

Finally, chapter 5 outlines the process of developing an antibody-based staining method for the detection of CSSV viral particles in CSSV-positive *T. cacao* leaf tissues. Using a novel CSSV antibody developed by Dr Jacqui Barnett and Sue Taylor (unpublished), and a secondary antibody with FITC fluorophore, an indirect staining method was tested and developed. *T. cacao* leaves with visible symptoms of CSSV were chosen for these experiments, with sectioning focussing on areas of chlorosis (Sagemann *et al*., 1985).

Autofluorescence from plant molecules proved a problem when developing this method. Plants are abundant in fluorescent molecules which can interfere with laser scanning microscopy techniques (García-Plazaola *et al*., 2015). In this case, the likely culprit of auto-fluorescence in samples of *T. cacao* leaf tissue are due to the flavin molecules, a component of cell walls (Donaldson, 2020). Despite the background fluorescence, areas of increased fluorescence in the FITC wavelength were imaged in almost all CSSV-positive *T. cacao* leaf samples stained with both antibodies, which were suggested to be the presence of CSSV virions.

CSSV particles were found consistently localised to spongy mesophyll tissues, a contrasting result to previous studies which used TEM to identify presence of CSSV particles in phloem companion cells (Jacquot *et al*., 1999). This study by Jacquot *et al*. (1999) did however use an agroinfection method, opposed to the study presented above which used mealybug inoculated *T. cacao*, giving a more representative comparison to a wild-type CSSV infection. Obok *et al*. (2018)

describe that *P. citri* spend the majority of their feeding time probing for and establishing stylet connection with vascular tissues, which when considered with the results presented in chapter 4 are analogous to CaMV. It was discovered that CaMV, which has been described as preferentially acquired from the phloem, can be acquired by its aphid vector after brief intracellular stylet punctures in non-vascular leaf tissues, such as spongy mesophyll, before the stylets reach the phloem (Moreno *et al.*, 2005). Considering the results described in chapter 3 where CSSV DNA was successfully detected in and/or around the vasculature of *T. cacao* stems, it is possible that CSSV particles can pass cell-to-cell from the spongy mesophyll tissues until they reach the vasculature, where they spread and infect the entire plant (Hipper *et al*., 2013). Considering previous results described in this thesis, for example in chapter 3 where CSSV DNA was successfully detected in and/or around the vasculature of *T. cacao* stems, it is possible that CSSV particles can pass cell-to-cell from the spongy mesophyll tissues until they reach the vasculature, where they spread and infect the entire plant (Hipper *et al*., 2013).

Considering previous results described in this thesis, for example in chapter 3 where CSSV DNA was successfully detected in and/or around the vasculature of *T. cacao* stems, it is possible that CSSV particles can pass cell-to-cell from the spongy mesophyll tissues until they reach the vasculature, where they spread and infect the entire plant (Hipper *et al*., 2013).

6.4 Final conclusions and recommendations for future work

The overall aims of this thesis were to further the understanding of CSSV pathogenesis and in doing so, develop new methods to aid in cocoa quarantine procedures. In this project, a new method for the quick and easy detection of CSSV DNA using samples of CSSV-positive *T. cacao* stem has been described for utilisation in quarantine scenarios. Further work on this method, for example to optimise a multiplex qPCR method for the detection of multiple CSSV species, a HRM method for an even quicker diagnosis, and the potential for this method to be used to extract DNA from other pathogens in a similar manner will go long way to improving on the current methods for *T. cacao* quarantine. A HRM method for the

detection and distinction between two eTcBVs has been developed, a method which also has the potential to provide valuable insights into the remaining 10 identified insert types and subsequently the evolution of these EVEs within *T. cacao* genomes. Finally, the location of CSSV virions were successfully imaged *in situ* using immunofluorescent confocal laser scanning microscopy. This method could be improved through the use of fixation steps and blocking solutions to reduce the presence of auto-fluorescent molecules, providing clearer images. There is also potential for further work using this method to identify the location of CSSV virions within the mealybug vector. An *in-situ* hybridisation method for detection of CSSV DNA should also be considered, which would provide further valuable information about the replication of CSSV *in planta*.

Quarantine procedures for cocoa germplasm could still be improved on further. For example, International Organization for Standardisation (ISO) certification could be implemented through a national accreditation body (e.g. United Kingdom Accreditation Service, UKAS) to ensure technical competence is achieved and that quarantine facilities adhere to recognised international standards. Alternative methods of pathogen diagnosis could also be explored, such as the use of Nanopore sequencing for the detection of plant viruses (Liefting *et al*., 2021; Sun *et al*., 2022). Recent advances in CRISPR-Cas based platforms could be investigated; these have been shown to provide a rapid and sensitive method for the detection of very low concentrations of nucleic acids with single-base mismatch specificity (Tripathi *et al*., 2020). A one-step CRISPR-Cas system has already been employed as a detection assay for diagnosis of plant RNA viruses in under 30 minutes at a single temperature in the field (Aman *et al*., 2020).

While initial costs involved in purchasing equipment, facilities and training may be costly, use of these molecular methods and updating the existing quarantine measures should be considered. Through the use of the molecular diagnostic techniques outlined in this thesis, quarantine time could be shortened with costs reduced. Along with further research into emerging methods for pathogen diagnostics, the transportation of cocoa germplasm across the globe will be improved and lead to greater advances in cocoa research.

REFERENCES

Abrokwah, F. *et al*., 2016. Geographical distribution of Cacao swollen shoot virus molecular variability in Ghana. *Plant Disease,* Volume 100, p. 2011–2017.

Ameyaw, G. A. *et al*., 2016. Appraisal of cocoa swollen shoot virus (CSSV) mild isolates for cross protection of cocoa against severe strains in Ghana. *Plant Disease,* Volume 100, p. 810–815.

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

Andres, C. *et al*., 2017. Combatting Cocoa Swollen Shoot Virus Disease: What do we know?. *Crop Protection,* Volume 98, p. 76–84.

Andret-Link, P. & Fuchs, M., 2005. Transmission specificity of plant viruses by vectors. *Journal of Plant Pathology,* Volume 87, p. 153–165.

Anning, A. K., Ofori-Yeboah, A., Baffour-Ata, F. & Owusu, G., 2022. Climate change manifestations and adaptations in cocoa farms: Perspectives of smallholder farmers in the Adansi South District, Ghana. *Current Research in Environmental Sustainability,* Volume 4, p. 100196.

Balan, S., Nithya, K., Cherian, K. A. & Viswanathan, R., 2022. True Seed Transmission of Sugarcane bacilliform virus (SCBV) in Sugarcane. *Sugar Tech,* Volume 24, p. 513–521.

Bejarano, E. R., Khashoggi, A., Witty, M. & Lichtenstein, C., 1996. Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution. *Proceedings of the National Academy of Sciences of the United States of America,* Volume 93, p. 759–764.

Bhat, A. I., Hohn, T. & Selvarajan, R., 2016. Badnaviruses: The current global scenario. *Viruses,* Volume 8.

Bhattacharjee, R., 2018. Achieving sustainable cultivation of cocoa Edited by Professor Pathmanathan Umaharan. *Burleigh Dodds Science,* p. 1 – 18.

Blanc, S., Drucker, M. & Uzest, M., 2014. Localizing viruses in their insect vectors. *Annual Review of Phytopathology,* Volume 52, p. 403–425.

Boutanaev, A. M. & Nemchinov, L. G., 2021. Genome-wide identification of endogenous viral sequences in alfalfa (Medicago sativa L.). *Virology Journal,* Volume 18, p. 1–5.

Bradamante, G., Scheid, O. M. & Incarbone, M., 2021. Under siege: Virus control in plant meristems and progeny. *Plant Cell,* Volume 33, p. 2523–2537.

Bragard, C. *et al*., 2013. Status and prospects of plant virus control through interference with vector transmission. *Annual Review of Phytopathology,* Volume 51, p. 177–201.

Chabannes, M. *et al*., 2021. Badnaviruses and banana genomes: a long association sheds light on Musa phylogeny and origin. *Molecular Plant Pathology,* Volume 22, p. 216–230.

Chatzidimopoulos, M. *et al*., 2019. High-Resolution Melting (HRM) Analysis Reveals Genotypic Differentiation of Venturia inaequalis Populations in Greece. *Frontiers in Ecology and Evolution,* Volume 7, p. 1–8.

Chen, X. *et al*., 2011. Rapid detection of isoniazid, rifampin, and ofloxacin resistance in mycobacterium tuberculosis clinical isolates using high-resolution melting analysis. *Journal of Clinical Microbiology,* Volume 49, p. 3450–3457.

Chen, Y., Evans, J. & Feldlaufer, M., 2006. Horizontal and vertical transmission of viruses in the honey bee, Apis mellifera. *Journal of Invertebrate Pathology,* Volume 92, p. 152–159.

Chiba, S. *et al*., 2011. Widespread endogenization of genome sequences of non-retroviral rna viruses into plant genomes. *PLoS Pathogens,* Volume 7.

Chiumenti, M. *et al*., 2016. Unusual genomic features of a badnavirus infecting mulberry. *Journal of General Virology,* Volume 97, p. 3073–3087.

Cornejo, O. E. *et al*., 2018. Population genomic analyses of the chocolate tree, *Theobroma cacao* L., provide insights into its domestication process. *Communications Biology,* Volume 1, p. 1–12.

Daymond, A., 2018. Safe handling and movement of cocoa germplasm for breeding. p. 101–110.

Demeke, T. & Jenkins, G. R., 2010. Influence of DNA extraction methods, PCR inhibitors and quantification methods on real-time PCR assay of biotechnology-derived traits. *Analytical and Bioanalytical Chemistry,* Volume 396, p. 1977–1990.

Dietzgen, R. G., Mann, K. S. & Johnson, K. N., 2016. Plant virus-insect vector interactions: Current and potential future research directions. *Viruses,* Volume 8.

Diop, S. I. *et al*., 2018. Tracheophyte genomes keep track of the deep evolution of the Caulimoviridae. *Scientific Reports,* Volume 8, p. 1–9.

Domfeh, O. *et al*., 2011. Cocoa swollen shoot virus disease situation in Ghana: A review of current trends. *African Journal of Agricultural Research,* Volume 6, p. 5033–5039.

Donaldson, L., 2020. Autofluorescence in plants. *Molecules,* Volume 25.

Dubey, S. C. *et al*., 2021. Plant quarantine for biosecurity during transboundary movement of plant genetic resources. *Indian Phytopathology,* Volume 74, p. 495–508.

Dumur, T. *et al*., 2019. Probing the 3D architecture of the plant nucleus with microscopy approaches: challenges and solutions. *Nucleus,* Volume 10, p. 181–212.

Dzahini-Obiatey, H., Ollennu, A. & Aculey, P. C., 2009. Cocoa swollen shoot virus in Ghana: A review of diagnostic procedures. *Ghana Journal of Agricultural Science,* Volume 41, p. 123–136.

End, M. J., Daymond, A. J. & Hadley, P., 2021. *The Alliance of Bioversity International and the International Center for Tropical.* s.l.:s.n.

Feschotte, C. & Gilbert, C., 2012. Endogenous viruses: Insights into viral evolution and impact on host biology. *Nature Reviews Genetics,* Volume 13, p. 283–296.

Franzen, M. & Borgerhoff Mulder, M., 2007. Ecological, economic and social perspectives on cocoa production worldwide. *Biodiversity and Conservation,* Volume 16, p. 3835–3849.

Frisch, C. *et al*., 2003. From EST to IHC: Human antibody pipeline for target research. *Journal of Immunological Methods,* Volume 275, p. 203–212.

García-Plazaola, J. I. *et al*., 2015. Autofluorescence: Biological functions and technical applications. *Plant Science,* Volume 236, p. 136–145.

Garzo, E., Bonani, J. P., Lopes, J. R. S. & Fereres, A., 2012. Arthropod Structure & Development Morphological description of the mouthparts of the Asian citrus

psyllid , Diaphorina citri Kuwayama (Hemiptera : Psyllidae). *Arthropod Structure and Development,* Volume 41, p. 79–86.

Geering, A. D. W., Scharaschkin, T. & Teycheney, P. Y., 2010. The classification and nomenclature of endogenous viruses of the family Caulimoviridae. *Archives of Virology,* Volume 155, p. 123–131.

Ghanim, M., Brumin, M. & Popovski, S., 2009. A simple, rapid and inexpensive method for localization of Tomato yellow leaf curl virus and Potato leafroll virus in plant and insect vectors. *Journal of Virological Methods,* Volume 159, p. 311–314.

Gosalvez-Bernal, B. *et al*., 2008. Distribution and pathway for phloem-dependent movement of Melon necrotic spot virus in melon plants. *Molecular Plant Pathology,* Volume 9, p. 447–461.

Gutiérrez, S., Michalakis, Y., Van Munster, M. & Blanc, S., 2013. Plant feeding by insect vectors can affect life cycle, population genetics and evolution of plant viruses. *Functional Ecology,* Volume 27, p. 610–622.

Hagen, L. S. *et al*., 1993. *Nucleotide sequence and genomic organization of cacao swollen shoot virus.* s.l.:s.n.

Harper, G., Hull, R., Lockhart, B. & Olszewski, N., 2002. Viral sequences integrated into plant genomes. *Annual Review of Phytopathology,* Volume 40, p. 119–136.

Hedman, J. & Rådström, P., 2015. *Overcoming inhibition in real-time diagnostic PCR.* s.l.:s.n.

Holmes, E. C., 2011. The evolution of endogenous viral elements. *Cell Host and Microbe,* Volume 10, p. 368–377.

Im, K., Mareninov, S., Diaz, M. F. P. & Yong, W. H., 2019. An introduction to performing immunofluorescence staining. *Methods in Molecular Biology,* Volume 1897, p. 299–311.

Jacquot, E. *et al*., 1999. In situ localization of cacao swollen shoot virus in agroinfected *Theobroma cacao*. *Archives of Virology,* Volume 144, p. 259–271.

Jiménez, J., Moreno, A. & Fereres, A., 2021. Semipersistently transmitted, phloem limited plant viruses are inoculated during the first subphase of intracellular stylet penetrations in phloem cells. *Viruses,* Volume 13.

Jonkman, J., 2020. Rigor and Reproducibility in Confocal Fluorescence Microscopy. *Cytometry Part A,* Volume 97, p. 113–115.

Katzourakis, A. & Gifford, R. J., 2010. Endogenous viral elements in animal genomes. *PLoS Genetics,* Volume 6.

Lockhart, B. E. L., 1990. *Evidence for a double-stranded circular DNA genome in a second group of plant viruses.* s.l.:s.n.

MacLeod, A., Pautasso, M., Jeger, M. J. & Haines-Young, R., 2010. Evolution of the international regulation of plant pests and challenges for future plant health. *Food Security,* Volume 2, p. 49–70.

Marelli, J. P. *et al*., 2019. Chocolate under threat from old and new cacao diseases. *Phytopathology,* Volume 109, p. 1331–1343.

Martin, B., Collar, J. L., Tjallingii, W. F. & Fereres, A., 1997. Intracellular ingestion and salivation by aphids may cause the acquisition and inoculation of non-persistently transmitted plant viruses. *Journal of General Virology,* Volume 78, p. 2701–2705.

McClelland, R. D., Culp, T. N. & Marchant, D. J., 2021. Imaging Flow Cytometry and Confocal Immunofluorescence Microscopy of Virus-Host Cell Interactions. *Frontiers in Cellular and Infection Microbiology,* Volume 11, p. 1–10.

Medberry, S. L., Lockhart, B. E. L. & Olszewski, N. E., 1992. The commelina yellow mottle virus promoter is a strong promoter in vascular and reproductive tissues. *Plant Cell,* Volume 4, p. 185–192.

Mette, M. F. *et al*., 2002. Endogenous viral sequences and their potential contribution to heritable virus resistance in plants. *EMBO Journal,* Volume 21, p. 461–469.

Montagna, M. T. *et al*., 2019. Chocolate, “food of the gods”: History, science, and human health. *International Journal of Environmental Research and Public Health,* Volume 16.

Moreno, A., Tjallingii, W. F., Fernandez-Mata, G. & Fereres, A., 2012. Differences in the mechanism of inoculation between a semi-persistent and a non-persistent aphid-transmitted plant virus. *Journal of General Virology,* Volume 93, p. 662–667.

Moreno, A., Uzest, M. & Fereres, A., 2005. A Single Amino Acid Position in the Helper Component of Cauliflower Mosaic Virus Can Change the Spectrum of Transmitting Vector Species. Volume 79, p. 13587–13593.

Motamayor, J. C. *et al*., 2008. Geographic and genetic population differentiation of the Amazonian chocolate tree (*Theobroma cacao* L). *PLoS ONE,* Volume 3.

Muller, E., 2009. Cacao Swollen Shoot Virus. In: s.l.:s.n., p. 255.

Muller, E. *et al*., 2018. Next generation sequencing elucidates cacao badnavirus diversity and reveals the existence of more than ten viral species. *Virus Research,* Volume 244, p. 235–251.

Muller, E. *et al*., 2021. Identification and distribution of novel badnaviral sequences integrated in the genome of cacao (*Theobroma cacao*). *Scientific Reports,* Volume 11, p. 1–13.

Ndowora, T. *et al*., 1999. Evidence that badnavirus infection in Musa can originate from integrated pararetroviral sequences. *Virology,* Volume 255, p. 214–220.

Ng, J. C. K. & Falk, B. W., 2006. Virus-Vector Interactions Mediating Nonpersistent and Semipersistent Transmission of Plant Viruses.

Ng, J. C. K. & Zhou, J. S., 2015. Insect vector – plant virus interactions associated with perspectives and future challenges. *Current Opinion in Virology,* Volume 15, p. 48–55.

Obok, E., Wetten, A. & Allainguillaume, J., 2018. Electropenetrography application and molecular-based virus detection in mealybug (Hemiptera: Pseudococcidae) vectors of Cacao swollen shoot virus on *Theobroma cacao* L.. *Annals of Agricultural Sciences,* Volume 63, p. 55–65.

Perrine-Walker, F., 2020. Phytophthora palmivora–cocoa interaction. *Journal of Fungi,* Volume 6, p. 1–20.

Quainoo, A. K., Wetten, A. C. & Allainguillaume, J., 2008. The effectiveness of somatic embryogenesis in eliminating the cocoa swollen shoot virus from infected cocoa trees. *Journal of Virological Methods,* Volume 149, p. 91–96.

Raccah, B., 2001. Plant Virus Transmission by Insects. *eLS.*

Raccah, B. & Fereres, A., 2009. Plant Virus Transmission by Insects. *eLS.*

Ramírez, A. M. H., Duque, H. J. S. & Trujillo, A. I. U., 2017. Quality of cocoa (*Theobroma cacao* L.) DNA from foliar tissue at different stages of development. *Acta Agronomica,* Volume 67, p. 311–318.

Reed, G. H., Kent, J. O. & Wittwer, C. T., 2007. High-resolution DNA melting analysis for simple and efficient molecular diagnostics. *Pharmacogenomics,* Volume 8, p. 597–608.

Rezadoost, M. H., Kordrostami, M. & Kumleh, H. H., 2016. An efficient protocol for isolation of inhibitor-free nucleic acids even from recalcitrant plants. *3 Biotech,* Volume 6, p. 1–7.

Richert-Pöggeler, K. R. *et al*., 2003. Induction of infectious petunia vein clearing (pararetro) virus from endogenous provirus in petunia. *EMBO Journal,* Volume 22, p. 4836–4845.

Roberts, K. *et al*., 2007. Long-distance movement of Cauliflower mosaic virus and host defence responses in Arabidopsis follow a predictable pattern that is determined by the leaf orthostichy. *New Phytologist,* Volume 175, p. 707–717.

Sagemann, W. *et al*., 1985. Detection and Comparison of some Ghanaian Isolates of Cacao Swollen Shoot Virus (CSSV) by Enzyme-Linked Immunosorbent Assay (ELISA) and Immunoelectron Microscopy (IEM) Using an Antiserum to CSSV Strain 1A. *Journal of Phytopathology,* Volume 114, p. 79–89.

Shargil, D. *et al*., 2015. Development of a fluorescent in situ hybridization (FISH) technique for visualizing CGMMV in plant tissues. *Journal of Virological Methods,* Volume 223, p. 55–60.

Staginnus, C. *et al*., 2009. Suggestions for a nomenclature of endogenous pararetroviral sequences in plants. *Archives of Virology,* Volume 154, p. 1189–1193.

Staginnus, C. & Richert-Pöggeler, K. R., 2006. Endogenous pararetroviruses: two-faced travelers in the plant genome. *Trends in Plant Science,* Volume 11, p. 485–491.

Strickland, A. H., 1951. The Entomology of Swollen Shoot of Cacao. II.–The Bionomics and Ecology of the Species involved. *Bulletin of Entomological Research,* Volume 42, p. 65–103.

Tamukra, K. a. N. M., 1993. *Estimation of the number of nucelotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees..* s.l.:s.n.

Thresh, J. M. & Fargette, D., 2001. The epidemiology of African plant viruses: basic principles and concepts.. *Virus and virus-like diseases of major crops in developing countries.,* p. 61–111.

Tong, S. Y. C. & Giffard, P. M., 2012. Microbiological applications of high-resolution melting analysis. *Journal of Clinical Microbiology,* Volume 50, p. 3418–3421.
Torbert, K. A. *et al*., 1998. Expression of the Commelina yellow mottle virus promoter in transgenic oat. *Plant Cell Reports,* Volume 17, p. 284–287.
Umber, M. *et al*., 2017. The genome sequence of Dioscorea bacilliform TR virus, a member of the genus Badnavirus infecting Dioscorea spp., sheds light on the possible function of endogenous Dioscorea bacilliform viruses. *Archives of Virology,* Volume 162, p. 517–521.
Vassilieff, H. *et al*., 2023. Endogenous Caulimovirids: Fossils, Zombies, and Living in Plant Genomes. *Biomolecules,* Volume 13, p. 1–18.
Wessel, M. & Quist-Wessel, P. M. F., 2015. Cocoa production in West Africa, a review and analysis of recent developments. *NJAS - Wageningen Journal of Life Sciences,* Volume 74-75, p. 1–7.
Wetten, A., Campbell, C. & Allainguillaume, J., 2016. High-resolution melt and morphological analyses of mealybugs (Hemiptera: Pseudococcidae) from cacao: Tools for the control of Cacao swollen shoot virus spread. *Pest Management Science,* Volume 72, p. 527–533.
Witte, R. *et al*., 2018. Concepts in light microscopy of viruses. *Viruses,* Volume 10, p. 1–31.

APPENDIX I

Raw qPCR data from stem sample experiments

Table 18: Concentration of CSSV in stem samples when undiluted and diluted - Log_{10} CSSV copies μL^{-1}, n = 6.

Plant Code[1]	**CSSV Concentration (Log10 CSSV copies μL^{-1})**																	
	Dilution 1:10						Dilution 1:20						Undiluted					
	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
G13	3.698	3.675	3.682	3.873	0.000	3.795	3.396	3.290	3.190	3.468	3.196	3.328	0.000	0.000	0.000	0.000	0.000	0.000
G20	3.107	2.852	2.743	2.497	2.802	2.644	1.900	2.265	1.525	0.000	2.086	2.562	0.000	0.000	0.000	0.000	0.000	0.000
G30	2.990	3.152	3.204	3.140	3.083	3.000	2.121	2.672	2.583	2.804	2.176	2.920	0.000	0.000	0.000	0.000	0.000	0.000
G91	3.671	3.763	3.714	3.867	3.694	3.660	3.393	3.336	3.210	3.330	3.310	3.336	0.000	0.000	0.000	0.000	0.000	0.000

[1] CSSV infected plant (Table in methods)

Table 19: Concentration of CSSV in stem samples following two hour and 30-minute incubation times - Log_{10} CSSV copies μL^{-1}, n = 6.

Plant Code	**CSSV Concentration (Log10 CSSV copies µL-1)**											
	2 Hour Incubation						30 Minute Incubation					
	1	2	3	4	5	6	1	2	3	4	5	6
G13	3.851	3.777	3.739	3.691	3.854	3.885	3.644	3.270	2.535	3.303	3.519	3.114
G30	3.076	3.338	2.868	2.962	2.844	2.723	2.912	2.964	2.121	1.199	1.594	2.833
G44	3.673	3.442	3.767	3.530	3.710	3.430	3.384	3.459	3.057	3.267	2.899	3.188

Table 20 Concentration of CSSV in stem samples following immediate extraction and after being stored for one week - Log_{10} CSSV copies μL^{-1}, n = 6.

Plant Code	**CSSV Concentration (Log10 CSSV copies µL-1)**											
	Immediate extraction						Stored for one week					
	1	2	3	4	5	6	1	2	3	4	5	6
G13	3.988	4.188	4.104	4.053	4.140	4.093	3.952	3.955	4.037	3.916	4.053	4.041
G30	3.489	3.641	3.386	3.477	3.146	3.413	3.508	0.000	3.111	3.025	3.248	3.146
G44	4.004	4.170	3.823	4.111	3.980	3.820	3.925	4.021	4.086	4.037	4.114	0.000

Table 21: CSSV concentration in separated stem components – middle heartwood and outer peel.

Stem section[1]	**CSSV Concentration (Log_{10} CSSV copies µL-1)**																	
	Biological replicate 1						Biological Replicate 2						Biological Replicate 3					
	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Middle Heartwood	2.170	2.394	0.000	0.000	0.000	0.000	0.000	0.000	2.274	0.000	2.207	2.442	0.000	0.000	2.045	0.000	0.000	0.000
Outer peel	2.582	2.903	2.733	2.752	2.303	2.567	0.000	0.000	2.025	2.520	1.455	2.493	3.121	3.140	2.820	2.701	2.923	2.818

[1] From CSSV infected plant G20 (Table in methods)

Table 22: Concentration of synthetic CSSV reference sequences when used to spike an uninfected *T. cacao* stem sample and diluted.

Spiked sample	**Log10 CSSV copies µL-1**					
	1	2	3	4	5	6
Undiluted	0.000	0.000	0.000	0.000	0.000	0.000
1:10	5.688	5.732	5.725	5.878	5.682	5.787
1:20	5.562	5.316	5.358	5.487	5.354	5.348
1:40	5.017	5.029	5.013	5.025	4.977	5.045
1:80	5.013	4.755	4.754	4.810	4.732	4.790

APPENDIX II

scientific reports

Check for updates

OPEN

Identification and distribution of novel badnaviral sequences integrated in the genome of cacao (*Theobroma cacao*)

Emmanuelle Muller[1,2,5], Ihsan Ullah[3,5], Jim M. Dunwell[3]✉, Andrew J. Daymond[3], Megan Richardson[4], Joël Allainguillaume[4] & Andy Wetten[4]

Theobroma cacao **is one of the most economically important tropical trees, being the source of chocolate. As part of an ongoing study to understand the diversity of the badnavirus complex, responsible for the cacao swollen shoot virus disease in West Africa, evidence was found recently of virus-like sequences in asymptomatic cacao plants. The present study exploited the wealth of genomic resources in this crop, and combined bioinformatic, molecular, and genetic approaches to report for the first time the presence of integrated badnaviral sequences in most of the cacao genetic groups. These sequences, which we propose to name eTcBV for endogenous *T. cacao* bacilliform virus, varied in type with each predominating in a specific genetic group. A diagnostic multiplex PCR method was developed to identify the homozygous or hemizygous condition of one specific insert, which was inherited as a single Mendelian trait. These data suggest that these integration events occurred before or during the species diversification in Central and South America, and prior to its cultivation in other regions. Such evidence of integrated sequences is relevant to the management of cacao quarantine facilities and may also aid novel methods to reduce the impact of such viruses in this crop.**

Viral integration within eukaryotic genomes is now a widely recognised phenomenon[1] and has been described and documented in many species during the last 15 years, thanks to the complete sequencing of many genomes. It occurs in both animals and plants but, unlike animal retroviruses, integration into a plant genome is not an obligatory step in the life cycle of plant viruses. It is proposed that such horizontal gene transfer (HGT) results from illegitimate recombination during the repair of double-stranded DNA breakages[1,2]. Within the plant kingdom, there is evidence that integration of viral sequences occurs in a wide range of families, while the integrated viruses belong mostly to the *Caulimoviridae* family. Such endogenous viral elements (EVEs) belonging to five genera of this family have been characterised in at least 27 plant species from nine different plant families[3–7].

The distribution and structure of EVEs within plant genomes are diverse and range from short, dispersed, and repetitive viral sequences to longer stretches of near-full length viral genomes. Most EVEs have no deleterious effect on their hosts because they comprise rearranged sequences with inactivating mutations, but in some cases, integrated sequences contain a functional full-length viral genome, which can be activated and lead to systemic infection of the host plant. These latter cases result in infective integration; examples include *Petunia vein clearing virus* (PVCV, genus *Petuvirus*) in petunia[8], *Tobacco vein clearing virus* (TVCV, genus *Solendovirus*) in tobacco[9] and *Banana streak virus* (BSV; genus *Badnavirus*) in banana[10–12].

Cacao (*Theobroma cacao*) is one of the most economically important tropical trees, being vital to the cocoa industry, and providing a livelihood for several million smallholder farmers in the developing world. Particularly in West Africa, growers' incomes are now threatened by potential reduction in yield caused by susceptibility of the crop to several different viruses of the genus *Badnavirus*[13–15]. However, the presence of most of these viruses has never been described in South and Central America, the geographical origin of the cacao tree.

The diversity of the viruses that infect cacao has been described and can be categorised into at least 11 Badnavirus species[16,17]. These include eight species associated with the commercially important cacao swollen shoot disease (*Cacao swollen shoot Togo A virus*—CSSTAV, *Cacao swollen shoot Togo B virus*—CSSTBV, *Cacao*

[1]CIRAD, AGAP Institut, 34398 Montpellier, France. [2]AGAP, Univ Montpellier, CIRAD, INRAE, Institut Agro, Montpellier, France. [3]School of Agriculture, Policy and Development, University of Reading, Earley Gate, Reading RG6 6EU, UK. [4]University of the West of England, Frenchay Campus, Coldharbour Lane, Bristol BS16 1QY, UK. [5]These authors contributed equally: Emmanuelle Muller and Ihsan Ullah. ✉email: j.dunwell@reading.ac.uk

swollen shoot CD virus—CSSCDV, *Cacao swollen shoot CE virus*—CSSCEV, *Cacao swollen shoot Ghana M virus*—CSSGMV, *Cacao swollen shoot Ghana N virus*—CSSGNV, *Cacao swollen shoot Ghana Q virus*—CSSGQV and the proposed Cacao swollen shoot Ghana T virus—CSSGTV) in West Africa, two species (*Cacao mild mosaic virus*—CaMMV, *Cacao yellow vein banding virus*—CYVBV) described in Trinidad, one of which was also found in Puerto Rico[18], and one species (*Cacao bacilliform Sri Lanka virus*—CBSLV) from Sri Lanka. One additional cacao badnavirus species [named S species[17]], distinct from all the other known badnaviruses, has been frequently detected in cacao trees in West Africa but the complete genome of this species has not yet been reconstructed.

The search for genetic resistance to *Cacao swollen shoot virus* (CSSV) and other internationally important diseases has stimulated much recent research into the genetic diversity of cacao germplasm available to the breeder. In general terms, the great diversity of cacao genotypes comprises two broad genetic groups "Criollo" and "Forastero" defined on morphological and geographical origins, and a third group "Trinitario" recognized as "Criollo" X "Forastero" hybrids. In recognition of the high genetic diversity within the Forastero grouping, Motamayor et al. in 2008 proposed a new classification of *T. cacao* germplasm into ten genetic groups based on Bayesian statistics applied to genotyping data obtained with microsatellite markers[19]. The ten groups, Amelonado, Contamana, Criollo, Curaray, Guiana, Iquitos, Marañon, Nacional, Nanay, and Purús reflect cacao diversity more accurately and this designation is now used for distribution studies of genetic and phenotypic traits. Recently, the genomes of 200 cacao accessions were sequenced and it was shown that along with their assignment to one of the ten genetic population groups, a high number of cacao genotypes correspond to admixed individuals with a differential contribution from several ancestral populations[20]. These analyses provide a more realistic picture of the very complex diversity of cacao as a consequence of genetic diversification and various domestication processes.

As part of an ongoing study of the relationship between CSSV diversity and the cacao host, sequences corresponding to the cacao badnavirus species S have been widely detected in asymptomatic cacao trees, including examples not originating from West Africa. These results raised the possibility of integration of viral sequences from this species in the host genomes. In order to address this question, the present study combined PCR-based analyses with a targeted in silico search for cacao badnavirus S sequences in the wide range of cacao genome sequence data. We show that sections of the genome of this species exist in a variety of integrated forms in the genome of many cacao genotypes. In addition, for the characterized insertion of type VI (Trinitario), a multiplex PCR was developed and used to confirm the Mendelian inheritance of this specific insert, and to characterise the accessions at the International Cocoa Quarantine Centre, the University of Reading, UK (ICQC, R) (http://www.icgd.reading.ac.uk/icqc/).

Results

Identification of reverse transcriptase (RT) RNase H badnaviral sequences from asymptomatic cacao plants. In an initial study, a fragment of the expected size of 628 bp was consistently obtained by PCR with the primer pair Badna1deg2/Badna4deg2 (corresponding to RT RNase H region of badnaviruses genomes) from apparently healthy leaf samples collected from the ICQC, R, from the Centre de Coopération Internationale en Recherche Agronomique pour le Développment, France, (CIRAD) collection, and from seedlings grown from germinated seeds. A phylogeny was constructed from the alignment of the amplified fragments (Fig. 1) showing that they correspond to the cacao badnavirus S species, except for one sequence from sample GU 114/P (Guiana group). This sequence is clearly different, with a nucleotide identity of less than 80%, and belongs to another species, provisionally named S prime. Among sequences belonging to cacao badnavirus S, we observed 12 consistent subgroups, each containing a characteristic insert, designated as type I to XII (Fig. 1). Supplemental Table S1 lists the cacao clones that produced PCR products with these primers, along with the viral type obtained. When sequences were amplified from samples of the same clones (CCN 51, EBC 10, PA 120, PA 137, PA 150, IMC 55 and IMC 47) maintained independently in two different collections (ICQC,R, CIRAD or Trinidad), the sequences were identical. Representative samples from the ten cacao genetic groups mentioned above were tested to determine the distribution of such sequences alongside the diversity of cacao genomes according to the previous description[19]. Surprisingly, we observed that this PCR analysis suggested that these badnaviral sequences were present in all cacao diversity groups but with a higher prevalence in the five groups Guiana, Iquitos, Marañon, Nanay, and Purús, and in the group of admixed clones (Table S1 online). By comparison, only two sequences were obtained from the Nacional group and only one sequence from each of the Amelonado, Contamana, Criollo and Curaray groups.

We also observed a correlation between the cacao genomic groups and the type of badnavirus S amplified (Fig. 1 and Table S1 online). All but one of the sequences from samples of the Guiana group contain identical type I sequences. Similarly, all sequences from samples of the Iquitos group contain identical type V sequences.

Long amplicons and assembly containing sequences of cacao badnavirus S. In order to amplify a more complete genome of the badnavirus S species, inverted abutting primers were designed. These primers, designed in the RT -RNase H region, allowed the amplification of identical 2421 bp amplicons from cacao clones NA 79 and NA 226.

From the bioinformatic analysis of the Illumina data from an ICS 76 sample, different sizes of contigs containing badnavirus S sequences were obtained. The sequence of the 2421 bp amplicon amplified from NA 79 and a contig of 3813 nt from ICS 76 were aligned and used to design primers S2465Fdeg /S4666R to amplify a ~ 2.2 kb fragment (Fig. 2). PCR performed with this primer pair amplified a fragment of the expected size from five different groups of cacao clones, namely PA 211 (Marañon), GU 230/C (Guiana), IMC 55 (Iquitos), EBC 5 (Purús) and NA 79 (Nanay), along with the Trinitario clone ICS 95 from an admixed group. BLASTN analysis of these amplicons identified a 1.4 kb region with ~ 60% nucleotide identity in the RT RNase H region of all other badnaviral species.

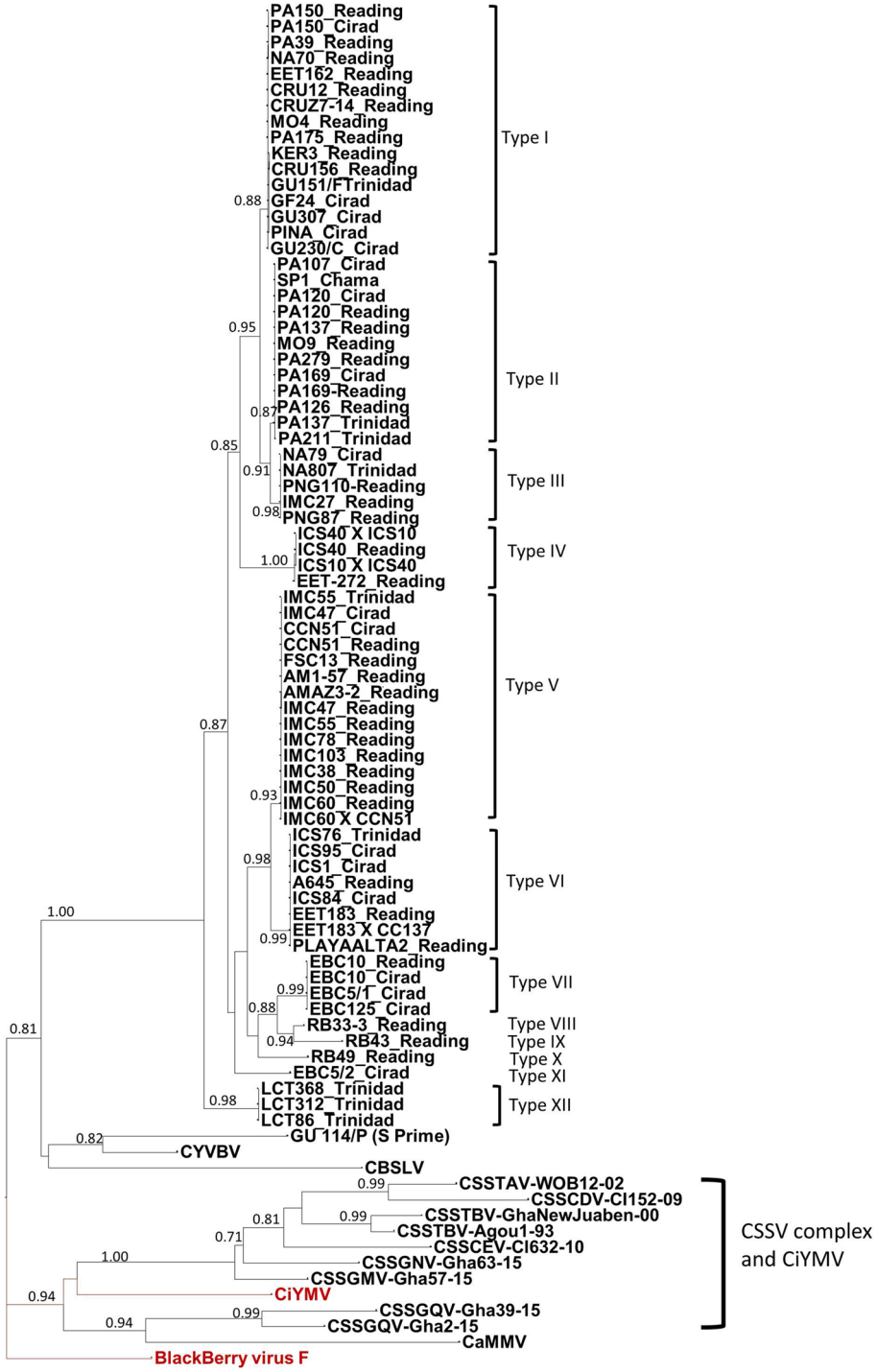


Figure 1. Maximum likelihood phylogenetic tree of badnavirus S sequences based on alignment of the RT RNase H region of open reading frame 3 (ORF3). Numbers on the branches represent the SH-aLRT (approximate Likelihood ratio test) branch supports over 0.7. The Citrus yellow mosaic Virus (CiYMV) (AF347695) and Blackberry virus F (YP009229919) in red colour are used as outgroups along with the other badnavirus infecting cacao trees [CYVBV (KX276640), CaMMV (KX276641), CBSLV (MF642736) and species from the Cacao swollen shoot complex (CSSTAV, AJ781003; CSSTBV, L14546 and AJ608931; CSSCDV, JN606110; CSSCEV, MF642719; CSSGMV, MF642724; CSSGNV, MF642725; CSSGQV, MF642726 and MF642733]. The names of sequences include the name of the cacao clone and the name of the collection from which they were obtained. The 12 different viral types of sequence are identified from I to XII.

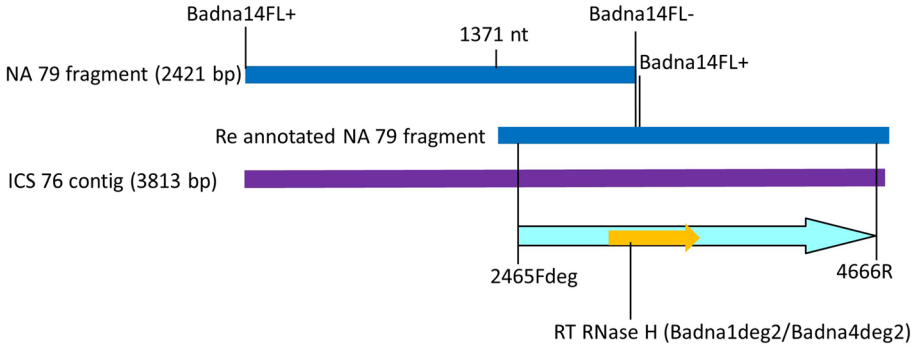


Figure 2. Illustration of design of primers used to amplify S type viral sequences. A fragment of 2421 bp amplified (blue bar) from cacao clone NA 79 with inverted abutting primers Badna14FL + and Badna14FL- was sequenced, re annotated (starting position at 1371) and aligned with a 3813 bp contig constructed from Illumina sequence of cacao clone ICS 76 (purple bar). This alignment was used to design a set of degenerate primers (2465Fdeg and 4666R) for amplification of a ~ 2.2 kb S type viral sequence. Vertical lines indicate locations of primers. Location of RT RNase H region is indicated as a yellow arrow.

In silico screening of whole genome sequencing (WGS) data from cacao. A series of independent WGS datasets were utilised in this analysis. The largest of these datasets was generated in BioProject PRJNA486011, the overall aim of which was to explore the cacao domestication process, and its conclusions provide valuable insights into the evolutionary history of the cacao populations and their population structure[20]. The data generated in this study comprises sequence information from 200 cacao genomes with 5.3–74.5X coverage. We aligned raw data from each of these genomes with viral reference sequences using Bowtie 2 to detect viral sequences. These reference data included sequences of ~ 2.2 kb viral fragments amplified from clones GU 230/C, PA 211, NA 79, IMC 55 and ICS 95, hereafter designated as type I, type II, type III, type V and type VI, respectively (a subset of the total of the 12 different types identified above). Our analysis identified viral sequences in datasets from 103 genomes, out of which a single viral sequence was detected in 95 genomes, and multiple viral sequences were found in eight genomes. Type VI was found to be the most prevalent, detected alone or in combination in 52 and one genomes, respectively. The V, II, III and I types were independently found in 15, 12, 10 and 6 genomes, respectively; these four types were also found in combination in four genomes (Fig. 3, supplemental Table 2). In their study, Cornejo et al.[20] placed 79 reference cacao clones into the ten major genetically distinct groups described above, namely Amelonado, Contamana, Criollo, Curaray, Guiana, Iquitos, Marañon, Nacional, Nanay and Purús. The remaining 121 admixed clones were placed in five arbitrary sub-groups (Fig. 3) based on the major contribution of genetic groups to their mixed ancestry. Our overall results revealed an interesting correlation between genetic grouping of reference clones and mapping of viral sequences. The genomic data search did not detect viral sequence in the reference clones designated as Contamana, Criollo, Curaray, or Nacional. All 11 Amelonado reference clones, except TRD 86, were also found to be free of viral sequences. All 14 Marañon reference clones contained type II viral sequence, whereas 80% of Nanay reference clones had type III sequence. Similarly, type I and V viral sequences were prevalent in Guiana and Iquitos reference clones, respectively. The type VI viral sequence was present in all Purús reference clones except CAB_71_PL3, whereas types V and VI were prevalent in the admixture sub-groups with 70% and 21% representation, respectively (Fig. 3, supplemental Table 2).

The second group of WGS datasets, 31 in total, were downloaded from European Nucleotide Archive (ENA) for BioProject PRJNA558793. In addition, data from the ICS 1 and Catongo Blanco clones were directly downloaded from the NSF (project submitter) website (https://plantscience.psu.edu/research/labs/guiltinan/nsf-plant-genome-research-program). These 33 datasets are notable for their very deep genome coverage (20.8–88.5X). Screening of the datasets detected viral sequences in 23 datasets. The clones used in the study were selected from six of the genetic groups, namely Amelonado, Guiana, Iquitos, Marañon, Nanay and Trinitario, and mapping of viral sequences was found to be correlated to the genetic grouping, as previously found above in BioProject PRJNA486011. Specifically, the clones in the Guiana, Marañon, Nanay and Iquitos genetic groups were mapped with type I, II, III and V viral sequences, the only exception being PA 13 (Marañon), which mapped with type I (Table 1). The Trinitario clone ICS 1 contained the viral type VI as expected. The reliability of the data analysis was confirmed on the basis that there are ten clones common to the PRJNA558793 and PRJNA486011 BioProjects. Common clones positive for the viral mapping, AMAZ 15/15, NA 33, PA 107 and ICS 1 contained the similar viral type V, III + V, II, and VI, respectively in both studies (Table 1). Interestingly, clone PA 13 (Marañon), which mapped to type I in this study also mapped in the analysis above to type I in the dataset from BioProject PRJNA486011, where it is defined as being admixed. It is noteworthy that although the samples for AMAZ 15/15, PA 107 and PA 13 in the two studies were obtained from two different germplasm centres, they produced the same viral type pattern (Table 1).

The third group of WGS datasets, ten in total, were downloaded and analysed from BioProject PRJNA77799[21] in which the raw read data had 4.2 to 11.0X coverage. This study, which includes six Trinitario clones, and one clone each from Amelonado, Criollo and Forastero, also includes a dataset from the cacao wild relative *T. grandiflorum*. Two Trinitario clones, EET 64 and ICS 1, were mapped with type VI (Table 1). There are four clones

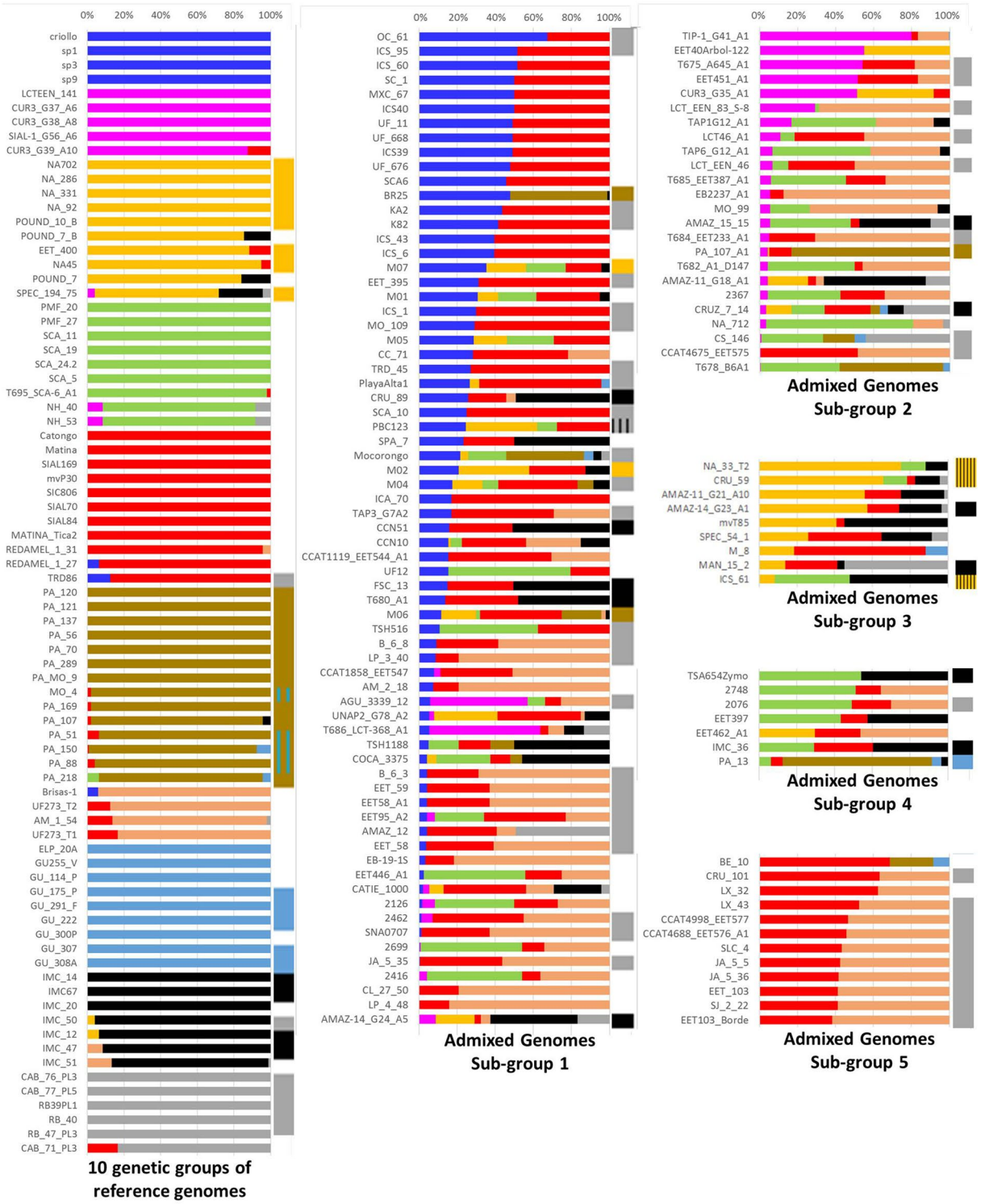


Figure 3. Summary of the bioinformatic search for the presence of the different types of viral sequences in cacao genomes. Bar graph representing population structure in cacao, redrawn from a previous study[20]. The first block on the left side represents 79 clones (with names) grouped into ten distinct genetic groups. The other two blocks represent 121 admixed clones grouped into five sub-groups based on shared proportion of ancestry (horizontal bars). The narrow column on the right side of the individual horizontal bars represents the type of viral sequence found in the clones positive for the viral mapping. The colours correspond to the following groups/**virus types**: Blue (Criollo), Magenta (Curaray), Golden (Nanay/**III**), Green (Contamana), Red (Amelonado), Dark brown (Marañon/**II**), Light brown (Nacional), Light blue (Guiana/**I**), Black (Iquitos/**V**) and Grey (Purús/**VI**). Boxes with hatching patterns represent two viral sequences.

PRJNA77799		C	PRJNA558793		C	PRJNA558793			
Clone	Viral type	Viral type	Clone	Viral type	Viral type	Clone	Viral type	Clone	Viral type
ICS 1	VI	VI	*AMAZ 12*	–	VI	GU 123V	I	NA 70	III + V
ICS 6	–	–	*AMAZ 15/15*	V	V	GU 195V	–	NA 710	III
ICS 39	–	–	*Catongo_ Blanco*	–	–	GU 257E	–	NA 807	III
SCA 6	–	–	*COCA 3370/5*	–	–	IMC 105	V	NA 916	III
Amelonado	–		*ICS 1*	VI	VI	IMC 31	–	OYA 2B	–
Criollo 22	–		*NA 33*	III + V	III + V	IMC 57	V	PA 16	II
EET 64	VI		*PA 107*	II	II	IMC 60	–	PA 279	II + VI
Pentagonum	–		*PA 13*	I	I	KER 1L	I	PA 299	II
Stahel	–		*POUND 7*	–	–	KER 6	I	PA 71	I
T. grandiflorum	–		*SPEC 54/1*	–	–	NA 246	II	PA 81	II
			ELP 37A	–		NA 34	III	Pina	I

Table 1. Summary of findings from BioProject PRJNA77799[21] and BioProject PRJNA558793[47], and comparison with findings from BioProject PRJNA486011[20], designated as "C". Italic font (except *T. grandiflora*) indicates clones common to two or more studies.

common to this study and the 200 genomes study[20]. Both studies confirmed that no viral sequences were found in clones SCA 6, ICS 6 and ICS 39, whereas ICS 1 mapped with type VI virus (Table 1). This latter clone, which is included in all three studies we examined, was mapped to the same viral type, despite the source material being obtained from different germplasm centres (Costa Rica, Puerto Rico, and Trinidad).

We also searched one dataset (MATINA 1/6) from BioProject PRJNA51633[19], and a long read PacBio dataset of POUND 7, BioProject PRJNA421343[22], for badnaviral S sequences; however, no viral sequences were identified in either dataset.

Amplification of virus-plant junction fragment in cacao clone PA 279. Based on evidence from the global bioinformatic analysis of the raw genome data from PA 279, a more targeted search for viral sequences was conducted in the preliminary assembly of this clone, downloaded from the NSF website mentioned above. This revealed a type VI sequence in jcf7180010890274, a 51,364 bp contig that contains the viral sequence from position 16,480 to 18,593 nt. A BLASTN search in the B97 *T. cacao* genome (i.e. the reference Criollo *T. cacao* genome available in GenBank) found high similarity of the contig region 1–15,454 and 21,505–51,364 with Chromosome V (GenBank number LT 594,792.1), whereas contig region 15,455–21,504 nt had no similarity with the cacao genome. A BLASTN search of this 6050 nt sequence in the virus database showed similarity with badnaviral sequences corresponding to the RT RNase H region of type VI and the 12 bases $tRNA^{met}$ binding site conserved for every virus of the *Caulimoviridae* family (Fig. 4a). The putative viral region is flanked by B97 *T. cacao* genome Chromosome V, region 32,875,968–32,885,707 nt at the left and 32,885,725- 32,891,127 nt at the right side without a gap.

To further analyse the position of the viral insert, two primer sets were designed flanking the insertion site on B97 *T. cacao* genome Chromosome V. Both sets successfully amplified two bands from PA 279 (accession number RUQ 1119) and one band of expected size (without insertion) from Criollo 11 [CRI] (accession number RUQ 1718). The upper band of around 6.5 kb amplified from PA 279 with primer set I (Fig. 4b) was cloned and sequenced. The complete assembled contig of the clone comprised 6279 bp. The sequencing data revealed 100% similarity with the targeted region of the PA 279 contig jcf7180010890274 mentioned above. An insertion of 6050 bp was found in the genome of PA 279 at Chr V: 32,885,707–32,885,725 nt (B97 *T. cacao* numbering). This insertion site is located in a non-coding region of "uncharacterized" locus LOC18599617. A BLASTN search in the NCBI non-redundant nucleotide database found 98% similarity with B97 Chromosome V at both termini (left terminus from position 1–104 nt, and right terminus from position 6155–6279 nt of the contig). The contig sequence from position 1685–2166 nt showed 100% similarity with Gha68-16 S sequence corresponding to a putative reverse transcriptase protein (ORF3) gene, partial cds (GenBank accession MF784038.1[17]). A 17 nt long deletion was found in the amplified fragment in the host genome immediately after viral sequence insertion site (32,885,708–724 nt, B97 *T. cacao* numbering). The inserted sequence contains two ORFs ≥ 600 nt in length, with the peptide coded by ORF1 consisting of 817 aa and containing a pepsin-like aspartate protease, RT LTR and RT RNase H like conserved domains (Fig. 4c). ORF1 (28–735 aa) shows 59% similarity with Blackberry virus F (GenBank accession YP_009229919.1;[23]) polyprotein. The peptide coded by ORF2 consists of 252 aa and shows similarity to a hypothetical protein in several badnaviral species.

Screening of germplasm for the type VI insert. Using information generated in the preceding section, a multiplex PCR assay was developed to screen the germplasm collection at ICQC, R. The assay comprises two primer sets, one positioned at the right junction of the insert, the other one at the left junction of the insert. In positive clones (containing the unmodified insert of 6050 nt), the assay amplifies three fragments of 679, 495 and 143 bp, the first two from the two junctions of the insert with host genome, and the third from the host only (in

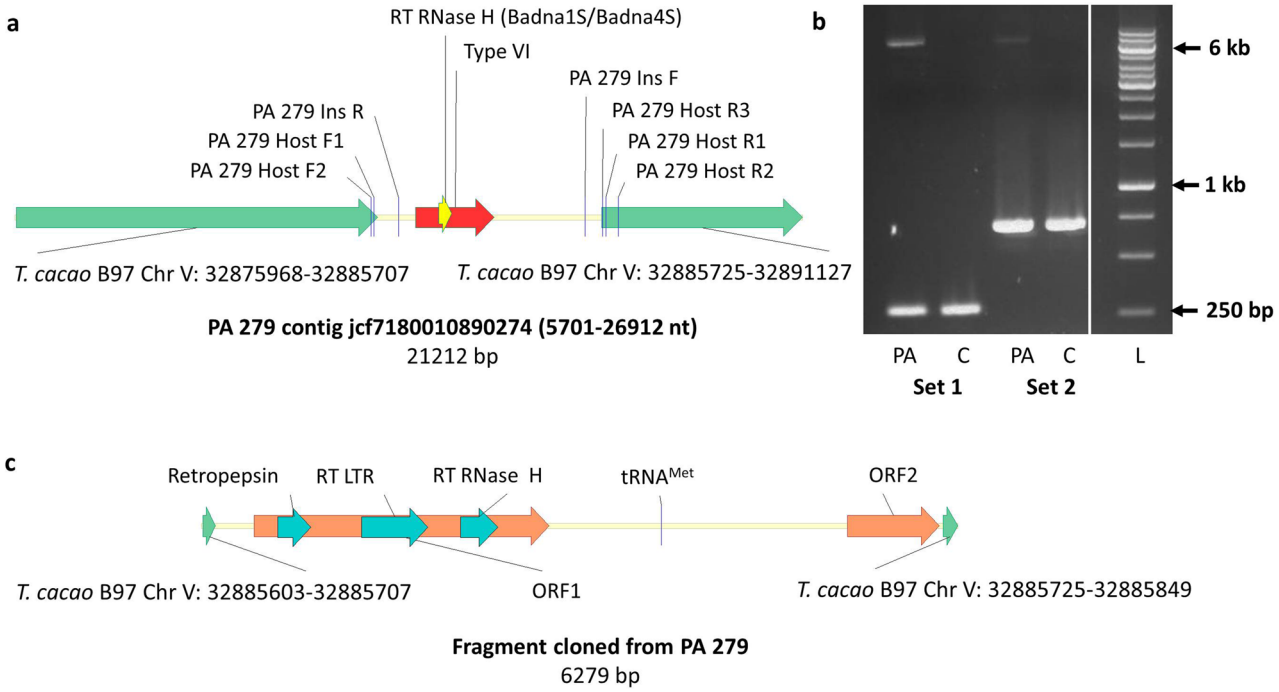


Figure 4. Amplification of type VI viral insertion from cacao clone PA 279. (**a**) Alignment of PA 279 contig jcf7180010890274 and B97 *T. cacao* genome Chromosome V. Green arrows represent genomic region of cacao clone B97 bordering 6050 bp of type VI viral insertion. Vertical lines indicate location of primers including primer set 1 (PA 279 Host F1, PA 279 Host R1) and set 2 (PA 279 Host F2, PA 279 Host R2) used to amplify the viral insert along with the bordering host sequence, and primers used in the multiplex assay to screen cacao germplasm for viral insertion type VI (PA 279 Host F1, PA 279 Ins R, PA 279 Ins F and PA 279 Host R3). (**b**) Amplification of the viral insert along with the bordering host sequence from PA 279 (PA) and Criollo 11 (C) clones. The 243 and 635 bp fragments amplified in both clones with primer set 1 and set 2 represent virus- allele (lacking viral insertion) whereas a fragment between 6 to 7 kb amplified in PA 279 clone in both primer sets represents virus + allele. (**c**) Open reading frames (brown arrows) and conserved domains (light blue arrows) present in the fragment amplified from PA 279 clone harbouring host genome and type VI viral sequence. Full-length gel of section (**b**) is presented in Supplementary Fig. 1.

case of a locus hemizygous for the insert). In clones lacking the insert, the assay only amplifies a 143 bp fragment; this also serves as an internal control (Table 2, Fig. 4a). Screening of 342 germplasm accessions from ICQC, R, revealed that 12 and 51 clones contained type VI viral sequence insertion in the homozygous and hemizygous forms, respectively (Table S1 online). The results on positive clones were verified by a second round of tissue sampling, DNA extraction and PCR amplification. Comparison of screening data of multiplex PCR assay and in silico analysis of WGS datasets for the virus insertion type VI showed confirmatory results. All nine clones common to both types of analysis contain the viral sequence except clone AMAZ 12. In the WGS data, the type VI viral sequence was mapped in AMAZ 12 in PRJNA486011; however, we could not detect any viral type in this clone, in either PRJNA558793, or the multiplex PCR study. This inconsistency may be due to accidental mislabelling in one of the source materials.

Inheritance study of type VI insert. Selfed progenies of two clones each from homozygous and hemizygous type VI virus locus groups (see above), and progenies from two independent crosses between a hemizygous and a negative clone were screened with the multiplex PCR assay to assess the inheritance pattern of the insert. No segregation was found in selfed progenies of clones B 9/10–25 and DOM 3, both homozygous for the virus locus. The progenies of clones EET 183 and PLAYA ALTA 2, both hemizygous for the virus insert, segregated in the classic Mendelian ratio of 1:2:1 (Fig. 5). The progenies of the crosses EET 183 X CC 137 and PLAYA ALTA 2 X APA 4, female parent hemizygous for the virus insert, male parent negative, segregated in the expected ratio of 1:1 (Fig. 5).

Discussion

Our results clearly demonstrate the first evidence for the presence of badnaviral sequences integrated in cacao genomes, as determined by a range of methodologies employing bioinformatics, molecular analysis and genetic analysis. Following PCR evidence of 12 different viral types belonging to badnavirus species S or species S prime in asymptomatic trees of different genetic groups of cacao, bioinformatic analysis identified a variety of these viral sequences in a large number of cacao clones for which genomic data are available. Correlation between genetic grouping of reference cacao clones and mapping of specific inserts, and discovery of the same viral type in the

Specificity	Primer name	Primer sequence†	Ta (°C)	Amplicon size (bp)
RT RNase H region				
CSSV + S species	Badna1deg2	CCATCCCTTGGACHGCNTTYTGGGT	63	628
	Badna4deg2	TTACATACGGCNCCCCAHCCYTCC AT		
S species	Badna 1S	GATATACTTGTYTTYTCYAACAGCG	59	350
	Badna 4S	GATAAGATTCCARTCRCTDGCCGA		
2.2 kb fragment	S2465Fdeg	ACAGCCCCTRCYRTAAAGGGTCA	63	2202
	S4666R	GCCATAATACGGTAGAGGGAATCATG		
Inverted abutting primers				
NA 79	Badna14FL +	GTCAAAAATCTACCTGATATGCGG	57	2421
	Badna14FL-	GTTTGGCTTTGATCTGTCGGATAAG		
PA 279 junction primers				
Complete insert and flanking sequence of host	PA 279 Host F1	GGACTATGTTATGCTTGTCCCTT	63	243*
	PA 279 Host R1	ACCAGCCAATGTGAAAATGA		6293**
	PA 279 Host F2	CCATGTGACCATCGTTTCAA	63	635*
	PA 279 Host R2	CGGCTCCTTATTCCAAACAC		6685**
Genotyping				
Multiplex	PA 279 Host F1	GGACTATGTTATGCTTGTCCCTTT	63	142*
	PA 279 Ins R	TGGATGCACCAGTATCCAGA		495**
	PA 279 Ins F	TGGCTCCAATGGAAGTTAGC		679**
	PA 279 Host R3	CGGTTGAGAAGAATCCACCT		

Table 2. PCR primers used in this study. †Letters in primer sequences refer to International Union of Pure and Applied Chemistry (IUPAC) codes for nucleotides. *expected size in cacao clone PA 279 without viral type VI insertion. **expected size in cacao clone PA 279 with viral type VI insertion.

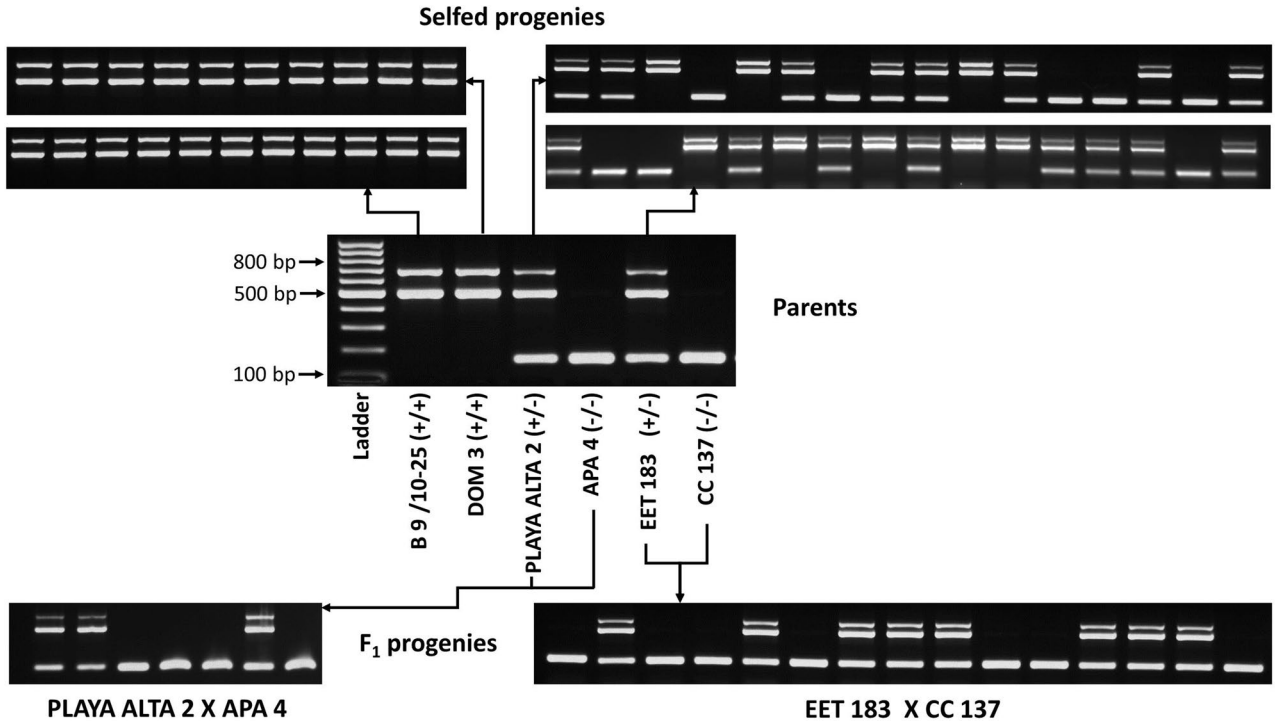


Figure 5. Segregation of viral type VI locus in selfed and crossed progenies of cacao clones. The parents and progenies were genotyped using a multiplex PCR assay including four primers (i.e. PA 279 Host F1, PA 279 Ins R, PA 279 Ins F and PA 279 Host R3).The 495 and 679 bp fragments represent virus + allele, left and right virus/host genome junctions, respectively. The presence of one 142 bp fragment represents virus–allele (−/−), lacking viral insertion. The presence of both 495 and 679 bp fragments in a genotype represents homozygous status (+/+) of the virus insertion locus, whereas amplification of 142, 495 and 679 bp fragments indicates hemizygous status (+/−) of the viral insertion locus. Full-length gels are presented in Supplementary Figs. 2–5.

same clone from multiple studies, confirm integration of the viral sequence in the host genome. The detailed study and sequencing of one specific type VI (Trinitario) insert in the PA 279 cacao clone revealed an insert of 6050 nt in the genome between position 32,885,707 and 328,857,725 of Chromosome V (referring to the numbering on B97 *T. cacao* genome). This approach could be extended to analyse other inserts in the future. We propose to call these integrations eTcBV1 and eTcBV2 for endogenous *Theobroma cacao* bacilliform virus 1 (species S defined as viral type I to XII) and 2 (species S prime) as suggested for the naming of endogenous viruses[5].

The overall results from the present study extend the list of plants harbouring EVEs, following the most recent description of EVEs in the genomes of *Citrinae*[24] and red raspberry[25]. Indeed, the cacao genome had already been mentioned as hosting caulimovirid (from the *Caulimoviridae* family) sequences, as reported in 27 plant species, in which sequences of a new proposed viral genus, Florendovirus, had been identified[26]. This particular viral genus is considered to have colonized the genomes of a diversity of species ranging from basal Angiosperms or ANITA grade to monocots and dicots. In 14 plant families, complete genomes were reconstituted, and Florendovirus-like sequences were also found to be present in EST databases[26].

Unlike these known viral species shown to be present in the genome of their host[10], in this work we have highlighted sequences belonging to unknown viral species whose existence as episomal particles and/or potential pathogenesis in cacao has not yet been demonstrated.

In summary, and in agreement with the findings for Citrinae[24] and Musaceae genomes[27], our results show a wide 'invasion' of cacao genomes by badnaviral sequences, significantly so for the cacao genetic groups Guiana, Iquitos, Marañon, Nanay, and Purús, along with the various admixed sub-groups. However, it is possible that we have in fact underestimated the presence of viral inserts, as the absence of PCR products could reflect a variation at the primer locus and does not exclude the presence of other, slightly different, badnaviral integration(s). Similarly, the absence of positive mapping in the four groups Contamana, Criollo, Curaray, and Nacional, and only in one genotype from the Amelonado group, is probably due to the specificity of the sequences used in the present analysis protocol.

As shown in the phylogeny that includes all badnaviral sequences found in cacao genomes (Fig. 1), the presence of PCR positives correspond to viral sequences significantly different from those associated with diseases caused by either CSSV in West Africa, CaMMV or CYVBV in Trinidad or CBSLV in Sri Lanka. Where present, such inserts are probably relics of ancient viral infections that occurred in the South American continent at the period of cacao genetic diversification[19]. As it has been discussed for badnaviral integration in bananas[27], it would be interesting, not only to extend the search to catalogue the complete range of viral inserts in cacao, but also to uncover both the time-line during which these various inserts took place, and any relationship to the concurrent diversification of the cacao genetic groups. There would seem to be two principal alternative possibilities for this process of integration and diversification. First, all the inserts could have occurred in the early cacao progenitor, followed by selective loss of most inserts during diversification, and consequential presence of a single predominant insert in each group. The second, and perhaps more likely alternative, is that specific inserts occurred in each group at the same time as the diversification into the various cacao groups, and their geographic separation[19]. At present, it would not seem possible definitively to distinguish between the alternatives. In addition, a search for badnavirus sequences in other co-located plant species in the upper Amazon basin may also allow the specific source of these cacao inserts to be identified.

Interestingly, it has been suggested previously that cacao swollen shoot viruses could be transmitted to cacao seeds[28] but the resultant adult cacao plants from the PCR-positive cacao seedlings never exhibit symptoms of swollen shoot. It could be hypothesized that the high sensitivity associated with the PCR technique could have amplified a virus titre that was too low for an infection to proliferate. PCR fragments were generated in those plants from a conserved region of ORF1 in CSSTBV and CSSTAV species but were sized via capillary electrophoresis and the products were not sequenced. Not all seedlings resulting from pollination were PCR-positive, and the study indicated that the seed transmission of functional, episomal CSSV was not probable. Results presented here on the crosses and progenies (including self progenies) clearly showed that, unlike episomal swollen shoot viruses, these integrated sequences are vertically transmitted to the progenies on a Mendelian basis.

We provide here, for the first time, definitive proof of viral sequence insertion in the cacao genome. This finding helps to explain the positive PCR for presence of virus in symptomless clones. However, the discovery of DNA sequences of the genus *Badnavirus* as integrated sequences in their plant host genome complicates the use of nucleic-acid based diagnostics for badnaviruses in the Cacao swollen shoot species complex that might infect the cacao tree and cause symptoms; this issue has been illustrated by the challenges experienced in reliable detection of banana streak viruses in *Musa* species[11,12,29]. Importantly, the presence of badnaviral inserts in most, if not all, cacao genetic groups, with no evidence of associated symptoms, suggests that there is no risk of spreading any disease by distributing biological material containing these inserted sequences. This issue has been discussed most recently in relation to the evidence for an integrated form of the Rubus Yellow Net Virus in the Red Raspberry genome[25].

Although this study provides valuable evidence for a variety of badnavirus sequences in the cacao genome, it leaves several important questions to be answered. One of the most obvious is whether the process of integration is completely random or whether there are specific sites into which such integration is most likely to occur. It is known that recombination can occur between badnavirus strains[30,31], and it may be that the integration process itself occurs preferentially into recombinogenic sequences such as remnants of transposons found widely in plant genomes.

The presence of viral sequences, in whole or in part, that are integrated into host genomes, also raises some interesting evolutionary questions. First, depending on the specific insertion site, one would expect that there may be a positive or negative impact on the phenotype of the plant. For any integrated sequences to be maintained within the genome, it must be assumed that there is a positive impact on the competitiveness of the host plant. If this were not the case, then the integrated sequence would be expected to degenerate over time by the

insertion of indels that would disrupt the sequence and then to be eliminated from the cacao genomes. In the present situation in cacao, it would seem that those viral sequences detected to date, though not complete, have maintained their integrity and perhaps therefore they provide some defensive role against the activity of other invasive and potentially damaging viruses by homology-dependent gene silencing[2,6], as has been suggested for example in yam[32]. This possibility offers the opportunity of manipulating the inserts, for example by gene editing as has been successfully demonstrated in certain genotypes of banana that host infective EVEs[33]. These targeted approaches can be considered as a supplement to the existing projects examining the potential for cross protection provided by mild strains of CSSTBV[34].

In summary, the data generated in this study add to the growing evidence for the dynamic status of plants (and other) genomes[35], such as evidence for the role of plant viruses as agents for horizontal gene transfer between species[36], and the role of such diversity in determining the response of plants to the continuous battle between plants and their pathogens.

Methods

Plant material. Fresh leaves of cacao clones from different genetic groups were supplied by CIRAD (Montpellier, France) and ICQC, R and ICG, T (International Cocoa Genebank, Trinidad and Tobago). To confirm whether there was evidence for sexual transmission of possible viral integrations, controlled self- and cross-pollinations were conducted at ICQC, R using clones determined to be positive and negative for the presence of badnaviral sequences. For cross-pollinations, petals and sepals, and then anthers were removed from the recipient flowers before pollination was conducted using the isolated anthers from the pollen donor clone. Following five to six months of pod development, the resultant seeds were collected and germinated. All experimental research conducted on plants complied with relevant institutional, national, and international guidelines and legislation.

Genomic DNA extraction. Total genomic DNA was isolated from cacao leaves using the Plant DNeasy kit (Qiagen) according to manufacturer's recommendations. Eighty milligrams of fresh leaves were ground with liquid nitrogen in a microcentrifuge tube in the presence of ceramic beads using a FastPrep-24 Classic (MP Biomedicals) homogenizer. Alternatively, five hundred milligrams of cacao leaves were frozen in liquid nitrogen and ground to a fine powder, which was then mixed with 5 mL of extraction buffer [100 mM tris(hydroxymethyl) aminomethane (Tris)-HCl, pH 8, 1.4 M NaCl, 20 mM ethylenediaminetetraacetic acid (EDTA), 2% w/v mixed alkyltrimethylammonium bromide, 1% w/v PEG6000 and 0.5% w/v Na_2SO_3 added freshly]. Samples were incubated at 74 °C for 30 min with 2 mg/mL RNase (Qiagen), extracted twice by 5 mL of chloroform–isoamyl alcohol (24:1) and precipitated with 5 mL of isopropanol at − 20 °C. DNA pellet was rinsed with EtOH and resuspended in 500 µL of sterile distilled deionized water. After quantification, DNA quality was assessed by PCR using the microsatellite mTC 351 primer pair (Table 2)[37].

Illumina sequencing of the ICS 76 clone from Trinidad. Extracted DNA from the ICS 76 clone (T3 tree) was sent to Fasteris S.A. (Geneva, Switzerland) for library preparation and sequencing using Illumina HiSeq rapid run technology, which resulted in paired-end reads of 250 bp mean length. Paired-end reads were trimmed using the Cutadapt script[38] to remove adaptors and filter for quality and were assembled using SPAdes v3.12.0[39] with k-mers ranging from 21 to 127 (21, 33, 43, 55, 77, 99, 127). All contigs were used to perform a BLAST analysis against a locally created database containing all available sequences representative of the cacao badnavirus S diversity to identify contigs (or scaffolds) containing cacao badnavirus S species.

PCR amplification. For detection of badnaviral sequences, two primer pairs were designed targeting RT RNase H badnaviral region. The first primer pair, which amplifies a 628 bp fragment, was designed by aligning all badnaviral sequences detected in cacao trees samples from West Africa as previously described[17]. This primer set is potentially able to detect all species associated with Cocoa swollen shoot disease, CSSTAV, CSSTBV, CSSCDV, CSSCEV, CSSGMV, CSSGNV, CSSGQV, CSSGTV, along with the badnaviral species S (not associated with complete genomes). The second primer pair was designed to amplify 366 bp, and specifically detects cacao badnavirus S species. In order to amplify a longer fragment of the genome of the cacao badnavirus species S, inverted abutting primers were designed in the RT RNase H fragment obtained from cacao clone NA 79. Alignment of sequence of the fragment amplified from NA 79 with inverted abutting primers and 3813 nt contig from ICS 76 allowed the design of new primers to amplify a fragment of ~ 2.2 kb fragment containing badnavirus S sequences (Fig. 2).

For detection, amplifications were performed using the Phire Hot Start II DNA Polymerase (Fisher Scientific) according to the manufacturer's recommendations. For long PCR, the Expand Long Template PCR system (Roche) was used following the manufacturer's instructions with the Expand Long Template buffer 3, an annealing of 57 °C and an elongation step of 6 min.

Two primer sets were designed flanking the insertion site by using information from alignment of PA 279 contig jcf7180010890274 and B97 *T. cacao* genome Chromosome V to amplify the viral insert along with the bordering host sequence from clone PA 279. The Phusion Green Hot Start II High-Fidelity PCR master mix (Fisher Scientific) was used to amplify the viral insert along with the bordering host sequence from PA 279. The amplicon was cloned with Zero Blunt TOPO PCR Cloning Kit for Sequencing (Fisher Scientific). The information obtained from sequencing of the viral insert amplified from PA 279 was then utilized to design a multiplex PCR assay for detection of the type VI insert in ICQC, R germplasm. Platinum Hot Start PCR Master Mix (Fisher Scientific) was used in the multiplex PCR assay.

BioProject	Number of Biosamples	Number of SRA experiments	Read length (bp)	Coverage/size (X)	Number of selected SRA experiments	References
PRJNA486011	200	200	74–202	5.3–74.5	200	[20]
PRJNA558793	31	93	275–300	20.8–88.5	31 + 2*	[47]
PRJNA77799	10	10	60	4.2–11.0	10	[21]
PRJNA421343	1	7	2981–12,905	5.4–14.0	1	[22]
PRJNA51633	8	14	361–869	0.6–2.1	1	[19]

Table 3. Whole Genome Sequence (WGS) datasets used in the study. *Dataset directly downloaded from the NSF (project submitter) website (https://plantscience.psu.edu/research/labs/guiltinan/nsf-plant-genome-research-program).

The PCR fragments and cloned PCR products were sequenced by Sanger technology (Eurofins Genomics, Germany and Source Bioscience, UK). Information about the primers used in this study including primer sequences, specific targets, and annealing temperatures for PCR are described in Table 2.

In silico screening of genomic data. The Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra) was searched for publicly available Whole Genome Sequence (WGS) and RNA sequence (RNA-Seq) datasets of *T. cacao* (as of April, 2020). The raw read files, of the searched datasets, in FastQ format, were downloaded from the ENA (https://www.ebi.ac.uk/ena/browser/home). The short reads were mapped using Bowtie2 v 2.3.4.1[40]. Minimap2 mapper[41] was used to align the long (MinION) reads. A database consisting of five badnavirus S sequences (derived from PCR fragments obtained with S2465Fdeg and S4666R primers, Table 2) was used as a reference in both cases. As described above, the five selected types, namely type I, type II, type III, type V and type VI, (a subset of the total of the 12 different clones initially identified) were chosen on the basis of ~ 2.2 kb viral fragments amplified from clones GU 230/C, PA 211, NA 79, IMC 55 and ICS 95, respectively. The mapped reads were compressed, sorted and indexed by Samtools v 1.10[41]. The alignment data were visualised in the Integrative Genomics Viewer (IGV) v.2.4.13[42]. Table 3 lists all the Bioprojects used in this study to search for viral sequences in cacao genomes.

Preliminary assembly of clone PA 279 was downloaded from the Penn State University website http://bigdata.bx.psu.edu/Cacao_NSF_data/. Blast analyses were conducted for the preliminary assembly of PA 279 (https://blast.ncbi.nlm.nih.gov/) using the badnavirus S type VI sequence as the query in order to identify the precise insertion site.

Phylogenetic analysis. Seaview version 4.0 software was used to analyse the DNA sequences, and these were aligned using the MUSCLE multiple alignment algorithm[43]. Phylogenetic relationships between CSSV sequences were estimated with PhyML (maximum likelihood method[44]) with SH-aLRT (approximate likelihood ratio test[45]) branch supports and phylogenetic trees were visualized with the Darwin 5 program[46].

Data availability

Sequencing data generated in this study were submitted to NCBI GenBank and have Accession Numbers MT993375 and MW009740-MW009807.

Received: 17 October 2020; Accepted: 22 March 2021
Published online: 15 April 2021

References

1. Feschotte, C. & Gilbert, C. Endogenous viruses: insights into viral evolution and impact on host biology. *Nat. Rev. Genet.* **13**, 283–296 (2012).
2. Staginnus, C. & Richert-Poggeler, K. R. Endogenous pararetroviruses: two-faced travelers in the plant genome. *Trends Plant. Sci.* **11**, 485–491 (2006).
3. Teycheney, P. Y. & Geering, A. D. W. Endogenous viral sequences in plant genomes. In *Recent Advances in Plant Virology* (eds Caranta, C. *et al.*) 343–362 (Caister Academic Press, Norfolk, 2011).
4. Chen, S. *et al.* Ancient endogenous pararetroviruses in *Oryza* genomes provide insights into the heterogeneity of viral gene macroevolution. *Genome Biol. Evol.* **10**, 2686–2696 (2018).
5. Geering, A. D. W., Scharaschkin, T. & Teycheney, P.-Y. The classification and nomenclature of endogenous viruses of the family *Caulimoviridae*. *Arch. Virol.* **155**, 123–131 (2010).
6. Mette, M. F. *et al.* Endogenous viral sequences and their potential contribution to heritable virus resistance in plants. *Embo J.* **21**, 461–469 (2002).
7. Staginnus, C., Iskra-Caruana, M. L., Lockhart, B., Hohn, T. & Richert-Poggeler, K. R. Suggestions for a nomenclature of endogenous pararetroviral sequences in plants. *Arch. Virol.* **154**, 1189–1193 (2009).
8. Richert-Poggeler, K. R. & Shepherd, R. J. Petunia vein-clearing virus: a plant pararetrovirus with the core sequences for an integrase function. *Virology* **236**, 137–146 (1997).
9. Gregor, W., Mette, M. F., Staginnus, C., Matzke, M. A. & Matzke, A. J. A distinct endogenous pararetrovirus family in *Nicotiana tomentosiformis,* a diploid progenitor of polyploid tobacco. *Plant Physiol.* **134**, 1191–1199 (2004).
10. Gayral, P. *et al.* A single Banana streak virus integration event in the banana genome as the origin of infectious endogenous pararetrovirus. *J. Virol.* **82**, 6697–6710 (2008).
11. Harper, G., Osuji, J. O., Heslop-Harrison, J. S. & Hull, R. Integration of banana streak badnavirus into the *Musa* genome: molecular and cytogenetic evidence. *Virology* **255**, 207–213 (1999).

12. Ndowora, T. *et al.* Evidence that Badnavirus infection in *Musa* can originate from integrated pararetroviral sequences. *Virology* **255**, 214–220 (1999).
13. Andres, C. *et al.* Combatting Cocoa Swollen Shoot Virus Disease: What do we know?. *Crop Prot.* **98**, 76–84 (2017).
14. Bhat, A. I., Hohn, T. & Ramasamy, S. Badnaviruses: The current global scenario. *Viruses* **8**, 177 (2016).
15. Lockhart, B. E. L. Evidence for a double-stranded circular DNA genome in a second group of plant viruses. *Phytopathology* **80**, 127–131 (1990).
16. Chingandu, N. *et al.* Molecular characterization of previously elusive badnaviruses associated with symptomatic cacao in the New World. *Arch. Virol.* **162**, 1363–1371 (2017).
17. Muller, E. *et al.* Next generation sequencing elucidates cacao badnavirus diversity and reveals the existence of more than ten viral species. *Virus Res.* **244**, 235–251 (2018).
18. Puig, A. *et al.* First report of Cacao mild mosaic virus (CaMMV) associated with symptomatic commercial cacao (*Theobroma cacao* L.) trees in Puerto Rico. *Plant Dis.* https://doi.org/10.1094/PDIS-04-20-0745-PDN (2020).
19. Motamayor, J. C. *et al.* Geographic and genetic population differentiation of the Amazonian chocolate tree (*Theobroma cacao* L.). *PLoS ONE* **3**, e3311 (2008).
20. Cornejo, O. E. *et al.* Population genomic analyses of the chocolate tree, *Theobroma cacao* L, provide insights into its domestication process. *Commun. Biol.* **1**, 167 (2018).
21. Kane, N. *et al.* Ultra-barcoding in cacao (*Theobroma spp.*; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. *Am. J. Bot.* **99**, 320–329 (2012).
22. Morrissey, J., Stack, J. C., Valls, R. & Motamayor, J. C. Low-cost assembly of a cacao crop genome is able to resolve complex heterozygous bubbles. *Hortic. Res.* **6**, 44 (2019).
23. Shahid, M. S., Aboughanem-Sabanadzovic, N., Sabanadzovic, S. & Tzanetakis, I. E. Genomic characterization and population structure of a Badnavirus infecting blackberry. *Plant Dis.* **101**, 110–115 (2017).
24. Yu, H. *et al.* Endogenous pararetrovirus sequences are widely present in *Citrinae* genomes. *Virus Res.* **262**, 48–53 (2019).
25. Diaz-Lara, A., Mosier, N. J., Stevens, K., Keller, K. E. & Martin, R. R. Evidence of Rubus Yellow Net Virus integration into the red raspberry genome. *Cytogenet. Genome. Res.* **160**, 329–334 (2020).
26. Geering, A. D. W. *et al.* Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution. *Nat. Commun.* **5**, 5269 (2014).
27. Chabannes, M. *et al.* Badnaviruses and banana genomes: a long association sheds light on *Musa* phylogeny and origin. *Mol. Plant Pathol.* **22**, 216–230 (2021).
28. Quainoo, A. K., Wetten, A. C. & Allainguillaume, J. Transmission of cocoa swollen shoot virus by seeds. *J. Virol. Methods* **150**, 45–49 (2008).
29. Le Provost, G., Iskra-Caruana, M.-L., Acina, I. & Teycheney, P.-Y. Improved detection of episomal *Banana streak viruses* by multiplex immunocapture PCR. *J. Virol. Methods* **137**, 7–13 (2006).
30. Ramos-Sobrinho, R., Chingandu, N., Gutierrez, O. A., Marelli, J. P. & Brown, J. K. A complex of Badnavirus species infecting cacao reveals mixed infections, extensive genomic variability, and interspecific recombination. *Viruses* **12**, 443–461 (2020).
31. Bomer, M., Rathnayake, A. I., Visendi, P., Silva, G. & Seal, S. E. Complete genome sequence of a new member of the genus *Badnavirus*, Dioscorea bacilliform RT virus 3, reveals the first evidence of recombination in yam badnaviruses. *Arch. Virol.* **163**, 533–538 (2018).
32. Umber, M. *et al.* The genome sequence of Dioscorea bacilliform TR virus, a member of the genus Badnavirus infecting *Dioscorea spp.*, sheds light on the possible function of endogenous Dioscorea bacilliform viruses. *Arch. Virol.* **162**, 517–521 (2017).
33. Tripathi, J. N. *et al.* CRISPR/Cas9 editing of endogenous banana streak virus in the B genome of *Musa* spp. overcomes a major challenge in banana breeding. *Commun. Biol.* **2**, 46 (2019).
34. Domfeh, O., Ameyaw, G. A. & Dzahini-Obiatey, H. K. Effects of cacao swollen shoot virus mild strains N1 and SS365B on growth and yield of cacao - a follow-up report. *Trop. Plant. Pathol.* **44**, 503–510 (2019).
35. Koonin, E. V., Makarova, K. S., Wolf, Y. I. & Krupovic, M. Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. *Nat. Rev. Genet.* **21**, 119–131 (2020).
36. Catoni, M. *et al.* Virus-mediated export of chromosomal DNA in plants. *Nat. Commun.* **9**, 5308 (2018).
37. Lanaud, C. *et al.* Isolation and characterization of microsatellites in *Theobroma cacao* L.. *Mol. Ecol.* **8**, 2141–2143 (1999).
38. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).
39. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
40. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
41. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
42. Robinson, J. T., Thorvaldsdottir, H., Wenger, A. M., Zehir, A. & Mesirov, J. P. Variant review with the integrative genomics viewer. *Cancer Res.* **77**, e31–e34 (2017).
43. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
44. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biol.* **52**, 696–704 (2003).
45. Anisimova, M., Gil, M., Dufayard, J-F., Dessimoz, C. & Gascuel, O. 2011. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Systematic Biol.*, syr041 (2011).
46. Perrier, X. & Jacquemoud-Collet, J. P. DARwin software, http://darwin.cirad.fr/darwin (2006)
47. Hamala, T. *et al.* Gene expression modularity reveals footprints of polygenic adaptation in *Theobroma cacao*. *Mol. Biol. Evol.* **37**, 110–123 (2020).

Acknowledgements

This project was supported by funding from Cocoa Research UK, The University of the West of England, CIRAD and Cocoa Research Association with United States Department of Agriculture as co-sponsors of the ICQC, R. We are also grateful to Claire Lanaud and Olivier Fouet (CIRAD) for providing DNA samples of cacao clones representative from the ten genetic groups and for helpful discussion. We acknowledge Gerben Martijn Ten Hoopen of the Cocoa Research Centre of the University of the West Indies (International Cocoa Genebank, Trinidad and Tobago) for providing leaf samples of the ICS 76 clone.

Author contributions

E.M., J.M.D., I.U., A.J.D., A.W. & J.A. conceived the project. E.M., I.U., M.R., A.W. and J.A. conducted all practical experiments and analysed results. E.M., I.U. & J.M.D. prepared the figures and wrote the first draft of the manuscript. All authors reviewed the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-87690-1.

Correspondence and requests for materials should be addressed to J.M.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.