2017 International Symposium on Cocoa Research (ISCR), Lima, Peru, 13-17 November 2017

A Next Generation Sequencing approach to elucidate CSSV species profiles

Emmanuelle Muller1, Sebastien Ravel1, Andrew Wetten2, Joel Allainguillaume3, Francis Abrokwah4, Koffié Kouakou5 Henry K. Dzahini-Obiatey6

1 CIRAD,Montpellier, France,

2 University of Reading, Reading,United Kingdom,

3 University of the West of England, Bristol, United Kingdom,

4 University of Cape Coast, Cape Coast, Ghana,

5 World Cocoa Foundation, Abidjan, Côte d'Ivoire,

6 Cocoa Research Institute of Ghana, Akim Tafo, Ghana.

Cacao swollen shoot virus (CSSV) is a member of the family Caulimoviridae, genus Badnavirus and is naturally transmitted to *Theobroma cacao* by several mealybug species. The virus is restricted to West Africa, while the cacao tree originates from the Americas, and has therefore most probably an indigenous origin on the West African subcontinent. The resultant disease has caused enormous economic damage in Ghana since the 1930's but was restricted to small areas in Togo and Côte d'Ivoire until recently. Now, renewed outbreaks in the main producing areas in Côte d'Ivoire, Ghana and Togo, cause serious yield losses and tree death.

CSSV populations in West African countries are genetically structured into several different groups according to the diversity in the first part of ORF3 corresponding to the movement protein. To unravel the extent of isolate diversity we used Illumina HiSeq technology and reconstructed 21 new complete genomes corresponding to the different groups of CSSV sequences. In this way we were able to compare the partial sequences of the RTase region (recognised as the taxonomical region by ICTV using a 20% threshold of nucleotide divergence to denote separate species), and thereby identifying nine different CSSV species. These results will now be used to improve the detection of all badnaviruses present in cacao leaf samples, a vital tool in efforts to halt the spread of the disease and confirm the healthy status of new plantations.

Introduction

Cacao swollen shoot disease (CSSD) which results from CSSV (Cacao swollen shoot virus) 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. 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.

PCR-based diagnosis holds quite 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.iseases in West Africa limiting cacao production.

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 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, B-C, D, 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.

Since 1999, attempts to describe the viral diversity present in the CRIG Museum collection have started using PCR amplification approaches 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.

We were able to reconstruct 21 new complete genome sequences corresponding to species A, B-C, D, M, N, R, Q, the species E-G-J-K-L 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.

Materials and Methods

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). The genome sequences have been deposited at DDBJ/EMBL/GenBank under the accession MF783897-MF784080.

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 in 2015 and also included in the analysis.

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. 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). 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). All the contigs obtained 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.

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 (tRNAMet, 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.

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. PCR products were sequenced by Sanger Technology.

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). 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.

Results

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.

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. 1 and 2. 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.

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.

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 or Ghana. 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 length of the complete genomes reconstructed de novo ranges from 6985bp (Gha53-15 isolate) to 7412bp (CI632-10 isolate) and are presented in Table 3. The sizes, numbers and arrangements of the different ORFs show similarity among the different CSSV genomes

These new complete genomes meant we were able to obtain sequences for the RT/RNase H region of isolates from group E, J, K and L and they were different from those previously obtained via Sanger sequencing of the RT/RNAse H region with 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 correspond 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. 1, 2 and 3).

Six complete genomes assembled form 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).

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 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 4).

Complex of species responsible for cacao swollen shoot disease

With respect to species responsible for cacao swollen shoot disease, Table 4 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, BC, D, EGJKL, M, N, Q, R, S and T (Fig. 2). 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.

Discussion

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 permitting the potential entry and movement of mealybugs, the vectors of CSSV. Differences observed between sampling results could 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.

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 were potentially present in the past in different 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.

Pathogenicity of species responsible for cacao swollen shoot disease

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.

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). 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.

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.

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., 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.

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.

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.

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.

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.

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

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 PCR ORF3A F/R	Group assignment PCR BADNA 1/4	Group assignment Illumina analysis	size of the *de novo* reconstructed viral sequence (group) when superior to 1kb
CIDivo-15	Côte d'Ivoire (Lôh-Djiboua)	2015	D	X	D	7205bp (D)
CI232-09	Côte d'Ivoire (Haut Sassandra)	2009	D	S	D	1663bp (D) partial
CI243-09	Côte d'Ivoire (Haut Sassandra)	2009	-	S	D	1952bp (D) partial
CI303-09	Côte d'Ivoire (Haut Sassandra)	2009	B	B	B	3544bp (B)partial, 1054bp (B) 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) partial
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. 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 for the new complete genomes.

Isolate name (group)	Number of amino acids (starting nucleotide-ending nucleotide*)						Sequence
	ORF 1	ORF2	ORF3	ORFX	ORF4	ORFY	Size (bp)
ToWobe12-02 (A)	**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)	**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)	**143** (441-869)	**132** (869-1264)	**1834** (1272- 6773)	**113** (2374-2712)	**NC**	**131** (6434-6826)	**7161**
CI152-09 (D)	**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) MF642734	**138** (295-708)	**124** (699-1070)	**1833** (1070-6568)	**NC**	**NC**	**132** (6199-6594)	**6985**
Gha54-15 (R) MF642735	**138** (429-842)	**124** (833-1204)	**1833** (1204-6702)	**NC**	**NC**	**132** (6333-6728)	**6990**
Cacao Sri Lanka Virus MF642736	**143** (184-612)	**131** (612-1004)	**1772** (1004-6319)	**NC**	**NC**	**133** (5995-6393)	**7215**

* Without including the stop codon

NC : No Corresponding

Table 4. 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 ToAgou 1-93	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