DOI: 10.5281/zenodo.10210009

Identification and Distribution of Novel Badnaviral Sequences Integrated in the Genome of Cacao (*Theobroma cacao*)

E. Muller[1], I. Ullah[2], J. Dunwell[2], D. Lopez[1], C. Mariac[3], A. Daymond[2], J. Allainguillaume[4], A. Wetten[4]

[1] CIRAD, AGAP Institut, Montpellier, France
AGAP, Univ Montpellier, CIRAD, INRAE, Institut Agro, Montpellier, France
[2] School of Agriculture, Policy and Development, University of Reading, Earley Gate, Reading, United Kingdom
[3] IRD, UMR DIADE, 911 Avenue Agropolis Montpellier, France
[4] University of the West of England, Frenchay Campus, Coldharbour Lane, Bristol, United Kingdom

ABSTRACT

As part of an ongoing study to understand the diversity of the badnavirus complex, responsible for the cacao swollen shoot disease in West Africa, evidence was found recently of virus-like sequences in asymptomatic cacao plants. The present study exploited the wealth of genomic resources in this crop, and combined bioinformatic, molecular, and genetic approaches to report for the first time the presence of integrated badnaviral sequences in most of the cacao genetic groups. These sequences, which we propose to name eTcBV for endogenous *Theobroma cacao* bacilliform viruses, varied in type with each predominating in a specific cacao genetic group. Additionally to the viral insert of type VI first identified, we recently described, with the help of Oxford Nanopore technology, a viral insert of type I and a viral insert of type III. A diagnostic multiplex PCR method was developed to identify the homozygous or hemizygous condition of the specific insert of type VI, which was inherited as a single Mendelian trait.

These data suggest that these integration events occurred before or during the species diversification in Central and South America, and prior to its cultivation in other regions. Such evidence of integrated sequences is relevant to the management of cacao quarantine facilities, and may also aid novel methods to reduce the impact of such viruses in this crop.

1. Introduction

Viral integrations within eukaryotic genomes are frequent events described in a wide range of plant families. Integrated viruses can take the form of partial and/or highly rearranged non-activatable sequences (dormant) or in rare cases, complete and activatable sequences that can cause systemic infection of the host plant.

As part of an ongoing study to understand the diversity of the badnavirus complex, responsible for the cacao swollen shoot disease in West Africa, evidence was found of virus-like sequences in asymptomatic cacao plants. The present study exploited the wealth of genomic resources in this crop, and combined bioinformatic, molecular, and genetic approaches to report for the first time the presence of integrated badnaviral sequences in most of the cacao genetic groups (**Figure 1**).

These sequences, which we propose to name eTcBV for endogenous Theobroma cacao bacilliform viruses, varied in type with each predominating in a specific cacao genetic group. The partial sequences, divided into 13 types (I to XIII), belong to two species called eTCBV1 and eTCBV2 for endogenous *Theobroma cacao* bacilliform virus 1 and 2 (**Figure 2**).

DOI: 10.5281/zenodo.10210009

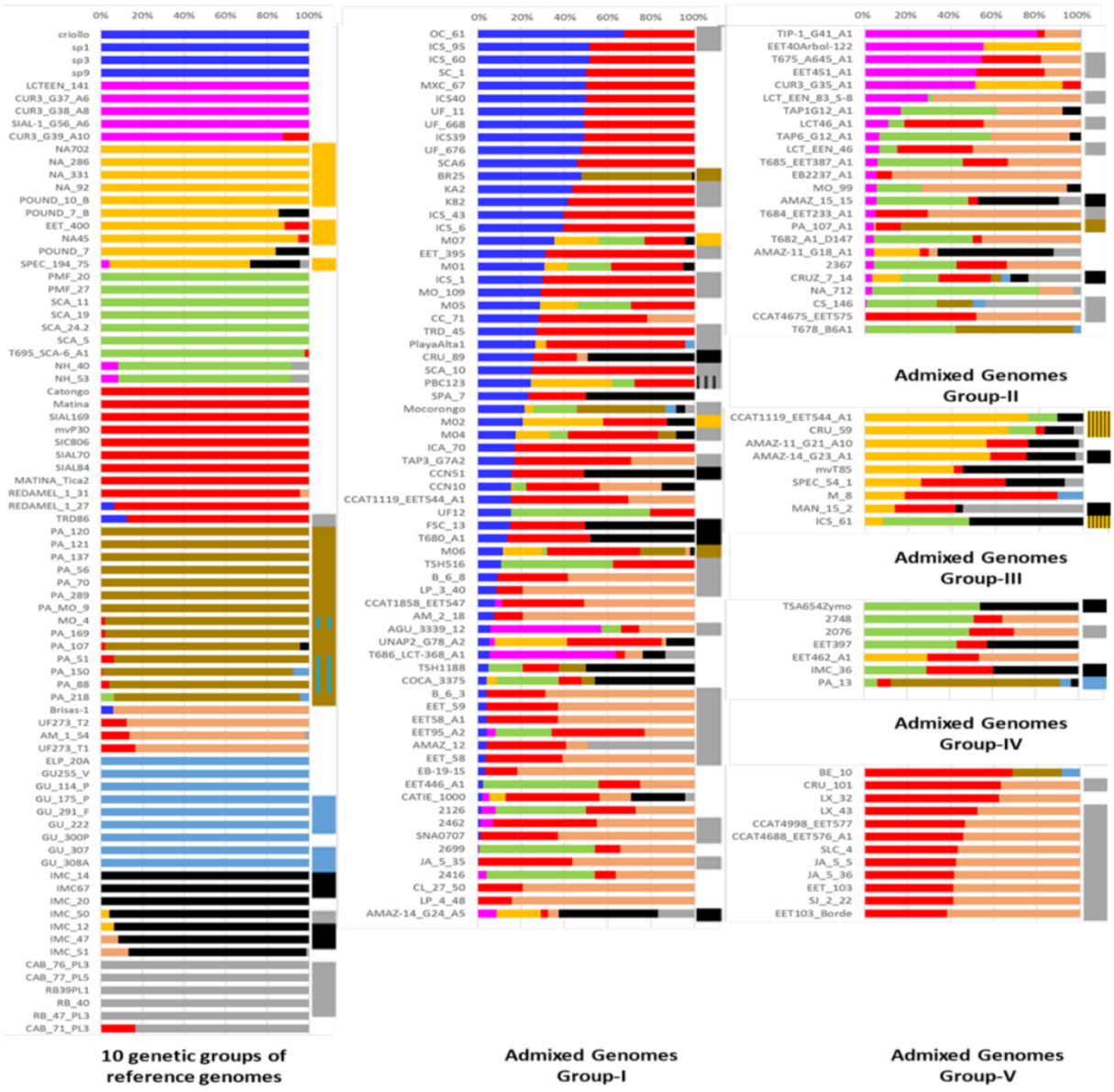

Figure 1: Summary of the bioinformatic search for the presence of the different types of viral sequences in cacao genomes. Bar graph representing population structure in cacao, redrawn from [Cornejo, O. E. et al. Population genomic analyses of the chocolate tree, Theobroma cacao L, provide insights into its domestication process. Commun. Biol. 1, 167 (2018)]. The first block on the left side represents 79 clones (with names) grouped into 10 distinct genetic groups. The other two blocks represent 121 admixed clones grouped into five sub-groups (horizontal bars). The narrow column on the right side of the individual horizontal bars represents the type of viral sequence (I to VI, see Figure 2) found in the clones positive for the viral mapping. The colours correspond to the following groups/virus types: Blue (Criollo), Magenta (Curaray), Golden (Nanay/III), Green (Contamana), Red (Amelonado), Dark brown (Marañon/II), Light brown (Nacional), Light blue (Guiana/I), Black (Iquitos/V) and Grey (Purús/VI). Boxes with hatching patterns represent two viral sequences.

DOI: 10.5281/zenodo.10210009

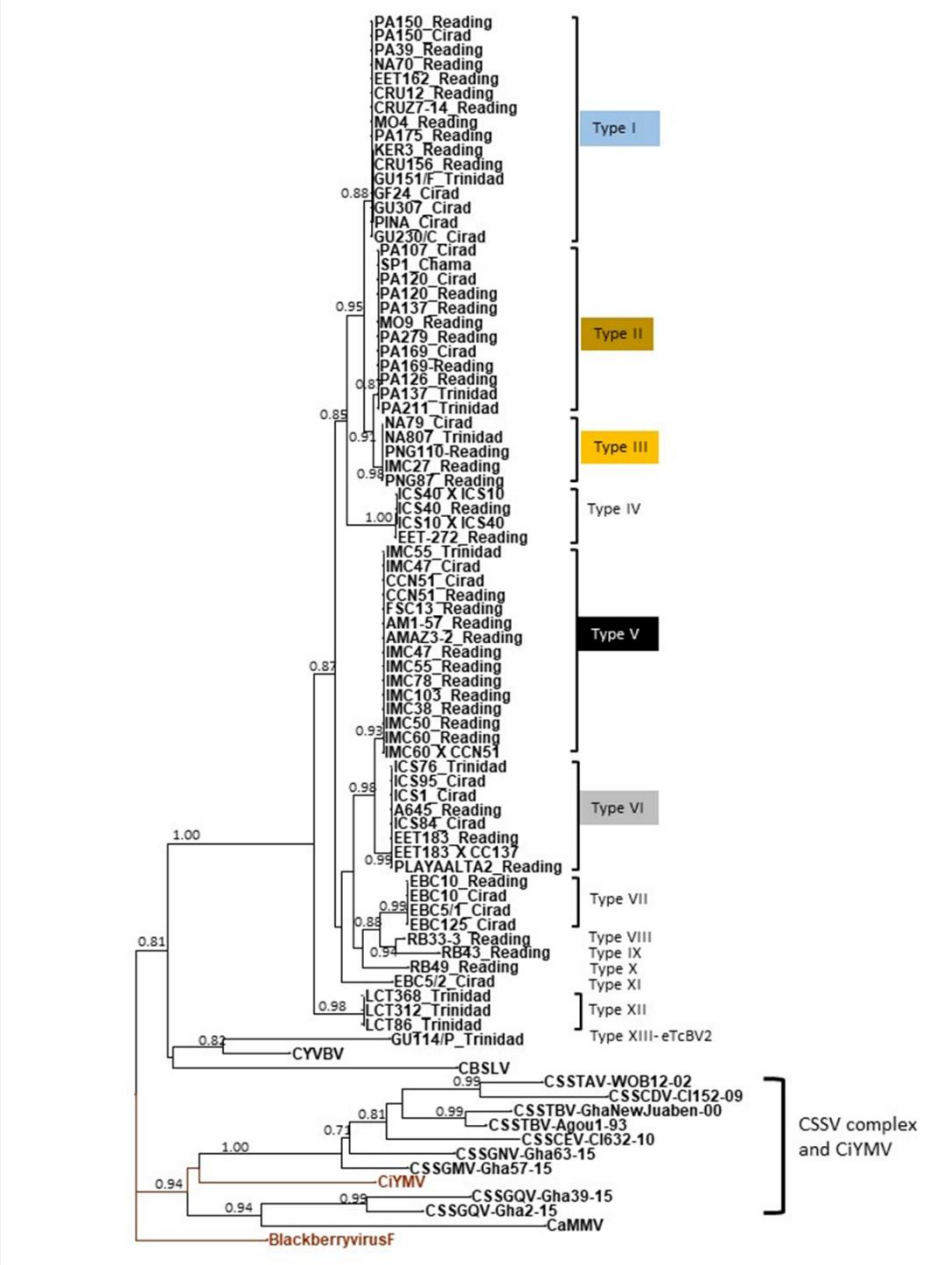

Figure 2: Maximum likelihood phylogenetic tree of badnavirus S sequences based on alignment of the RT RNase H region of open reading frame 3 (ORF3). Numbers on the branches represent the SH-aLRT (approximate Likelihood ratio test) branch supports over 0.7. The Citrus yellow mosaic Virus (CiYMV) (AF347695) and Blackberry virus F (YP009229919) in red colour are used as outgroups along with the other badnavirus infecting cacao trees [CYVBV, CaMMV, CBSLV and species from the Cacao swollen shoot complex (CSSTAV, CSSTBV, CSSCDV, CSSCEV, CSSGMV, CSSGNV, CSSGQV). The names of sequences include the name of the cacao clone and the name of the collection from which they were obtained. The 12 different viral types of sequence are identified from I to XII (eTcBV1) and XIII (eTCBV2).

DOI: 10.5281/zenodo.10210009

Results

The integration of the type VI sequences was demonstrated by bioinformatics analyses on complete cacao genomes, followed by molecular biological confirmation. Additionally, we recently described, with the help of Oxford Nanopore technology, viral inserts of type I, II, III and V longer than 10kb (**Figure 3**). A diagnostic multiplex PCR method was developed to identify the homozygous or hemizygous condition of the specific insert of type VI, which was inherited as a single Mendelian trait.

These data suggest that these integration events occurred before or during the species diversification in Central and South America, and prior to its cultivation in other regions.

Perspectives

Such evidence of integrated sequences is relevant to the management of cacao quarantine facilities. These results may also aid novel methods to reduce the impact of such viruses in this crop. In particular, this could be highly significant for cocoa because as already observed for other viruses, the integration of these two badnaviral species into the cacao genome could modify their resistance to certain viral species of the swollen shoot associated species complex (CSSVs) via processes that may involve interfering RNAs (siRNAs, responsible for gene silencing, which constitutes one of the plant's antiviral defences).

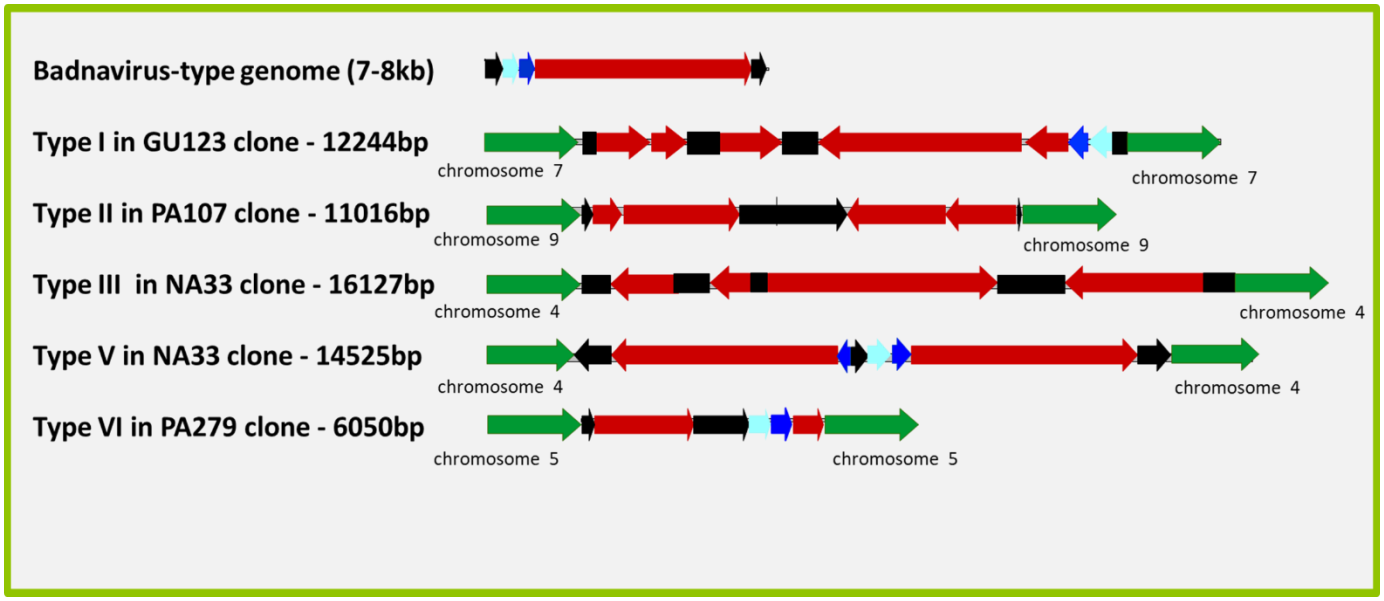

Figure 3: Overview of eTcBV1 structures of type I, II, III, V and VI in cacao genomes. Cacao genomic sequences are in green, with chromosome number. The viral genome is represented in linear view with light blue, dark blue and red solid arrows indicating ORF1, ORF2 and ORF3 of the badnaviral genome, respectively. The intergenic regions are in black.