Accepted Manuscript

Title: Next generation sequencing elucidates cacao badnavirus diversity and reveals the existence of more than ten viral species

Authors: E. Muller, S. Ravel, C. Agret, F. Abrokwah, H. Dzahini-Obiatey, I. Galyuon, K. Kouakou, E.C. Jeyaseelan, J. Allainguillaume, A. Wetten

PII: S0168-1702(17)30674-3
DOI: https://doi.org/10.1016/j.virusres.2017.11.019
Reference: VIRUS 97294

To appear in: *Virus Research*

Received date: 7-9-2017
Revised date: 15-11-2017
Accepted date: 18-11-2017

Please cite this article as: Muller, E., Ravel, S., Agret, C., Abrokwah, F., Dzahini-Obiatey, H., Galyuon, I., Kouakou, K., Jeyaseelan, E.C., Allainguillaume, J., Wetten, A., Next generation sequencing elucidates cacao badnavirus diversity and reveals the existence of more than ten viral species.Virus Research https://doi.org/10.1016/j.virusres.2017.11.019

Next generation sequencing elucidates cacao badnavirus diversity and reveals the existence of more than ten viral species

E. Muller[1], S. Ravel[1], C. Agret[2], F. Abrokwah[3], H. Dzahini-Obiatey[4], I. Galyuon[5], K. Kouakou[6], E.C. Jeyaseelan[7], J. Allainguillaume[8] A. Wetten[9].

1 CIRAD, UMR BGPI, 34398 Montpellier, France.
BGPI, Univ Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France.

2 CIRAD, UMR AGAP, 34398 Montpellier, France
AGAP, Univ Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France.

3 Department of Biochemistry, University of Cape Coast, Ghana.

4 Cocoa Research Institute of Ghana, P. O. Box 8, Akim Tafo, GHANA.

5 Department of Molecular Biology and Biotechnology, University of Cape Coast, Ghana.

6 World Cocoa Foundation, Abidjan, Côte d'Ivoire.

7 University of Jaffna, Faculty of Science, Department of Botany, Jaffna, Sri Lanka.

8 University of the West of England, Frenchay Campus, Coldharbour Lane, Bristol BS16 1QY, United Kingdom.

9 School of Agriculture, Policy and Development, University of Reading, Whiteknights, Reading, RG6 7BE, United Kingdom.

Corresponding author at : CIRAD, UMR BGPI, TA A-54/K, Campus international de Baillarguet, 34398 Montpellier Cedex 5, France. Tel.:+33 499 624 648; fax:+33 499 624 848
E-mail address: emmanuelle.muller@cirad.fr (E. Muller)

Highlights

- 21 new complete cacao badnavirus sequences reconstructed *de novo* thanks to the NGS.

- The taxonomic status of the molecular groups of cacao badnaviruses elucidated

- Ten species associated with cacao swollen shoot disease

Abstract

Cacao swollen shoot virus is a member of the family *Caulimoviridae*, genus *Badnavirus* and is naturally transmitted to *Theobroma cacao* (L.) by several mealybug species. CSSV populations in West African countries are highly variable and genetically structured into several different groups based on the diversity in the first part of ORF3 which encodes the movement protein. To unravel the extent of isolate diversity and address the problems of low titer and mixed viral sequences in samples, we used Illumina MiSeq and HiSeq technology. We were able to reconstruct *de novo* 20 new complete genomes from cacao samples collected in the Cocoa Research Institute of Ghana (CRIG) Museum and from the field samples collected in Côte d'Ivoire or Ghana. Based on the 20% threshold of nucleotide divergence in the reverse transcriptase/ribonuclease H (RT/RNase H) region which denotes species demarcation, we conclude to the existence of seven new species associated with the cacao swollen shoot disease. These new species along with the three already described leads to ten, the total number of the complex of viral species associated with the disease. A sample from Sri Lanka exhibiting similar leaf symptomology to West African CSSD-affected plants was also included in the study and the corresponding sequence represents the genome of a new virus named cacao bacilliform SriLanka virus (CBSLV).

Abbreviations

CSSV cacao swollen shoot virus
CSSCDV cacao swollen shoot CD virus
CSSTAV cacao swollen shoot Togo A virus
CSSD cacao swollen shoot disease
PCR Polymerase chain reaction
RT Reverse transcriptase
RNase H Ribonuclease H
ORF Open reading frame
$^{t}RNA^{Met}$ Methionyl transfer RNA
CRIG Cocoa Research Institute of Ghana
ICTV International Committee on Taxonomy of Viruses
CBSLV cacao bacilliform SriLanka virus
NGS Next generation sequencing

Keywords

cacao swollen shoot virus
Complete genomes
Badnavirus

Cacao
Phylogeny
Illumina sequencing

1. Introduction

Cacao swollen shoot disease (CSSD) which results from cacao swollen shoot virus (CSSV) infection is now regarded as the major viral disease affecting cacao and has been recognized as one of the most important diseases in West Africa limiting cacao production. CSSD was first described in Ghana at Effiduase in the New Juabeng district of the Eastern region in 1936 (Steven) although the disease was probably present in the nearby Nankese township of Ghana from 1922 (Paine, 1945). The disease subsequently appeared in all major cacao growing areas in West Africa with CSSD reported in Côte d'Ivoire in 1943 (Burle, 1961, Mangenot *et al.*, 1946), in Nigeria in 1944 (Thresh 1959), in Togo in 1949 (Partiot *et al.*, 1978) and in Sierra Leone in 1958 (Attafuah *et al.*, 1963). In addition, West African Amelonado cacao, planted uniformly throughout West Africa, appeared to be highly susceptible and sensitive to CSSV and has favored the rapid spread of the disease. CSSD has always been described as a disease endemic to West Africa, as it has never been reported in South America, the cacao tree's centre of origin. Additionally, CSSV has not been reported in Sao Tome, nor in Fernando Po (Tinsley, 1971), islands which were the main stepping stones of cacao introduction from the American continent towards West Africa. A viral disease causing similar leaf symptoms was reported in Trinidad (Kirkpatrick, 1953; Swarwick, 1961), but it is not associated with swellings. Following a government mandated-eradication program, this disease reappeared 14 years ago in the International Cocoa Genebank, Trinidad (ICGT) and two new badnaviruses have been characterized (Chingandu *et al*., 2017). The existence of CSSD in Malaysia, Indonesia and Sri Lanka (Kenten and Woods, 1976; Peiris, 1953; Crop Protection Compendium, 2002) has been mentioned but only as an attenuated form of CSSD. Additionally, in Malaysia, the disease is likely due to the importation of infected clones (Liu and Liew, 1979). To date, beyond West African cacao, swellings were only reported in Sri Lanka (Orellana and Peiris 1957).

Symptom variability between many different viral isolates has been noted from the first description of the disease in parallel with distinct designations in the different West African countries. However, isolate description by symptomology alone is inadequate for an understanding of the biology, origin and relationships between these different viruses.

Different types of serological diagnosis have been developed but to date these have insufficient polyvalence and sensitivity to address the high variability of the virus (review in Muller, 2008). PCR-based diagnosis holds more promise for the detection of latent infections (Muller *et al.*, 2001) and distinct molecular groups corresponding to different viral species (Kouakou *et al.* 2012, Abrokwah *et al.*, 2016) though even the use of degenerate primers has not consistently achieved viral fragment amplification from symptomatic leaves (Kouakou *et al.,* 2012; Abrokwah *et al.*, 2016). Furthermore, PCR diagnosis with degenerate primers cannot resolve the existence of mixed viral infections.

PCR primers have been designed in different parts of the genome, particularly, the first and third part of ORF3 corresponding respectively to the putative movement protein (ORF3A area) and to the reverse transcriptase/ribonuclease H (RT/RNase H) region. The first part of ORF3 (primers ORF3A-putative movement protein) was found to be highly conserved between the first six complete CSSV genomes and was therefore used both as a source of diagnostic primers (Muller and Sackey, 2005) and in variability studies to describe different CSSV molecular groups (Kouakou *et al.,* 2012; Abrokwah *et al*., 2016). To date, using the 80% nucleotide identity threshold in the RT/RNase H region (primers Badna 1/4 CSSV) and according to the recommendations of the International

Committee on Taxonomy of viruses (ICTV) (https://talk.ictvonline.org/ictv-reports/ictv_online_report/), we have described five different species responsible for CSSD: A (CSSTAV), B (CSSV), D (CSSCDV), G and M. However, for some samples, discrepancies between the two reconstructed phylogenies from ORF3A and RT/RNase H regions were observed indicating either recombination events between the two regions or the presence of mixed infection. There is a need therefore to study the complete genomic sequences corresponding to these isolates to characterise this diversity. Next generation sequencing technologies now offer an opportunity to resolve this dilemma and to complete the detection of cacao viruses without a priori sequence knowledge.

In 1944, twenty years after the discovery of CSSD in Ghana, the West African Cocoa Research Institute (WACRI) was established at Tafo, Ghana (later becoming the Cocoa Research Institute of Ghana-CRIG). Surveys across Ghana for CSSD began at the same time and a collection of viral isolates based on symptom description was established and named the CRIG Museum. This museum currently comprises more than 70 isolates in the form of potted, symptomatic cacao plants var. Amelonado (up to ten plants per isolate), which are maintained by regularly regrafting. This collection represents a valuable inventory of CSSV isolates from a range of past infected sites (albeit a collection that will have experienced sequence evolution in the intervening period).

Since 1999, attempts to describe the viral diversity present in the collection have been made using PCR amplification of specific regions of the CSSV genome followed by Sanger sequencing. This strategy has been hindered by a number of issues including the absence of symptoms for many plants, the lack of young leaves and the existence of mixed infections within the collection. However, plant screening with next generation sequencing technologies (Hiseq Illumina) can potentially address these problems of low titer and mixed viral sequences.

In the present study, 31 samples from the CRIG Museum and 14 samples recently collected from field sites in Côte d'Ivoire and Ghana and corresponding to new species and /or corresponding to samples with sequence inconsistency between the first and third part of ORF3 were analyzed via Illumina sequencing. A sample from Sri Lanka exhibiting similar leaf symptomology to West African CSSD-affected plants was also included in the study.

We were able to reconstruct 21 new complete genome sequences corresponding to species A (CSSTAV), B (CSSV), D (CSSCDV), E, M, N, R, Q along with the viral species responsible for the cacao disease in Sri Lanka. We therefore confirm that cacao swollen shoot disease is caused by a complex of species, all of which should be taken into account for effective control of the disease in the different West African countries.

2. Material and methods

2.1. Sample description, DNA extraction and PCR-sequencing analysis

Multiple samples from the CRIG Museum were collected in 1999, 2000, 2012, 2015, and 2016 (Table 1). Total DNA was extracted from symptomatic dried leaves with the Plant DNeasy kit (Qiagen) according to manufacturer's recommendations. Twenty milligrams of dried leaves was ground in a microcentrifuge tube in the presence of ceramic beads with a MP disrupter. To confirm the presence of CSSV in the samples, CSSV sequences were obtained by direct Sanger sequencing (Eurofins, MWG Operon) of PCR products amplified from the two regions ORF3A and Badna1/4 CSSV according to

Abrokwah *et al.* (2016). When direct sequencing failed, PCR products were cloned and several clones sequenced. The genome sequences have been deposited at DDBJ/EMBL/GenBank under the accession MF783897-MF784080.

In 2016, the number of plants harbouring the same isolate was recorded in the CRIG collection. To assess potential variability between replicate plants, two to ten leaves from separate plants were collected from 9 isolates/accessions (Gha26, Gha28, Gha30, Gha36, Gha39, Gha40, Gha53, Gha72 and Gha73).

Thirty one samples from the CRIG Museum collected in 1999, 2012, 2015 or 2016 were selected to be sequenced via Illumina technology. In addition, fourteen field samples from Ghana analysed in Abrokwah *et al.* (2016) and from Côte d'Ivoire analysed in Kouakou *et al.* (2012), corresponding to groups B, D, E, F, J, K or L have been included in this analysis (Table 2). A sample from Sri Lanka, supplied by the University of Jaffna's Department of Botany, exhibiting similar leaf symptomology to West African CSSD-affected plants was sourced from the Matale district (7°27'25.0"N 80°38'15.7"E) in 2015 and also included in the analysis.

2.2. Illumina DNA sequencing and *de novo* assembly

Extracted DNA underwent rolling circle amplification (RCA, TempliPhi kit, GE Healthcare Life Science) to concentrate/enrich the sample with circular forms and was sent to Fasteris S.A. (Geneva, Switzerland) for library preparation and sequencing using Illumina MiSeq along with Illumina HiSeq rapid run technology which resulted in paired-end reads of 250-bp mean length (Table 3 and 4). Paired-end reads were trimmed using the cutadapt script (Martin, 2011) to remove adaptors and filter for quality. The resulting reads were mapped with BWA (Li and Durbin, 2010) against the *Theobroma cacao* reference genome (Argout *et al.*, 2011; http://cocoa-genome-hub.southgreen.fr/). Unmapped reads were assembled using SPAdes v3.6.2 (Bankevich *et al.* 2012) with k-mers ranging from 21 to 127 (21, 33, 43, 55, 77, 99, 127), giving the number of contigs described in Table 3 and 4. All contigs were used to perform a BLAST analysis against a locally created database containing all available CSSV sequences in order to identify contigs demonstrating clear CSSV origin (Table 3 and 4). This CSSV database was continuously updated with the newly obtained sequences during the analysis.

The libraries were prepared with two different protocols depending on the DNA concentration of the sample after RCA: the Nano protocol was used from starting material greater than 100ng (samples HUX8 to HUX61, HUX104, HUX106 to HUX108, Tables 3 and 4) and the Nextera XT protocol is used from starting material less than 100ng (samples HUX68 to HUX88 and HUX105). The sequencing cover of the libraries obtained by the Nano protocol was double that of the sequencing cover of libraries obtained by the Nextera XT protocol.

As shown in Tables 3 and 4, the percentage of reads mapped to *T. cacao* varies from 28.6% (sample Gha52-15) to 66.6% (sample CI632-10). The number of contigs obtained with the SPAdes assembly varies from 36986 (corresponding to the samples with the lowest number of reads) to 137411 and from 0 to 3782 of these contigs have positive results in the BLAST analysis against a CSSV database.

2.3. Genome annotation

The Vector NTI Suite software (Vector NTI Advance® 11.5.2, Invitrogen, Life technology) was used to manually analyse the contigs, to assemble the smaller contigs, analyse ORFs with coding capacity for proteins larger than 10kDa, detect specific badnaviral motifs ($tRNA^{Met}$, RT and RNase H) and to confirm their badnaviral origin. The full genome sequences have been deposited at DDBJ/EMBL/GenBank under the accession MF642716- MF642736.

2.4. PCR and Sanger sequencing of contig junctions

When circularizing the viral contigs of each approximately complete sequence, ORFs situated on the junction between the end and the beginning of the linear contigs were sometimes interrupted. The sequences in these regions were obtained by designing a further set of PCR primers positioned on either side of the junction to amplify a product covering the relevant area (Supplemental Table 1). PCR products were sequenced by Sanger Technology.

2.5. Phylogenetic studies

Seaview version 4.0 software was used to analyze the DNA sequences and these were aligned using the MUSCLE multiple alignment algorithm (Edgar, 2004) and treated by Gblocks (Castresana, 2000). Phylogenetic relationships between CSSV sequences were estimated with PhyML (maximum likelihood method, Guindon and Gascuel, 2003) with SH-aLRT (approximate likelihood ratio test, Anisimova *et al*., 2011) branch supports and phylogenetic trees were visualized with the Darwin 5 program (Perrier and Jacquemoud-Collet., 2006).

3. Results

3.1. Analysis of the composition of the CRIG collection by direct PCR-sequencing or Illumina sequencing

From a total of 151 samples collected from 1999 to 2016 and nominally corresponding to 72 isolates, 130 were positive for CSSV (Sanger or Illumina sequencing) and corresponded to 71 different isolates. Only one out of the 72 sampled isolates (Gha74 collected in 2016) was negative for CSSV following PCR screen or Illumina sequencing (Table 1). Twenty plant samples were negative but found to be positive in other replicate plants corresponding to the same isolate.

For some samples, different sequences were observed for the same sample amplified twice in succession. For instance, samples Gha10-00 and Gha30-2-16 exhibited different sequences in the ORF3A section, and samples Gha2-00, Gha2-16, Gha29-16, Gha69-16 exhibited differences in the RT/RNase H region. Where a second sequence was obtained it was named using the format GhaXbis-XX. Some isolates (Gha8-00, Gha18-00, Gha35-15 and Gha49-15) also exhibited a mix of two sequences within a single sample as observed through Sanger sequencing of cloned PCR products.

For isolates Gha30 and Gha53, sequences were found to differ, depending on the replicate plant sampled in 2016. Gha30-1-16 and Gha30-2-16 contain sequences of group B and E+H respectively. Gha53-2 and Gha53-4 contain sequences of group P+R and B respectively (Table 1). For 22 isolates/accessions out of 72 (Gha2, Gha6, Gha13, Gha15, Gha16, Gha18, Gha22, Gha25, Gha26, Gha28, Gha29, Gha30, Gha34, Gha36, Gha40, Gha45, Gha49, Gha52, Gha53, Gha54, Gha64 and Gha68), sequences were found to differ depending on the year of collection.

Partial sequences have been obtained from most of the samples collected in 2000, 2011, 2012, 2015 and 2016 by Sanger technology. Sequences aligned in the ORF3A region (movement protein) and the

Badna1/4 CSSV region (RT/RNase H region) led to the construction of phylogenetic trees presented in Fig. 1A and 2A. In these phylogenies, the 14 field samples from Ghana and Côte d'Ivoire (Kouakou *et al.*, 2012; Abrokwah *et al.*, 2016) have been included for comparison. As with some field samples, for some samples from the CRIG Museum, sequences are only available for the RT/RNase H or the ORF3A region. For many other samples (27 samples from CRIG Museum and 6 field samples), sequences obtained from ORF3A and RT/RNase H region involve distortions between the two reconstructed phylogenies indicating recombination between the two regions or presence of mixed infection.

Partial and complete sequences were obtained by Illumina from 31 samples collected in 2012, 2015 or 2016 (Table 1). As expected, Illumina technology was able to detect more efficiently all the sequence groups present in one sample. Among the 151 samples analysed from the CRIG collection from 1999 to 2016, six of them (Gha58-16, Gha61-15, Gha67-12, Gha68-12, Gha69-12, Gha73-1-16) did not yield Sanger sequences but were analysed by Illumina technology. However, we observed three inconsistencies between sequencing approaches for samples Gha4-15, Gha66-12 and Gha2-16, with Illumina technology not able to detect the same sequences as those obtained by Sanger sequencing.

Of the 130 positive samples from the CRIG Museum, 80 samples contain sequences belonging to group B (and B-C species), 40 contain sequences belonging to group R, 16 to group G, 12 to group Q, 11 to group S, eight to group N, six to group A, five to group P, four to group E, four to group K, three to group M, one to group H and one to group T.

Forty seven samples had mixed infections (belonging to up to four groups) and these mixtures are more specifically associated with species R, Q and S. Of the 40 plants containing R species, only 12 plants had single infections with R species while on 12 plants containing Q species, only two plants had single infection. Species S, for which only partial sequences have been obtained, always occurred as part of a mixed infection. In contrast, only 16 plants out of a total 80 containing B species exhibited mixed infection. Other species are possibly present in the "mono infected "plants but were not detected.

3.2. New complete CSSV genomes reconstructed *de novo* by Illumina analysis

Twenty complete sequences were reconstructed *de novo* from cacao samples collected in the CRIG Museum and from the field samples collected in Côte d'Ivoire and Ghana. Numbering of the sequence follows badnavirus convention and refers to the plus-strand beginning at the 5'end of the minus strand replication priming site, the $tRNA^{Met}$ binding site (Harper and Hull, 1998). For GWR3-14, CI632-10 and GCR329-14, the junctions were sequenced by Sanger because ORF3 was interrupted in the *de novo* reconstructed draft sequence. The lengths of the complete genomes reconstructed *de novo* range from 6985bp (Gha53-15 isolate) to 7412bp (CI632-10 isolate) and are presented in Table 5. Open reading frames capable of encoding proteins larger than 10kDa with their locations on the plus-strand of the genome of CSSV isolates are described in Table 5. The sizes, numbers and arrangements of the different ORFs show similarity among the different CSSV genomes with two exceptions: the ORF1 of Gha37-15 which is slightly longer than that of other CSSV isolates (161 amino

acids instead of 138 to 156) and the slightly smaller ORF2 of Gha39-15 (98 amino acids instead of 120 to 149).

These new complete genomes meant we were able to obtain sequences for the RT/RNase H region of isolates from groups E, J, K and L which differed from those previously obtained via Sanger sequencing of that region using Badna 1/4 CSSV primers. Phylogenies constructed with the newly obtained RT/RNAse H sequences of E, J, K and L isolates removed the discrepancies in the phylogeny of the ORF3A region. Indeed those new RT/RNase H sequences belong to the same viral genomes as the ORF3A sequences obtained by Sanger previously in contrary to RT/RNase sequences obtained by Sanger methodology corresponding to another co-infecting viral sequence. We were also able to obtain sequences from the ORF3A region of isolates of the R and Q groups identified with the RT/RNase H phylogeny which have not been amplified with current ORF3A primers. To determine the relationships between the newly sequenced CSSV isolates, three phylogenetic trees were constructed from alignment of complete nucleotide sequences and the partial nucleotide sequences of the first part of ORF3 (ORF3A, movement protein) and the RT/RNase H region (Fig. 1B, 2B and 3).

Sequences from the RT/RNase H regions of isolates belonging to groups R, Q and N were identical regardless of the sequencing technology employed. This was also the case for sequences from ORF3A regions of isolates belonging to A, B, D, E, J, K, L, M and N groups.

Six complete genomes assembled from the NGS data belong to group R, four to group Q, one to group M, one to group N, one to group L, two to group E, one to group K, one to group J, one to group B (CI569-10), one to group D (CIDivo-15) and one to group A (Gha25-15). Two samples, Gha34-15 and GWR198, allowed the *de novo* assembly of two additional complete CSSV genomes belonging to two groups different from each other (differentiated by the letter corresponding to the sequence group). Sequences belonging to groups R and Q are clearly distant from the other CSSV groups on all the different phylogenies (Fig.1, 2 and 3) and seem to constitute a distinct new clade of CSSV sequences.

3.3. Complete sequence of a cacao virus from Sri Lanka

A complete sequence has been also reconstructed *de novo* from the sample collected in Sri Lanka. In Table 5, the size (7215bp), number (4) and arrangement of the different ORFs are described and are not different from CSSV isolates. The complete sequence has been included in the phylogenetic study of complete CSSV sequences and appears in another clade along with cacao yellow vein-banding virus (CYVBV) infecting cacao in Trinidad and recently sequenced (Chingandu *et al*., 2017) (Fig. 3). cacao mild mosaic virus (CaMMV) another complete sequence of a cacao virus infecting cacao in Trinidad (Chingandu et al., 2017) appears in the same clade as CSSV but very far from all CSSV sequences. Cacao bacilliform SriLanka virus (CBSLV) shares from 55.8% to 60.9% nucleotide sequence identity in the RT/RNase H region with other groups of CSSV sequences (Table 6).

3.4. Complex of species responsible for cacao swollen shoot disease

With respect to species responsible for cacao swollen shoot disease, Table 6 shows that by considering the 20% divergence threshold in the RT/RNase H region indicated by ICTV for the creation of new badnaviral species, we could define 10 distinct species: A, B (including subgroups B and C), D, E (including subgroups E, F, G, J, K and L) M, N, Q, R, S and T (Fig. 2B). Species A

corresponds to the newly created *Cacao swollen shoot Togo A virus* species (CSSTAV reference isolate ToWOB12, AJ781003, Muller and Sackey, 2005), species B corresponds to the first described *Cacao swollen shoot virus* species (reference isolate ToAgou1-93, L14546, Hagen *et al*., 1993) and species D corresponds to the newly described *Cacao swollen shoot CD virus* species (CSSCDV reference isolate CI152-09, JN606110, Kouakou *et al*., 2012). We propose to name 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 cacao swollen shoot Ghana R virus (CSSGRV), the new species E, M, N, Q, and R respectively for which we have the complete genomes. However, S and T species do not have complete sequences available as yet and we were not able to obtain the RT/RNase H sequence for the group P amplified from five different samples from the CRIG Museum by ORF3A primers and are not able to state yet if this group constitutes a new species.

3.5. Comparison of CSSV groups representations in the cacao farms and in the CRIG Museum

The calculation of the proportion of species present in the CRIG Museum considered mixed infection as the sum of different single infections. The total number of single infections in the collection corresponds to 191, the sum of infections by each species.

When considering only the region ORF3A, no mixed infection was detected in the field samples. However, all of these samples belonging to groups E, J, K and L when amplified with Badna1/4 CSSV primers exhibited sequences belonging to group S. Calculation of the proportion of isolates from the different groups present in the cacao farms therefore took into account only the sequences of the region ORF3A (115 sequences in total representative from the 830 samples collected in the six cacao growing regions).

As shown in Fig. 4, we observed that the B group is predominant in the CRIG collection (80 samples contain this group out of the 130 samples positive for CSSV) compared to the cacao farms (Supplemental Table 2 with group assignment of characterized CSSV samples from Abrokwah *et al.*, 2016). Additionally, the actual CSSV diversity in the cacao farms of the six cacao growing regions (8 groups + S group) is lower than the one observed in the CRIG Museum (13 groups). Additionally some groups were only present in the CRIG collection (G, N, P, Q and R) with others (C, L, K and J) exclusively present in the cacao farms and while E is present in both sets it is relatively underrepresented in the CRIG Museum.

4. Discussion

4.1. CSSV diversity in West Africa and in the CRIG Museum

Samples were collected several times between 1999 and 2016 at the CRIG Museum and most of the isolates were studied by direct Sanger sequencing. The fact that this collection contains all the main groups of CSSV isolates so far discovered in the cacao farms in West Africa signifies what a valuable resource this is for cacao researchers.

Successful detection of CSSV in samples was dependent on the year of collection and was probably influenced by the age of the leaves (Table 1). One out of the 72 isolates was negative and, for 22 isolates, depending on the sampling date or on the replicate plant collected in 2016, we did not detect the same CSSV group (Table 1). In the collection, each viral accession is maintained as multiple potted plants (two to ten) in a caged enclosure. While at least 50 m from other mealybug host plants,

the collection perimeter is not insect-proof and it is possible that small, wind-blown juvenile mealybugs could have occasionally alighted on the plants leading to inter-plant CSSV isolate transmission. Differences observed between sampling results could also be explained, by the possible mislabeling of some plants (when older plants are grafted on new seedlings), or by contamination due to accidental entry of mealybugs moving from one plant to another, or also by the presence of mixed infection and fluctuating concentrations of the two types of sequence. Virus might also disappear and then the plant could become re-infected.

The diversity of CSSV observed in the CRIG Museum compared to cacao farms in Ghana is interesting from a historical point of view, because this range can be seen as a picture of the diversity of the CSSV population at the time the Museum was established. The possibility of more recent external infections that have occurred after the establishment of the CRIG Museum should not be ruled out, but only with groups detected in the locality of CRIG. The groups G, N, P, Q and R currently only detected in the CRIG Museum may have been present previously in cacao farms but were absent or only present at a frequency that avoided detection in the farms sampled in the study of Abrokwah *et al.* (2016). Conversely, the groups of isolates C, J, K and L appear to correspond to recently emerged groups in the cacao farms. Group C has been detected in Togo since 1993 (Hagen *et al.*, 1994) but could have emerged either in Togo or in Ghana, since the establishment of the CRIG Museum. Group E, is underrepresented in the CRIG Museum but present in a high proportion in field samples (29% of the total isolates characterized) and appears to correspond to a group of isolates that has recently spread to cacao plots especially in the Western and Brong Ahafo regions of Ghana (Abrokwah *et al.*, 2016).

Considering these results, there would now be value in new CSSV prospections and further expansion of the collection to include all isolates detected recently in field samples. Cacao breeders need access to the full range of virus diversity to be found in the cacao field samples in order to develop genuinely disease resistant varieties for future replanting.

4.2. Mixed infection in the CRIG Museum

The presence of many mixed infections in the CRIG Museum compared to the situation in the cacao farms (where the only mixed infection detected was with the S species), may be explained by the fact that the plants have been infected for a long time and are surrounded by many other plants infected with different isolates with the potential for accidental cross contamination to occur via mealybug vectors. Ideally the Museum plants would be maintained within an insect-proof enclosure but this would require a barrier mesh size of ≤100 µm to exclude the smallest mealybug juveniles and has not proved to be practical in this region. The occurrence of mixed infection with sequences of different groups or species suggests the absence of cross protection between two isolates belonging to different groups (Folimova, 2013), though this would require further quantitative assessment of respective viral isolate titers.

For many sample sequences impacted by distortions between the two reconstructed phylogenies in ORF3A and RT/RNase H regions, a mixed infection was confirmed by two distant species, one amplified by one primer pair, the other amplified by the second. The possibility of a recombination event between two divergent viral isolates has therefore not been confirmed for any new isolate sequenced in this study. Additionally, recombination analysis on complete viral genome alignment

has been conducted using RDP4 software (Martin *et al.*, 2005) but without any conclusive results (data not shown).

4.3. Pathogenicity of species responsible for cacao swollen shoot disease

While the present study did not look for the presence of virions, all the full badnaviral genomes detected were derived from cacao leaf tissues exhibiting CSSD symptoms of the type shown by cacao infected with the CSSV isolate Agou 1 which was shown to contain viral particles with badnavirus morphology (Jacquot et al., 1999). The occurrence of distinct badnavirus sequences in West Africa cacao reported here supports the suggestion that CSSD arose from multiple instances of mealybug vectored infection of cacao from indigenous malvaceous species (Posnette et al., 1950) and it would be informative in future to examine such putative alternative hosts for badnaviral virions.

The complexity of the molecular diversity of viral species found in symptomatic cacao leaves means that there is a need to study the range of aggressiveness of the different species or subgroups. In future the symptomatology/aggressiveness associated with single strain infections arising from representative species should be explored as should the impact on hosts from mixed infections of these genomes. The species S, for example, never observed in single infection but present in the CRIG Museum and in the field samples in both Ghana and now Côte d'Ivoire needs to be studied more closely.

The two new species R and Q are peculiar because their sequences are quite distant from the other CSSV species and only detected in the CRIG Museum with many of them occurring in mixed infections. The complete sequences obtained will now facilitate the design of new primers specific for these species and support their detection in field samples. Since samples containing Q, R and S cannot be associated with typical symptomatology of CSSV even when in single infection in the CRIG Museum (11 samples for R, two samples for Q) and are not found in the field as single infections, there is a need to verify the particular pathogenicity of Q, R and S species.

4.4. Diagnostic of the complex of CSSV species

The results presented in this study allowed us to estimate the full diversity of the pathogens responsible for cacao swollen shoot disease and to evaluate the likely number of distinct species responsible for the disease in West Africa. Coincidentally, as with banana streak disease, cacao swollen shoot disease also appears to be a badnaviral disease caused by a complex of 10 different viral species (Iskra-Caruana *et al.*, 2014). This situation complicates the development of pathogen detection and diagnostic tools should now be improved to take into account all the highlighted diversity. With regard to DNA-based pathogen screening, it is unlikely that a single PCR assay can be developed to detect all the species responsible for the disease simultaneously. To be sufficiently sensitive for indexing purposes, QPCR could be established using a range of PCR primers able to detect all suspected CSSV species, though in the short term this is likely to be a laboratory rather than field tool.

As well as clarifying the diversity of West African CSSV isolates the NGS approach employed here also allowed for the first genome sequencing of a virus affecting cacao in Sri Lanka. The first report of viral symptoms on Sri Lankan cacao trees dates from 1953 (Peiris 1953) and described vein clearing patterns suggestive of a badnavirus disease. While there was no apparent reduction in vigour among

the affected plants, stem swellings were subsequently reported (Orellana and Peiris 1957) as were sporadic rounded pod symptoms and trials demonstrated that, as with CSSV, the disease was transmissible by multiple species of mealybugs (Carter, 1956). The isolate sequenced in our study also derives from a cacao orchard in the Central Province of Sri Lanka, within 20 km of the site of the 1953 disease report. While the recently affected trees are still not thought to be experiencing reduced vigour, genome information that facilitates the detection of this pathogen is of particular value since it appears to share more symptoms with the pathogenic West African forms than any other cacao badnavirus so far found outside that continent.

The new sequences generated in this work will help refine current molecular detection approaches for CSSV and support the development of novel field based screening methodologies for CSSV.

Acknowledgments

We wish to thank Sammy Sackey for samples collected from CRIG Museum in 2000.
This work was supported by the European Cocoa Association (ECA), CAOBISCO and the FCC.

References

Abrokwah, F., Dzahini-Obiatey, H., Galyuon, I., F., O.-A. & Muller, E. (2016) Geographical distribution of cacao swollen shoot virus molecular variability in Ghana. Plant Dis. 100, 2011-2017.

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. Syst. Biol., syr041.

Argout, X., Salse, J., Aury, J.M., Guiltinan, M.J., Droc, G., Gouzy, J., Allegre, M., Chaparro, C., Legavre, T., Maximova, S.N., Abrouk, M., Murat, F., Fouet, O., Poulain, J., Ruiz, M., Roguet, Y., Rodier-Goud, M., Barbosa-Neto, J.F., Sabot, F., Kudrna, D., Ammiraju, J.S., Schuster, S.C., Carlson, J.E., Sallet, E., Schiex, T., Dievart, A., Kramer, M., Gelley, L., Shi, Z., Berard, A., Viot, C., Boccara, M., Risterucci, A.M., Guignon, V., Sabau, X., Axtell, M.J., Ma, Z., Zhang, Y., Brown, S., Bourge, M., Golser, W., Song, X., Clement, D., Rivallan, R., Tahi, M., Akaza, J.M., Pitollat, B., Gramacho, K., D'Hont, A., Brunel, D., Infante, D., Kebe, I., Costet, P., Wing, R., McCombie, W.R., Guiderdoni, E., Quetier, F., Panaud, O., Wincker, P., Bocs, S. & Lanaud, C. (2011) The genome of *Theobroma cacao*. Nat. Genet. 43(2), 101-108.

Attafuah, A., Blencowe, J. & Brunt, A.A. (1963) Swollen shoot disease of cocoa in Sierra Leone. Tropical Agriculture, Trinidad 40, 229-232.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A. & Pevzner, P.A. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19(5), 455-477.

Burle, L. (1961). Le cacaoyer. G.-P. Maisonneuve et Larose, Paris.

Carter, W. (1956) Notes on some mealybugs (Coccoidae) of economic importance in Ceylon. FAO Plant Prot. Bull. VI(4), 49-52.

Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetics analysis. Mol. Biol. Evol. 17, 540-552.

Chingandu, N., Zia-Ur-Rehman, M., Sreenivasan, T.N., Surujdeo-Maharaj, S., Umaharan, P., Gutierrez, O.A. & Brown, J.K. (2017) Molecular characterization of previously elusive badnaviruses associated with symptomatic cacao in the New World. Arch. Virol. 162(5), 1363-1371.

Crop Protection Compendium (2002). *Cacao swollen shoot virus*. In Cacao swollen shoot virus. CABI Publishing, Wallingford, UK.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5), 1792-1797.

Folimonova, S.Y. (2013) Developing an understanding of cross-protection by Citrus tristeza virus. Front. Microbiol. 4, 76.

Guindon, S. & Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52(5), 696-704.

Hagen, L.S., Jacquemond, M., Lepingle, A., Lot, H. & Tepfer, M. (1993) Nucleotide sequence and genomic organization of cacao swollen shoot virus. Virology 196(2), 619-628.

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

Harper, G. & Hull, R. (1998) Cloning and sequence analysis of banana streak virus. Virus Genes 17, 271-278.

Iskra-Caruana, M.-L., Chabannes, M., Duroy, P.-O. & Muller, E. (2014) A possible scenario for the evolution of banana streak virus in banana. Virus Res. 186, 155-162.

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

Kenten, R.H. & Woods, R.D. (1976) A virus of the cocoa swollen shoot group infecting cocoa in North Sumatra. PANS 22(4), 488-490.

Kirkpatrick, T.W. (1953). *Insect pests of cacao and insect vectors of cacao virus diseases.* In Insect pests of cacao and insect vectors of cacao virus diseases. pp. 130-131. Imperial College of Tropical Agriculture.

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

Li, H. & Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinform. 26(5), 589-595.

Liu, P.S.W. & Liew, P.S.C. (1979). *Transmission Studies of a Cocoa Virus Disease (Yellow Vein-Banding) in Sabah*. In Transmission Studies of a Cocoa Virus Disease (Yellow Vein-Banding) in Sabah. Department of Agriculture, Sabah, Malaysia.

Mangenot, G., Alibert, H. & Basset, A. (1946) Sur les caractères du swollen shoot en Côte d'Ivoire. Rev. Int. Bot. Appl Tropical bulletin 283, 13.

Martin, D.P., Williamson, C. & Posada, D. (2005) RDP2: recombination detection and analysis from sequence alignments. Bioinform. 21(2), 260-262.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.j 17(1), 10-12.

Muller, E. (2008). *Cacao swollen shoot virus*. In Cacao swollen shoot virus Eds Mahy, B.W.J. & Van Regenmortel, M.H.V. pp. 403-409. Elsevier, Oxford.

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

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

Orellana, R.G. & Peiris, J.W.L. (1957) The swollen shoot phase of the virus disease of cacao in Ceylon. FAO Plant Prot. Bull. V(11), 165-168.

Paine, J. (1945). *Report of Agronomy Division, W.A.C.R.I.* In Report of Agronomy Division, W.A.C.R.I., unpublished.

Partiot, M., Djiekpor, E.K., Amefia, Y.K. & Bakar, K.A. (1978) Le "swollen shoot" du cacaoyer au Togo. Inventaire préliminaire et première estimation des pertes causées par la maladie. Café, Cacao, Thé 22(3), 217-228.

Peiris, J.W.L. (1953) A virus disease of cacao in Ceylon. Tropical Agriculture, Trinidad 109, 135-138.

Perrier, X. & Jacquemoud-Collet, J.P. (2006). *DARwin software*. http://darwin.cirad.fr/darwin.

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

Steven, W.H. (1936) A new disease of cacao in the Gold Coast. Gold Coast Farmer 5, 122-144.
Swarbrick, J.T. (1961) Cacao virus in Trinidad. Tropical Agriculture, Trinidad 38(3), 245-249.
Thresh, J.M. (1959) The control of cacao swollen shoot disease in Nigeria. Tropical Agriculture, Trinidad 36, 35-44.
Tinsley, T.W. (1971) The ecology of cacao viruses. I. The role of wilds hosts in the incidence of swollen shoot virus in West Africa. J. Appl. Ecol. 8(2), 491-495.

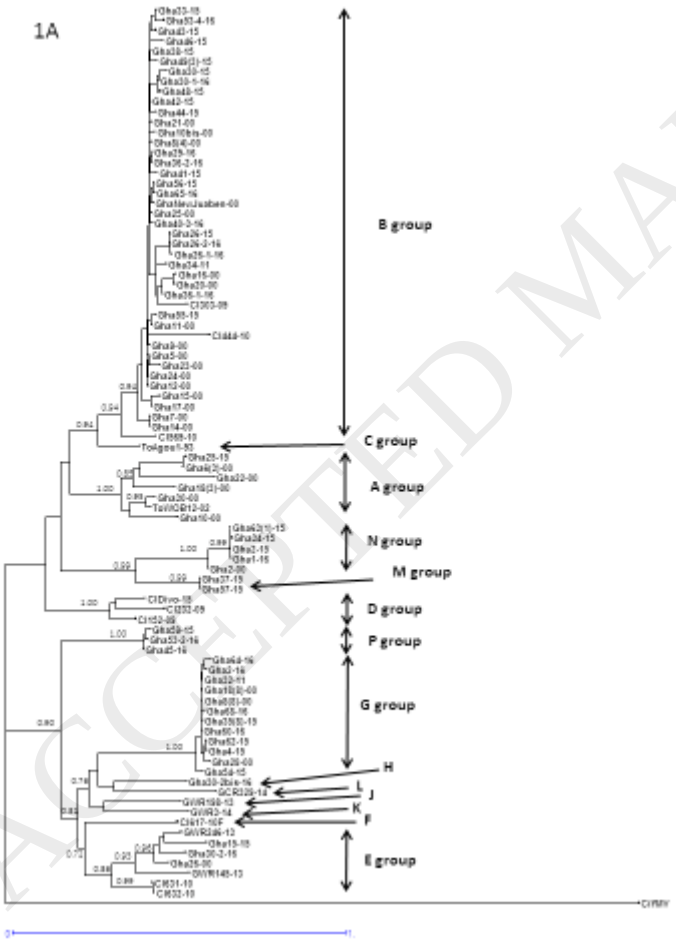
Figure 1. Maximum likelihood phylogenetic tree of CSSV sequences based on alignment of the 5' end of open reading frame 3 (ORF3). A. sequences obtained only by the Sanger technology. B. Sequences obtained by the Sanger and the Illumina technology. Numbers on the branches represent the SH-aLRT (approximate 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 to 2016 coded as 93 to 16).When two sequences are obtained for the same isolate with the two technologies, NGS is mentioned for the sequences obtained by Illumina technology and sequence is highlighted in grey. When sequences belonging to different groups are obtained for the same isolate, the letter corresponding to the group is affixed to the name of the isolate.

Figure 2. Maximum likelihood phylogenetic tree of CSSV sequences based on alignment of the RT/RNase H region of open reading frame 3 (ORF3). A. sequences obtained only by the Sanger technology. B. Sequences obtained by the Sanger and the Illumina technology. 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 to 2016 coded as 93 to 16).When two sequences are obtained for the same isolate with the two technologies, NGS or Sanger are indicated with the sequence and the sequence obtained with NGS is highlighted in grey. When sequences belonging to different groups are obtained for the same isolate, the letter corresponding to the group is affixed to the name of the isolate.

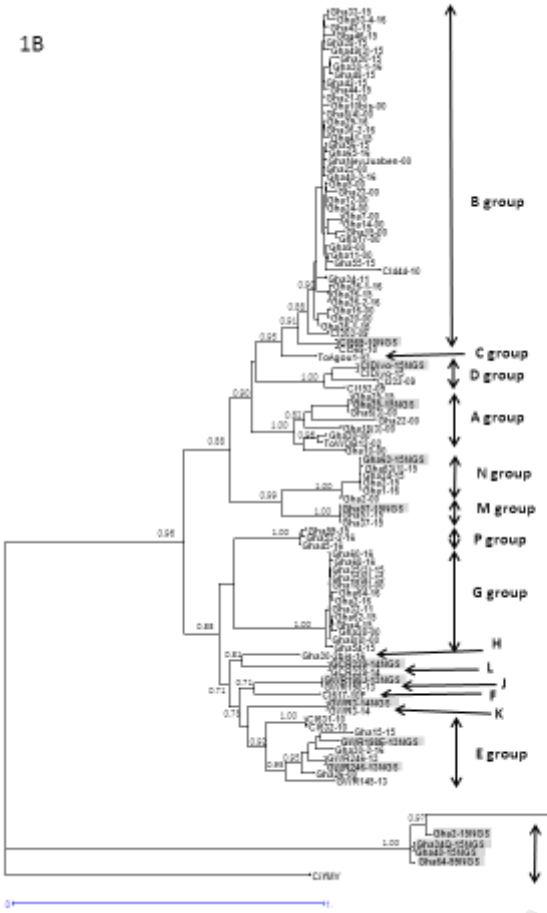
Figure 3. Maximum likelihood phylogenetic tree of CSSV sequences based on alignment of the complete viral genomes. Sequences in green correspond to viral species associated with CSSD, sequences in red correspond to other cocoa badnaviruses. Numbers on the branches represent the SH-aLRT (approximate likelihood ratio test) branch supports over 0.7. Genbank accession numbers of the additional badnaviral complete sequences used for comparative analysis are AJ002234 (*Banana streak OL virus*-BSOLV), AY805074 (BSMYV), AY750155 (BSVNV), AY493509 (BSGFV), HQ659760 (BSIMV), HQ593111 (BSCAV), HQ593107 (BSUAV), HQ593108 (BSUIV), HQ593109 (BSULV), HQ593110 (BSUMV), DQ092436 (BSAcYUV), EU034539 (*Bougainvillae chlorotic vein banding virus*, BCVBV), L14546 (CSSV-ToAgou1-93), JN606110 (CSSCDV-CI152-09), AJ781003 (CSSTAV-ToWOB12-02), KX276641 (cacao mild mosaic virus- CaMMV-), KX276640 (cacao yellow vein-banding virus- CYVBV), X52938 (*Commelina yellow mottle virus* -ComYMV-), AF347695 (*Citrus yellow mosaic virus* -CiYMV-), EU708317 (CiYMV IndiaAP), JN006806 (CiYMV-ROL), EU853709 (cycad leaf necrosis virus -CLNV-), X94576-581 (*Dioscorea bacilliform AL virus* -DBALV-), DQ822073 (*Dioscorea*

bacilliform SN virus -DBSNV-), DQ473478 (dracaena mottle virus –DrMV-), JF411989 (*Fig badnavirus 1* -FBV-1), HQ852249 (*Gooseberry vein banding associated virus* RC isolate -GVBV-RC-), HQ852251 (GVBV- RIB9001), HQ852250 (GVBV-BC), HQ852248 (GVBV-GB1), KT965859 (*Grapevine Roditis leaf discoloration-associated virus* BN-GRLDaV-), JF301669 (*Grapevine vein clearing virus* –GVCV-), KJ725346 (GVCV-VRU) KF875586 (hibiscus bacilliform virus-HiBV), AY180137 (*Kalanchoë top-spotting virus* -KTSV-), NC024301 (*Pagoda yellow mosaic associated virus*- PYMaV-), GQ428155 (pelargonium vein banding virus –PVBV-), GU121676 (*Pineapple bacilliform CO virus* –PBCoV-), GQ398110 (PBCoV HI1), KC808712 (*Piper yellow mottle virus*-PYMoV), M89923 (*Sugarcane bacilliform MO virus* -SCBMOV-), AJ277091 (*Sugarcane bacilliform IM virus* -SCBIMV-), FJ824813 (*Sugarcane bacilliform Guadeloupe A virus* -SCBGAV-), FJ439817 (*Sugarcane bacilliform Guadeloupe D virus* -SCBGDV-), NC012728 (*Sweet potato pakakuy virus* -SPPV-B), AF357836 (*Taro bacilliform virus*, TaBV), KP710177 (*Taro bacilliform CH virus*, TaBCHV1), KF241951 (*Rubus yellow net virus*-RYNV), KM078034 (RYNV-BS), KM229702 (*Yacon necrotic mottle virus*-YNMoV-).

Figure 4. Comparison of the molecular diversity of CSSV isolates between the CRIG Museum sampled from 1999 to 2016) and the cocoa farms in Ghana (sampled in 2013 and 2014)

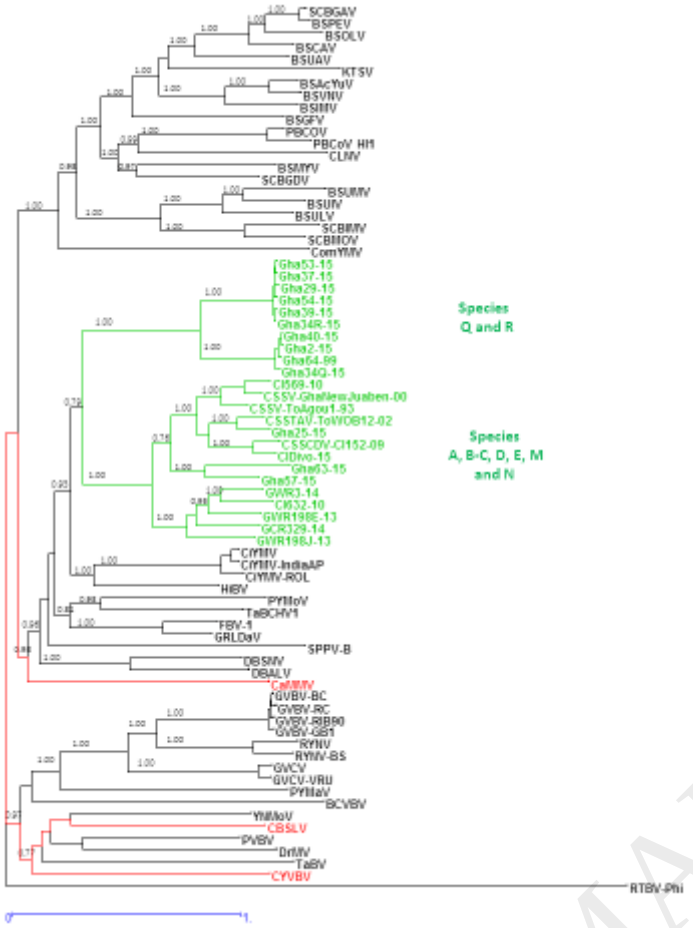


1B



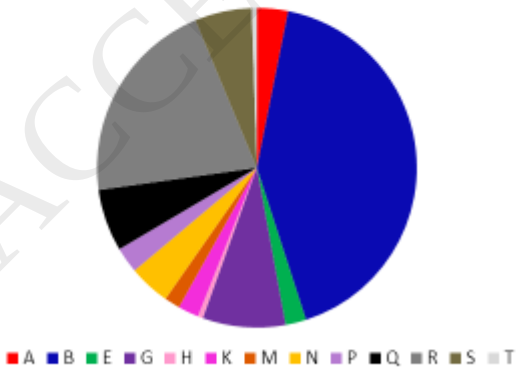
2A

2B



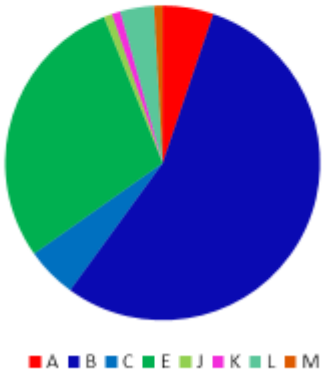
CRIG Museum diversity versus field CSSV diversity in Ghana

71 isolates (and 130 samples) analysed from CRIG (with 47 mixed infection)



Tafo CRIG Museum
13 groups

115 isolates from the 830 field samples of the six cacao growing regions



Cacao farms
8 groups

Table1. Group assignment and Illumina sequencing results for the CRIG Museum samples collected from 1999 to 2016.

ID	Isolate Name (region of origin)	Year of sampling	number of plants harbouring the same isolate recorded in 2016	Group assignment PCR ORF3A F/R	Group assignment PCR BADNA 1/4	Group assignment Illumina analysis	summary of analysis	size of the *de novo* reconstructed viral sequence (group) when superior to 1kb
1	Kpeve(Volta)	2000	3	N	Q		N+Q	
		2015		-	-			
		2016		N	-		N	
2	Gavekpe Todzi (Volta)	2000	1	N	T + Q		N +T +Q	
		2015		N	Q	Q + N+R	N+Q+R	7091bp (Q)
		2016		G	S +R	R+N	G+S +R+N	
3	Peki (Volta)	2000	3	B	X		B	
		2012		B	X		B	
4	Bisa (Eastern)	2000	4	-	-			
		2015		G	R	R +M +N	G+R+ M+N	3144bp (R) partial
5	Worawora (Volta)	2000	1	B	X		B	
		2012		B	B		B	
6	Domkorkrom (Eastern)	2000	2	A	B		A+B	
		2015		-	R		R	
		2016		-	-			
7	Tease Aduadum (Eastern)	2000	2	B	X		B	
8	Miaso (Eastern)	2000		B+G	X		B+G	
		2012	6	B	X		B	
9	Asamankese (Eastern)	2000	3	B	B		B	
10	Pamen (Eastern)	2000	3	A+B	B		A+B	
11	Djindji (Volta)	2000	4	B	X		B	
12	Kofi Pare (Eastern)	2000	7	B	B		B	
		2012		B	X		B	
13	Krofa Juansa F4T1 (Ashanti)	2000	5	B	B		B	
		2015		-	R		R	
14	Madjida Nkwanta (Ashanti)	2000	3	B	X		B	
		2012		B	X		B	
15	Bosomtwe Juaso (Ashanti)	2000	2	B	B		B	
		2015		E	R		E+R	
16	Bobiriso Juaso(Ashanti)	2000	4	B	B		B	
		2012		X	R		R	
		2015		-	R		R	
17	Konongo (Ashanti)	2000	5	B	X		B	
		2012		B	B		B	
18	Koben (Ashanti Region)	2000	3	A+G	X		A+G	
		2012		B	B		B	
19	Oyimso Agogo (Ashanti Region)	2000	6	B	B		B	
		2012		B	X		B	
20	Kwakoko Juansa (Ashanti)	2000	5	B	B		B	
		2012		B	B		B	
21	Bechem F1T1(Brong Ahafo)	2000	6	B	B		B	
		2012		B	B		B	
22	Okerikrom (Brong Ahafo)	2000	10	A	A		A	
		2015		-	R		R	
		2016		-	-			
23	Nkwanta (Near Dorma Ahenkro) (Brong Ahafo)	2000	9	B	B		B	
24	Nkrankwanta T1 (Brong Ahafo)	2000	5	B	X		B	
		2012		B	B		B	
25	Sankore T3/3 (Brong Ahafo)	2000	2	B	B		B	
		2012		B	B		B	
		2015		A	B	A+B	A+B	7229bp (A)
26	Punekrom (Western)	2000	8	E	S		E +S	
		2015		B	B		B	
26 1		2016		B	B		B	
26 2				B	-		B	
27	Surowno (Western)	2000	5	B	B		B	
		2012		B	X		B	
28	Bakukrom CC (Western)	2000	2	G	Q		G+Q	
		2011		G	S		G +S	
		2012		B	X		B	
		2015		G	R		G+R	
28 1		2016		-	-			
28 2				-	R		R	

29	SS167 (Eastern Region – CRIG PLOT)	2000	3	-	-			
		2015		-	R	R + B + S	R + B + S	7155 bp (R)
		2016		B	B+S		B +S	
30	SS365B (Eastern Region – CRIG PLOT)	2000	3	A	X		A	
		2015		B	B		B	
30 1		2016		B	B		B	
30 2				E + H	-		E+ H	
30 3				-	-			
32	CC Aboboya (Western)	2011	4	G	R		G+R	
		2015		-	R		R	
33	CC644 (Western)	2015	5	B	B		B	
34	CC Adembra (Western)	2011	3	B	X		B	
		2015		N	R	R+Q+N	N+ R +Q	6996bp (R) , 7343bp (Q)
35	Duasi Bosomtwe (Ashanti)	2015	2	G+ P	S		G+P +S	
36	Krofa Juansa F2T2 (Ashanti)	2015	5	-	R	R +G-K	R +G-K	1940bp (R) partial
36 1		2016		B	B		B	
36 2				B	B		B	
37	Amakom Bosomtwe (Ashanti)	2015	8	M	R	R+M	M+R	7012bp (R), 3707bp (M), 1223bp (M)partial, 1209bp (M) partial
38	Tease Adeakye (Eastern)	2015	1	B	B		B	
39	Dawa 1H (Estern)	2015	5	-	R	R	R	7097bp (R)
39 1		2016		-	-			
39 2				-	R		R	
40	Tafo Yellows (Eastern)	2015	6	-	Q	Q +S	Q+S	7102bp (Q)
40 1		2016		-	Q		Q	
40 2				-	-			
40 3				B	B		B	
41	Agyepomaaa (Eastern)	2012	4	B	B		B	
		2015		B	B		B	
42	Mampong 1M (Eastern)	2012	4	B	X		B	
		2015		B	B		B	
43	Dochi 1G (Eastern)	2012	4	B	X		B	
		2015		B	R		B+R	
44	Nkawkaw=1D (Eastern)	2015	4	B	B		B	
45	Nsba (Central)	2015	3	-	Q	Q +S	Q+S	2511bp (Q) partial
		2016		P	-		P	
46	Kwadzo Kumikrom J2/A (Brong Ahafo)	2015	.	B	B		B	
47	Kwadzo Kumikrom T2/2 (Brong Ahafo)	2015	4	B	B		B	
48	Kwadzo Kumikrom T1/6 (Brong Ahafo)	2015	not found in 2016 in the CRIG MUSEUM	B	B		B	
49	Takyimantia outbreak T3-15 (Brong Ahafo)	2012	6	B	B		B	
		2015		B +P	-		B+P	
50	Datano (Western)	2015	4	-	R	R +Q	R+Q	2513bp (R) partial
		2016		-	-			
51	Achiasi (Western)	2015	4	-	R	R+B	R+B	6243bp (R)
		2016		-	-			
52	Ayiboso (Western)	2012	4	B	B		B	
		2015		-	R	R+K	R+K	3657bp (R) partial
53	Bosomuoso 2 (Western)	2015	6	-	R	R	R	6985bp (R)
53 1		2016		-	-			
53 2				P	R		P+R	
53 3				-	-			
53 4				B	B		B	
54	Suhuma (Western)	2012	3	B	X		B	
		2015		G	B	R+B+G	B+G+R	6990bp (R)
55	Enchi E1 A3 (Western)	2015	4	B	R		B+R	
56	Enchi1/A/155 (Western)	2015	4	B	R		B+R	
57	Adjakaa - Enchi (Western)	2015	6	M	B	M+B	M+B	7009bp (M), 1745bp (B) partial, 1103bp (B) partial
58	Jamesi (Western)	2012	4	B	B		B	
		2015		-	R		R	
		2016		-	-	B	B	
59	CC Anibil (Western)	2015	4	P	-	P	P	1179bp (P) partial
60	CC Achechere	2015	3		R	R	R	4193bp (R) partial
		2016		G	R		G +R	
61	AD 196 (Eastern)	2012	1	B	B		B	
		2015		-	-	R	R	6377bp (R)
		2016		-	-			

62	AD 75 (Eastern)	2015	3	G	E	G+E	G +E	7239bp (G) discontinuousORF3, 5283bp (E) partial
63	AD 7 (Eastern)	2015	4	N	N	N+R+G	N+ R +G	7173bp (N)
64	Aiyim CC (Western)	1999	4	-	Q	Q	Q	7186bp (Q)
		2012		B	B		B	
		2016		G	R		G+R	
65	Amafié (Western)	2012	3	B	X		B	
		2016		B	R		B +R	
66	AD14 (Eastern)	2012		B	B	R	R + B	2862bp (R) partial
67	AD135 (Eastern)	2012	2	-	-	R+B	R+B	1849bp (R) partial
		2016		-	-			
68	Amanchia	2012	3	-	-	R	R	2348bp (R) partial
		2016		G	S		G + S	
69	Tease Atomsu Abuom (Eastern)	2012	5	-	-	Q+S	Q+S	
		2016		-	Q +S		Q +S	
71	SS75	2016		-	R	R + B +K	R + B + K	
72 1	Kwaku Anyan T1 (Brong Ahafo)	2016	4	-	-			
72 2				-	B	R+B+K	B+R+K	
72 3				-	-			
72 4				-	-			
73 1	Wiase (Western)	2016	5	-	-	B	B	
73 2				-	-			
73 3				-	-			
74	Enchi E/3/A (Western)	2016	3	-	-			

X not done, - PCR negative

Table 2. Group assignment and Illumina sequencing results for the field samples analysed from cocoa farms in Ghana and Côte d'Ivoire (Abrokwah *et al*., 2016, Kouakou *et al*., 2012).

Isolate Name	Country (Region)	Year of sampling	Group assignment ORF3A F/R	Group assignment BADNA 1/4	Group assignment NGS analysis	size of the *de novo* reconstructed viral sequence (group)
CI Divo-15	Côte d'Ivoire (Lôh-Djiboua)	2015	D	X	D	7205pb (D)
CI232-09	Côte d'Ivoire (Haut Sassandra)	2009	D	S	D	1663bp (D)
CI243-09	Côte d'Ivoire (Haut Sassandra)	2009	-	S	D	1952bp (D)
CI303-09	Côte d'Ivoire (Haut Sassandra)	2009	B	B	B	3544bp (B), 1054bp partial
CI444-10	Côte d'Ivoire (Marahoué)	2010	B	B	B	
CI569-10	Côte d'Ivoire (Guémon)	2010	B	B	B +S	7005bp (B)
CI617-10	Côte d'Ivoire (Mé)	2010	F	-	F	1837bp (F)
CI631-10	Côte d'Ivoire (Sud Comoé)	2010	E	-	E	5486bp (E) partial
CI632-10	Côte d'Ivoire (Sud Comoé)	2010	E	E	E	7412bp (E)
GCR329-14	Ghana (Central Region)	2014	L	S	L + S	6994bp (L)
GWR198-13	Ghana (Western Region)	2013	J	S	J+ E	7167bp (J), 7131bp (E)
GWR145-13	Ghana (Western Region)	2013	E	S	E+S	5766bp (E) partial
GWR3-14	Ghana (Western Region)	2014	K	S	K +S	7119bp (K)
GWR246-13	Ghana (Western Region)	2013	E	S	E+S	6971bp (E) partial

X not done, - PCR negative

Table 3. Results from the intermediary bioinformatics analysis of samples sequenced with MiSeq technology. Along with the library number, are indicated the numbers of paired reads obtained, the percentage of reads mapped against the *Theobroma cacao* reference genome, the number of contigs assembled from unmapped reads and the number of contigs resulting from the Blast search against the CSSV database.

Sample	Library name	Number of paired reads after adapter trimming	% of reads mapped to *Theobroma cacao*	number of contigs assembly with SPAdes	number of contigs with positive Blast against CSSV database
Sri Lanka	HUX-8	106 121	44.74	5068	1
G66-12	HUX-9	99 829	56.71	5110	1
G67-12	HUX-10	94 646	51.66	4613	0
G68-12	HUX-11	109 357	54.47	5595	1
G69-12	HUX-12	139 831	22.38	4254	0
CI303-09	HUX-13	115 172	43.42	7391	1
CI444-10	HUX-14	88 572	51.42	6286	0
CI569-10	HUX-15	111 485	59.61	5799	2
CI243-09	HUX-23	112 634	68.95	5583	0
GCR329-14	HUX-24	126 626	63.36	6082	3
GWR3-14	HUX-26	119 178	52.11	7646	2
G4-15	HUX-49	111 719	55.03	7125	1
G34-15	HUX-50	131 673	47.52	9257	6
G37-15	HUX-51	121 641	53.33	6804	11
G57-15	HUX-52	104 241	60.07	5821	1
G62-15	HUX-53	136 401	57.11	8338	11
G29-15	HUX-54	132 693	56.56	8327	11
G39-15	HUX-55	107 644	60.95	6729	2
G40-15	HUX-56	131 728	48.91	8826	1
G45-15	HUX-57	111 566	53.83	7148	1
G53-15	HUX-58	117 154	54.11	6889	1
G60-15	HUX-59	88 966	50.46	6119	5
G61-15	HUX-60	140 778	70.46	6723	7
G64-99	HUX-61	34 399	70.21	1962	2
CI617-10	HUX-68	113 634	58.80	4423	2
CI631-10	HUX-69	83 550	45.43	3735	5
CI632-10	HUX-70	119 549	71.60	3761	4
CI232-09	HUX-71	94 631	75.56	2775	0
GWR145-13	HUX-72	108 351	64.22	3973	2
GWR198-13	HUX-73	88 034	50.73	4239	5
GWR246-13	HUX-74	96 792	51.92	3491	11
G2-15	HUX-80	83 752	49.57	3619	151
G25-15	HUX-81	86 268	55.99	3523	12
G54-15	HUX-82	85 838	51.75	2451	30
G59-15	HUX-83	59 463	50.76	2510	0
G63-15	HUX-84	104 565	55.06	3314	5
G36-15	HUX-85	109 897	49.25	4552	6
G50-15	HUX-86	105 154	50.30	4346	5
G51-15	HUX-87	28 954	59.87	980	1

G-52-15	HUX-88	98 694	30.76	2909	57

Table 4. Results from the intermediary bioinformatics analysis of samples sequenced with HiSeq technology. Along with the library number, are indicated the numbers of paired reads obtained, the percentage of reads mapped against the *Theobroma cacao* reference genome, the number of contigs assembled from unmapped reads and the number of contigs resulting from the Blast search against the CSSV database.

Sample	Library name	Number of paired reads after adapter trimming	% of reads mapped to *Theobroma cacao*	number of contigs assembly with SPAdes	number of contigs with positive Blast against CSSV database
CIDivo-15	HUX 6	3 871 266	29.5	324*	6
Sri Lanka	HUX-8	7 651 347	33.6	88986	2
G66-12	HUX-9	7 259 732	40.6	89277	4
G67-12	HUX-10	6 634 320	39.3	80477	6
G68-12	HUX-11	7 833 981	42.6	93656	6
G69-12	HUX-12	10 765 222	15.3	79917	4
CI303-09	HUX-13	8 331 341	31.9	129914	2
CI444-10	HUX-14	6 081 563	38.5	109523	1
CI569-10	HUX-15	7 511 965	46.3	107737	2
CI243-09	HUX-23	7 661 156	53.1	89677	2
GCR329-14	HUX-24	8 455 948	50.6	111110	6
GWR3-14	HUX-26	7 648 799	41.6	116905	4
G4-15	HUX-49	7 764 075	39.4	108776	8
G34-15	HUX-50	8 772 851	33.8	122556	84
G37-15	HUX-51	8 354 315	38.2	124983	27
G57-15	HUX-52	7 120 254	42.9	95711	6
G62-15	HUX-53	9 006 793	42.5	120841	89
G29-15	HUX-54	8 795 657	41.1	133270	710
G39-15	HUX-55	7 496 393	42.7	102297	72
G40-15	HUX-56	9 158 600	36.6	137411	4
G45-15	HUX-57	7 872 645	38.0	111953	4
G53-15	HUX-58	7 844 567	39.1	102588	25
G60-15	HUX-59	6 021 394	36.9	93326	6
G61-15	HUX-60	9 824 866	50.6	108354	4
G64-99	HUX-61	2 287 035	52.2	36986	4
CI617-10	HUX-68	4 561 202	49.2	102620	65
CI631-10	HUX-69	3 398 624	38.3	56504	91
CI632-10	HUX-70	4 819 129	59.2	112930	11
CI232-09	HUX-71	4 579 631	66.6	102975	1
GWR145-13	HUX-72	4 517 773	58.1	109927	3
GWR198-13	HUX-73	3 342 286	42.9	94395	25
GWR246-13	HUX-74	3 167 751	45.5	43221	179
G2-15	HUX-80	3 960 559	44.0	81357	1067
G25-15	HUX-81	3 851 226	47.0	85891	8

G54-15	HUX-82	3 700 614	45.0	63992	3782
G59-15	HUX-83	2 999 875	42.1	67724	3
G63-15	HUX-84	4 278 299	52.4	58305	525
G36-15	HUX-85	5 306 802	40.6	104246	22
G50-15	HUX-86	4 179 889	45.5	94750	27
G51-15	HUX-87	3 648 139	52.7	79073	173
G-52-15	HUX-88	4 621 421	28.6	45350	3649
G2-16	HUX-104	2'217'329	71.8	222937	566
G58-16	HUX-105	3'558'575	56.9	32001	73
G72-16-2	HUX-106	3'722'639	64.2	107031	189
G71-16	HUX-107	3'643'270	70.8	194291	2419
G73-16-1	HUX-108	3'269'021	65.17	107156	1

* Assembly pipeline different from SPAdes

Table 5. Protein-coding regions located on the plus-strand of the genome of CSSV isolates. Highlighted gray lines are previously sequenced complete genomes. Genbank accession numbers are provided after the isolate names.

Isolate name	**Number of amino acids** (starting nucleotide-ending nucleotide*)						**Sequence**
(group)	**ORF 1**	**ORF2**	**ORF3**	**ORFX**	**ORF4**	**ORFY**	**Size (bp)**
ToWobe12-02 (A) AJ781003	**143** (432-860)	**149** (860-1306)	**1868** (1275-6878)		**NC**	**131** (6563-6955)	**7297**
Gha25-15 (A) MF642716	**143** (407-835)	**143** (835-1263)	**1862** (1232-6817)	**NC**	**NC**	**126** (6517-6894)	**7229**
GhaNewJuaben-00 (B) AJ608931	**143** (294-722)	**145** (722-1156)	**1847** (1125-6665)	**91** (2308-2580)	**NC**	**131** (6308-6700)	**7024**
CI569-10 (B) MF642717	**143** (296-724)	**143** (724-1152)	**1841** (1121-6643)	**NC**	**101** (4176-4478)	**131** (6286-6678)	**7005**
ToAgou1-93 (C) L14546	**143** (441-869)	**132** (869-1264)	**1834** (1272- 6773)	**113** (2374-2712)	**NC**	**131** (6434-6826)	**7161**
CI152-09 (D) JN606110	**143** (318-746)	**139** (746-1162)	**1872** (1152-6767)	**NC**	**97** (4063-4353)	**130** (6455-6844)	**7203**
CIDivo-15 (D) MF642718	**153** (285-743)	**144** (743-1174)	**1888** (1146-6809)	**145** (2212-2646)	**NC**	**130** (6497-6886)	**7205**
CI632-10 (E) MF642719	**154** (266-727)	**144** (727-1158)	**1855** (1124-6688)	**NC**	**NC**	**130** (6373-6762)	**7412**

GWR198E-13 (E) MF642720	**156** (476-943)	**145** (943-1377)	**1822** (1343-6808)	**NC**	**NC**	**146** (6442-6879)	**7131**
GWR198J-13 (J) MF642721	**143** (273-701)	**144** (701-1132)	**1868** (1098-6701)	**107** (2194-2514)	**NC**	**130** (6335-6724)	**7167**
GWR3-14 (K) MF642722	**143** (276-704)	**142** (704-1129)	**1881** (1095-6737)	**93** (2248-2526)	**NC**	**130** (6371-6760)	**7119**
GCR329-14 (L) MF642723	**148** (233-676)	**142** (676-1101)	**1853** (1067-6625)	**114** (2157-2498)	**NC**	**130** (6259-6648)	**6994**
Gha57-15 (M) MF642724	**143** (331-759)	**146** (759-1196)	**1831** (1174-6666)	**NC**	**NC**	**130** (6300-6689)	**7009**
Gha63-15 (N) MF642725	**144** (287-718)	**125** (718-1092)	**1881** (1092-6734)	**NC**	**NC**	**132** (6215-6610)	**7173**
Gha2-15 (Q) MF642726	**138** (284-697)	**122** (700-1065)	**1853** (1065-6623)	**NC**	**NC**	**115** (6305-6649)	**7091**
Gha34Q-15 (Q) MF642727	**138** (279-692)	**125** (683-1057)	**1941** (1057-6879)	**NC**	**NC**	**132** (6510-6905)	**7343**
Gha40-15 (Q) MF642728	**154** (236-697)	**125** (688-1062)	**1859** (1062-6638)	**NC**	**NC**	**132** (6269-6664)	**7102**
Gha64-99 (Q) MF642729	**138** (284-697)	**121** (700-1062)	**1858** (1062-6635)	**NC**	**NC**	**115** (6317-6661)	**7186**
Gha29-15 (R) MF642730	**138** (298-711)	**120** (714-1073)	**1839** (1073-6589)	**NC**	**NC**	**125** (6241-6615)	**7155**
Gha34R-15 (R) MF642731	**138** (295-708)	**124** (699-1070)	**1833** (1070-6568)	**NC**	**NC**	**132** (6199-6594)	**6996**
Gha37-15 (R) MF642732	**161** (243-725)	**124** (716-1087)	**1833** (1087-6585)	**NC**	**NC**	**132** (6216-6611)	**7012**
Gha39-15 (R) MF642733	**138** (280-693)	**98** (696-989)	**1833** (1182-6680)	**NC**	**NC**	**125** (6332-6706)	**7097**

Gha53-15 (R)	**138**	**124**	**1833**	**NC**	**NC**	**132**	**6985**
MF642734	(295-708)	(699-1070)	(1070-6568)			(6199-6594)	
Gha54-15 (R)	**138**	**124**	**1833**	**NC**	**NC**	**132**	**6990**
MF642735	(429-842)	(833-1204)	(1204-6702)			(6333-6728)	
cacao Sri Lanka virus MF642736	**143**	**131**	**1772**	**NC**	**NC**	**133**	**7215**
	(184-612)	(612-1004)	(1004-6319)			(5995-6393)	

* Without including the stop codon

NC : No Corresponding ORF

Table 6. Nucleotide sequence identity of pairwise combinations of representative CSSV isolates of each molecular group in the RT/RNase H region of ORF3 (region Badna 1/4 CSSV). The 4 proposed clades consist of the species A, BC, D, EGJKL, M and N on the one hand, Q and R on the other hand, S for the 3rd and T for the 4th.

CSSV groups	A ToWOB12-02	B GhaNew Juaben-00	C ToAgou1-93	D CI152-09	E CI632-10	G Gha62-15	J GWR198-13	K GWR3-14	L GCR329-14	M Gha57-15	N Gha63-15	Q Gha2-15	R Gha29-15	S CI232-09	T Gha2-00	Cacao SriLanka Virus
A	100	73.4	75.5	75.1	72.1	69.7	70.1	71.5	70.9	70.1	68.6	62	59.1	57.8	55.3	60.9
B		100	88.5	71.7	73.4	73.4	69.5	75.7	74.4	73.6	69.9	58.9	55.6	58.4	58.9	57.8
C			100	71.3	73.2	74.2	69.5	76.1	73.8	74.4	70.9	61.4	59.3	59	60.3	60.7
D				100	68.8	68.8	67.6	68.8	68.8	68.8	67.2	63.6	62	58.4	59.9	59.3
E					100	81.1	81.5	86	81.7	71.5	70.7	59.9	55.3	60.9	59.3	57.2
G						100	78.4	82.1	82.9	70.1	70.7	59.9	56	59	57.6	55.8
J							100	80.4	79.8	70.9	68	60.5	58	61.1	55.3	55.8
K								100	83.4	70.5	72.1	59.1	57	60.1	58	57.4
L									100	71.9	72.8	58.9	57.8	62.8	59.7	57.2
M										100	75.7	61.6	58.5	59.9	55.8	56.2
N											100	59.1	57.6	61.7	55.3	56.6
Q												100	76.7	61.3	61.2	59.1
R													100	58	63.2	57
S														100	59.7	56.6
T															100	59.3
CSLV																100