Characterization of the Complete Genome and P0 Protein for a Previously Unreported Genotype of Cotton Leafroll Dwarf Virus, an Introduced Polerovirus in the United States

Sofia Avelar,¹ Roberto Ramos-Sobrinho,¹ Kassie Conner,² Robert L. Nichols,³ Kathy Lawrence,⁴ and Judith K. Brown^{1,†}

¹ School of Plant Sciences, University of Arizona, Tucson, AZ 85721

² Plant Diagnostic Laboratory, Alabama Cooperative Extension System, Auburn University, Auburn, AL 36849

³ Cotton Incorporated, Cary, NC 27513

⁴ Department of Entomology and Plant Pathology, Auburn University, Auburn, AL 36849

Abstract

Virus-like disease symptoms consisting of leaf cupping, shortened internodes, and overall stunting were observed in commercial cotton fields in Alabama in 2017 to 2018. To determine the complete genome sequence of the suspected causal polerovirus, symptomatic leaf samples were collected in Macon County, Alabama, and subjected to Illumina RNA sequencing. Based on BLASTn analysis, the Illumina contig of 5,771 nt shared the highest nucleotide identity (approximately 95%) with members of the species *Cotton leafroll dwarf virus* (CLRVDV) (genus *Polerovirus*; family *Luteoviridae*) from Argentina and Brazil. The full-length viral genome sequence was verified by reverse transcription (RT)-PCR amplification, cloning, and Sanger sequencing. The complete CLRVDV genome of 5,865 nt in length shared 94.8 to 95.2% nucleotide identity with six previously reported CLRVDV isolates. The genome of the CLRVDV isolate amplified from Alabama samples (CLRVDV-AL) has seven predicted open reading frames (ORFs). Viral proteins 1 to 5 (P1

to P5) shared 91.9 to 99.5% amino acid identity with the six CLRVDV isolates from Argentina and Brazil. However, P0, the suppressor of host gene silencing, shared 82.4 to 88.5% pairwise amino acid identity with the latter CLRVDV isolates. Phylogenetic analysis of the seven full-length CLRVDV genomes resolved three sister clades: CLRVDV-AL, CLRVDV-typical, and CLRVDV-atypical, respectively. Three recombination events were detected by the recombination detection program among the seven CLRVDV isolates with breakpoints occurring along the genome. Pairwise nucleotide identity comparisons of ORF0 sequences for the three CLRVDV-AL field isolates indicated that they were >99% identical, suggesting that this previously unknown CLRVDV genotype represents a single introduction to Alabama.

Keywords: aphid vector, cotton blue disease, emerging plant virus, ssRNA plant virus

Cotton is among the most important fiber crops worldwide, and the United States is one of the largest cotton producers and a leading exporter. In Alabama, upland cotton *Gossypium hirsutum* (L.) is valued at approximately \$335 million (U.S. Department of Agriculture 2018). During the 2017 and 2018 cotton-growing seasons, viral disease symptoms were observed in commercial cotton fields in Alabama, Georgia, and Mississippi (Aboughanem-Sabanadzovic et al. 2019; Avelar et al. 2019; Tabassum et al. 2019). Symptoms consisted of curling and downward cupping of leaves exhibiting a bluish-green and/or a rust to red color, petiole and vein reddening, shortened internodes, dwarfing of plants, and reduced boll set. Among the cotton cultivars and genotypes affected were Delta Pine 1558 and 1646, M-315, and Pima S7. Yield loss was estimated in 2017 at \$19 million for the Alabama cotton crop alone (Avelar et al. 2019). Preliminary Illumina RNA sequencing (RNA-seq) of total RNA isolated from symptomatic cotton plants in Alabama resulted in a viral partial sequence that was most closely related to members of the species *Cotton leafroll dwarf virus* (CLRVDV) (genus *Polerovirus*; family *Luteoviridae*) reported from Argentina and

Brazil (Avelar et al. 2019). The presence of CLRVDV-like genome sequences was later confirmed in cotton plants from Georgia and Mississippi exhibiting similar symptoms (Aboughanem-Sabanadzovic et al. 2019; Tabassum et al. 2019).

A disease referred to as cotton blue disease (CBD) was reported affecting cotton for the first time in the Central African Republic in 1949. Disease symptoms were leaf rolling and vein yellowing, shortened internodes, and stunting of plants. In addition, the leaves developed an intense green to blue color (Cauquil 1977; Cauquil and Vaissayre 1971). The virus was shown to be transmitted by the cotton aphid *Aphis gossypii* (Glover) in a persistent manner (Cauquil and Vaissayre 1971; Mayo and Ziegler-Graff 1996). The causal agent associated with blue disease-like symptoms was identified in 2005, when a partial viral genome sequence was determined from cotton plants in Brazil, and shown to be most closely related to other well-characterized poleroviruses (Corrêa et al. 2005). The virus was named CLRVDV, and the complete genome was sequenced and characterized as a positive-sense, single-stranded RNA polerovirus of 5.8 kb in length (Distéfano et al. 2010).

In Brazil, the disease of cotton associated with CLRVDV has been controlled by cultivating CBD-resistant cotton varieties. However, in 2006, virus-like symptoms were observed in previously CBD-resistant cotton varieties (da Silva et al. 2015). This “new” disease was referred to as atypical cotton blue disease (ACBD) (da Silva et al. 2015) to distinguish it from the typical disease symptoms observed initially in Argentina and Brazil caused by typical CLRVDV (referred to hereafter as the CLRVDV-typical genotype). In 2009, an outbreak occurred in cotton in Argentina, where blue disease had been similarly managed. The previously CBD-resistant cotton plants exhibited foliar reddening and withering, and leaves were held at an abnormally upright position (Agrofoglio et al. 2017). A comparative sequence analysis of the ACBD-inducing genome (referred to hereafter as the CLRVDV-atypical [CLRVDV-at] genotype) with previously

†Corresponding author: J. K. Brown; jbrown@ag.arizona.edu

Funding: This research was supported by Cotton Incorporated (grant 06-829).

*The e-Xtra logo stands for “electronic extra” and indicates that two supplementary figures and one supplementary table are published online.

The author(s) declare no conflict of interest.

Accepted for publication 2 October 2019.

reported CLRDV-typical isolates indicated that CLRDV-at isolates in Argentina and Brazil were very similar to one another and therefore represented a unique CLRDV genotype.

The poleroviruses encode for seven open reading frames (ORFs) (Domier 2012; Smimova et al. 2015). For CLRDV, the protein 0 (P0) encoded by the ORF0 of the typical and atypical genotypes has been shown to act as a viral suppressor of RNA silencing (VSR), possibly by degrading Argonaute1 (Agrofoglio et al. 2019; Cascardo et al. 2015; Delfosse et al. 2014). The CLRDV P0 proteins harbor an F-box-like motif, LPxx(L/I)x₁₀₋₁₃P, previously identified as essential for VSR activity in polerovirus (Pazhouhandeh et al. 2006; Zhuo et al. 2014). Compared with the CLRDV-typical F-box-like motif, CLRDV-at isolates have a single amino acid substitution, in which the isoleucine (I) at position 72 has been replaced by valine (V) (Agrofoglio et al. 2017). Although CLRDV P0 has been implicated as an avirulence determinant associated with resistance-breaking CLRDV-at genotypes, the I-72→V-72 substitution does not solely explain the range of phenotypic variability associated with all CLRDV-at isolates (Agrofoglio et al. 2019). Thus, the contribution of CLRDV P0 to CBD resistance in cotton requires additional characterization to dissect differential virus–genotype interactions involved in pathogenesis and selection in the face of resistance and resistance breaking in cotton.

The objective of this study was to determine and characterize the full-length genome sequence of a CLRDV isolate from Macon County, Alabama, where a CLRDV partial sequence was reported (Avelar et al. 2019). In addition, the ORF0 sequence was determined and compared at the nucleotide and amino acid levels for field isolates collected from three Alabama counties: Barbour, Elmore, and Macon.

Materials and Methods

Plant samples. Leaves with petioles were collected from three symptomatic cotton plants during the 2018 cotton-growing season: Delta Pine 70 (Barbour), Next Gen 5711 B3XF (Macon), and Delta Pine 1646 (Elmore). The tissue was immediately frozen in liquid nitrogen and shipped the following day on dry ice to the University of Arizona School of Plant Sciences in Tucson, Arizona. The frozen plant samples were stored at –80°C until processing.

Purification of RNA and Illumina RNA-seq. Total RNA was isolated and purified from petioles from three individual symptomatic cotton samples using a modified silica capture RNA isolation method, adapted from Rott and Jelkmann (2001), followed by DNase treatment with the DNA-free DNA Removal Kit (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions. The RNA samples were tested for CLRDV presence by reverse transcription (RT)-PCR with the previously reported primers, CLRDV3675F/Pol3982R (Sharman et al. 2015). One CLRDV-positive sample from Macon County that resulted in the highest quality of RNA based on Bioanalyzer RNA integrity number (>7.0) was submitted to the Arizona State University Genomics Core facility for RNA-seq library preparation and Illumina shotgun sequencing (NextSeq version 2, 2 × 75 high output, paired-end reads, 75 bp). The CLRDV isolate amplified from Alabama samples was named CLRDV-AL.

Adapter sequences were removed, and the reads were trimmed using the BBDuk Adapter/Quality Trimming version 37.64 implemented in Geneious Prime version 2019.0.4 (<https://www.geneious.com/>). The trimmed reads were mapped to the CLRDV reference sequence with which they shared the highest nucleotide identity (GenBank accession number HQ827780) using Bowtie2 (Langmead and Salzberg 2012). Mapped reads were subjected to reference-guided assembly using Geneious Prime version 2019.0.4.

Amplification and Sanger sequencing of the CLRDV-AL genome. To evaluate the integrity of the CLRDV-AL genome assembled from RNA-seq reads, the Illumina consensus sequence was used to guide primer design for RT-PCR amplifications covering the full-length viral genome, with a minimum overlap between amplicons of >150 bp. The 5'- and 3'-untranslated regions were determined by RT-PCR amplification of the viral genome using the forward primer P20F and reverse P4R primer, respectively, spanning a length of 24 and 19 nt (Supplementary Table S1). In addition,

primer pair P0_51F/P0_916R was designed to amplify the complete ORF0. Primers, coordinates, and primer annealing temperature are provided in Supplementary Table S1.

Total RNA from a CLRDV-AL isolate collected in Macon County was obtained as described above. First-strand cDNA synthesis was carried out using the SuperScript IV First-Strand Synthesis System (Invitrogen, Carlsbad, CA), according to the manufacturer's instructions, with random hexamer and virus-specific primers (Primer P4; Supplementary Table S1). Amplification of the different CLRDV-AL genomic regions was performed using 1× JumpStart REDTaq ReadyMix (Sigma-Aldrich, St. Louis, MO), or LongAmp Taq 1× Master Mix (New England Biolabs, Ipswich, MA), 0.2 μM of each primer, 1 μl of cDNA, and nuclease-free water, in a 25-μl reaction volume. Cycling conditions were as follows: initial denaturation at 94°C for 2 min; 40 cycles of denaturation at 94°C for 30 s, annealing (see Supplementary Table S1 for temperatures for different primers) for 30 s, and amplification at 72°C (REDTaq) (or at 65°C [LongAmp]) for 1 min/kb, with a final cycle at 72°C (REDTaq) (or at 65°C [LongAmp]) for 10 min. The expected amplicon sizes were confirmed by electrophoresis in agarose gel (0.8%) containing 1× GelRed (Biotium, Hayward, CA), in 1× Tris-acetate EDTA buffer, pH 8.0.

Amplicons were gel purified using the Illustra GFX PCR and DNA Gel Band Purification kit (GE Healthcare Bio-Sciences, Pittsburgh, PA) according to the manufacturer's instructions. The purified fragments were individually ligated to pGEM T-Easy vector (Promega, Madison, WI) and transformed into competent *Escherichia coli* DH5α cells. Plasmid DNA was extracted using the GeneJET Plasmid Miniprep Kit (Thermo Fisher Scientific, Waltham, MA), and the viral inserts were bidirectionally Sanger sequenced by primer walking with an overlap of at least 150 bases. Sequences were assembled into a complete genome using Geneious Prime version 2019.0.4, followed by BLASTn analysis (GenBank) for initial verification.

Genome characterization. The CLRDV-AL ORFs were predicted in Geneious Prime version 2019.1.3 and verified by comparison with previously annotated ORFs from CLRDV reference sequences from GenBank. Predicted frame shifts, unconventional start codons, and ORF read-throughs were considered. The EMBOSS plugin (Rice et al. 2000), implemented in Geneious Prime version 2019.1.3, was used to predict the secondary amino acid structure for the seven P0 proteins encoded by the seven available CLRDV complete genome isolates.

Pairwise distances and phylogenetic analyses. The complete CLRDV-AL genome sequence, and six full-length genomes of CLRDV downloaded from GenBank (accessed in April 2019), were used for delimiting species boundaries. The nucleotide and amino acid sequences for ORFs 0 to 5 were selected from the complete genomes. In addition, partial ORF3 (approximately 250 nt) sequences and complete ORF3 (~605 nt) sequences available for CLRDV-like isolates identified in Mississippi (accession numbers MK512765 to MK512770; Aboughanem-Sabanadzovic et al. 2019) and Georgia (accession numbers MK290759 and MK290760; Tabassum et al. 2019), respectively, were downloaded from GenBank.

The pairwise nucleotide sequence comparisons for the full-length genome and ORFs 0 to 5 sequences were carried out using the Sequence Demarcation Tool (SDT) version 1.2 (Muhire et al. 2014). The amino acid sequences (P0 to P5) were aligned separately using MUSCLE (Edgar 2004), and the pairwise amino acid identities were calculated using SDT version 1.2 (Muhire et al. 2014).

The full-length CLRDV-AL sequence determined herein, and the six complete CLRDV genome sequences available in GenBank, were aligned using MUSCLE (Edgar 2004) implemented in Geneious Prime version 2019.1.3. The optimal nucleotide and amino acid substitution models were determined using MEGAX software (Kumar et al. 2018) and the Bayesian information criterion. The optimal evolutionary model for the genome nucleotide sequences was identified as the Kimura 2-parameter with gamma distribution rates among sites.

The nucleotide and amino acid sequences of CLRDV-AL ORF0 and P0, respectively, were aligned using MUSCLE (Edgar 2004).

The optimal evolutionary models for the nucleotide and amino acid sequences were identified as the Kimura 2-parameter and Jones-Taylor-Thornton, respectively, with uniform substitution rates among sites.

Finally, the nucleotide and amino acid sequences for CLRDV ORF3 and P3, including the U.S. isolates from Mississippi and Georgia, were aligned using MUSCLE (Edgar 2004). The 5′- and 3′-end sequences of the partial ORF3/P3 (250 nt) from the Mississippi isolates were interpreted as missing data. The optimal evolutionary model for the nucleotide and amino acid sequences were the Kimura 2-parameter and Jones-Taylor-Thornton, respectively, with uniform substitution rates among sites.

The maximum likelihood (ML) trees were reconstructed based on the complete genome, ORF0/P0, and ORF3/P3 data sets in MEGAX (Kumar et al. 2018), with 1,000 bootstrap iterations. The phylogenetic trees were edited using FigTree version 1.4.3 (<http://tree.bio.ed.ac.uk/software/figtree/>) and Inkscape (<https://inkscape.org/pt/>). For the full-length genome tree, when the most closely related polerovirus reference sequences were used as outgroups, the effects of homoplasy appeared to influence in-group relationships (i.e., spurious tree topology) (Lyons-Weiler et al. 1998; Smith 1994); thus, the ML trees were midpoint rooted.

Recombination analysis. The CLRDV complete genome sequences were aligned in MUSCLE (Edgar 2004). The aligned sequences were analyzed for evidence of predicted recombination using seven methods available in the recombination detection program (version 4.9; Martin et al. 2015), including RDP, Geneconv, Bootscan, Maximum Chi Square, Chimaera, SisterScan, and 3Seq, using default parameters, except that sequences were considered linear. A *P* value lower than a Bonferroni-corrected cutoff of 0.05 was used to evaluate the statistical significance of the results obtained using the different recombination methods. Recombination events that were detected by at least four different methods were considered valid predictions. Recombination analysis was carried out using a second approach, the neighbor-net method, implemented in Split-Tree4 version 4.14.8 (Huson and Bryant 2006), with the default parameters.

Results

Genome characterization. The CLRDV-AL Illumina RNA-seq reads (43,255) were mapped to the most closely related sequence available in GenBank, referred to as CLRDV-typical (accession number HQ827780) from Brazil and assembled into a consensus sequence of 5,771 nt in length. Alignment of the Illumina-assembled CLRDV-AL genome with six previously published isolates of CLRDV indicated that the terminal 94 nt of the noncoding 3′ end were absent. When the 3′ and 5′ ends of the CLRDV-AL genome were amplified by RT-PCR, cloned, and sequenced, a full-length genome of 5,865 nt in length was obtained accounting the previously missing 94 nt. The CLRDV-AL complete genome sequence of the Macon County genome was deposited in GenBank as accession number MN071395.

CLRDV-AL encodes seven predicted viral ORFs (ORF0, ORF1, ORF1-2, ORF3, ORF3a, ORF4, and ORF3-5; Fig. 1) and has a

genome organization characteristic of previously reported CLRDV isolates (Agrofoglio et al. 2017; da Silva et al. 2015; Distéfano et al. 2010). CLRDV-AL was identified as a single-stranded, positive-sense RNA virus and a member of the genus *Polerovirus* (family *Luteoviridae*). The closest known relatives of the CLRDV-AL genome sequence are six CLRDV isolates previously described from Argentina and Brazil, the only CLRDV genomes available thus far, with which it shared 94.8 to 95.2% nucleotide identity.

Pairwise identity analysis of seven viral coding regions, ORFs 0 to 5, indicated that the predicted ORF0 of CLRDV-AL was the most divergent, at 90.7 to 92.4% nucleotide identity (P0 at 82.8 to 86.6% amino acid identity). ORF1, encoding the predicted 70.1-kDa P1 protein, was the second most divergent region, at 92.9 to 94.2% shared nucleotide identity (P1 at 91.9 to 93.6% amino acid identity). The predicted 118.7-kDa viral RNA-dependent RNA polymerase, a fusion protein of viral proteins P1 and P2 encoded by ORFs 1 and 2 (Domier 2012), shared 94.5 to 95.1% nucleotide identity (P1-P2 at 95.2 to 96.4% amino acid identity). The predicted 22.4-kDa viral coat protein (P3), encoded by ORF3 (Domier 2012), shared 97.0 to 98.0% nucleotide identity (P3 at 98.5 to 99.5% amino acid identity) (Table 1). CLRDV ORF3 (coat protein) partial sequences, previously reported from cotton plants in Mississippi and Georgia (Aboughanem-Sabanadzovic et al. 2019; Tabassum et al. 2019), shared 98.4 to 100% nucleotide identity (P3 at 98.8 to 100% amino acid identity) with the different CLRDV genotypes (*data not shown*). ORF4, which encodes the predicted 19.4-kDa movement (P4) protein, shared 97.0 to 98.5% nucleotide identity (P4 at 92.5 to 96.1% amino acid identity) (Domier 2012). ORF3-5, encoding the predicted 77.2 kDa P3-P5 protein, involved in aphid transmission and virus accumulation, shared 95.3 to 96.0% nucleotide identity (P3-P5 at 96.0 to 97.3% amino acid identity) (Bruyère et al. 1997). Finally, the CLRDV-AL genome contained the non-AUG start codon CUG, used to initiate translation of ORF3a, that encodes the predicted protein 3a (P3a), a putative long-distance movement protein of 5.2 kDa in size (Smirnova et al. 2015). ORF3a shared 97.8 to 99.3% nucleotide identity; however, P3a was identical across genotypes as a result of synonymous mutations (Table 1).

The amino acid sequence of the viral P0 of 28.9 kDa in size, reported to have VSR activity among other poleroviruses and for CLRDV (Agrofoglio et al. 2019; Cascardo et al. 2015; Delfosse et al. 2014; Pazhouhandeh et al. 2006), varied at a number of amino acid residues among CLRDV-at, CLRDV-typical, and CLRDV-AL. Fifteen amino acid substitutions within P0 were uniquely present in the CLRDV-AL isolate, compared with the two previously described CLRDV-at and CLRDV-typical genotypes for which complete genome sequences are available (Fig. 2). The 15 unique amino acid substitutions in P0 of CLRDV-AL were also observed on the sequenced DNA fragment obtained with primers P0_51F/P0_916R (Supplementary Fig. S1). Given these unique amino acid substitutions, the predicted secondary structure for CLRDV-AL P0 differed from that of both CLRDV-at and CLRDV-typical genotypes in South America, which each encode a unique P0. In particular, the replacement of lysine (Lys) with a glutamic acid (Glu) residue at amino acid

Fig. 1. The genome organization of the cotton leafroll dwarf virus-Alabama isolate. The predicted viral open reading frames (ORFs) were identified, and the protein sizes (in kilodaltons) were determined, using Geneious Prime software version 2019.1.3 (<https://www.geneious.com/>). UTR = untranslated region.

residue 107 in CLRDV-AL yielded a unique, predicted secondary structure harboring two additional alpha helices (Fig. 2). In addition, based on amino acid alignment of the F-box-like motif previously identified in P0 of the typical and atypical CLRDV genotypes (Agrofoglio et al. 2017), the CLRDV-AL F-box-like motif had a mutation in which the isoleucine (Ile) was replaced with a valine (Val) at amino acid residue 72. An identical replacement was previously

reported in the F-box-like motif of the CLRDV-at genotype (Agrofoglio et al. 2017; da Silva et al. 2015).

The complete nucleotide sequences for ORF0 from Barbour and Elmore counties were submitted to GenBank under accession numbers MN046205 and MN046206, respectively. A comparison of the P0 amino acid sequence among the three CLRDV-AL isolates, representing Barbour, Elmore, and Macon counties, indicated that they

Table 1. Results of the pairwise nucleotide and amino acid distance analyses of the viral open reading frames (ORFs) for cotton leafroll dwarf virus (CLRDV)-Alabama, and the CLRDV-typical and CLRDV-atypical genotypes from Argentina (ARG) and Brazil (BRA) using Sequence Demarcation Tool software version 1.2

Isolate	Sequence	Nucleotide/amino acid percent identity						
		ORF0/P0	ORF1/P1	ORF1-2/P1-P2	ORF3/P3	ORF3a/P3a	ORF4/P4	ORF3-5/P3-P5
CLRDV-typical	GU167940_ARG06	92.1/86.6	94.0/92.7	95.1/95.8	97.5/98.5	99.3/100	98.1/94.8	95.3/96.0
CLRDV-typical	HQ827780_BRA07	92.4/86.2	94.2/93.6	95.2/96.4	97.5/99.5	97.8/100	98.1/94.8	95.4/96.5
CLRDV-atypical	KF906260_BRA11	90.7/82.8	92.9/91.9	94.5/95.3	98.0/99.5	99.3/100	98.5/96.0	95.5/97.1
CLRDV-atypical	KF906261_BRA06	91.1/83.5	93.3/92.1	94.8/95.3	98.0/99.5	99.3/100	98.5/96.0	95.7/97.1
CLRDV-atypical	KF359946_ARG10	91.6/85.8	93.6/93.0	94.8/95.2	97.0/99.0	98.6/100	97.0/92.5	95.9/97.3
CLRDV-atypical	KF359947_ARG10	91.3/84.7	93.8/92.8	94.6/95.3	98.0/99.5	98.6/100	98.1/94.8	96.0/97.1

Fig. 2. The predicted amino acid secondary structure for the open reading frame encoding protein 0 (P0) for the seven available cotton leafroll dwarf virus (CLRDV) isolates. The black circles indicate the amino acid residues unique to CLRDV-Alabama (CLRDV-AL) at residues that are conserved for the CLRDV-typical and CLRDV-atypical genotypes. The pink structure represents the predicted alpha helices, whereas the blue symbols represent turns, the yellow arrows indicate beta sheets, and the gray symbols indicate the coil-coil secondary structure. The P0 F-box-like motif is indicated by a gray rectangle, and the red circles indicate the conserved amino acid residues within the motif, according to Agrofoglio et al. (2019). The black asterisk indicates amino acid residue 72, where the isoleucine residue in CLRDV-typical P0 has been substituted for a valine in the P0 of both the CLRDV-atypical genotype and CLRDV-AL. USA = United States, ARG = Argentina, and BRA = Brazil.

were identical except at residues 153 and 247, at which amino acid substitutions occurred in Elmore and Macon field isolates, respectively (Supplementary Fig. S1). This observation, taken together with the similar symptoms observed in cotton in Alabama during 2017 to 2018, suggests that the isolates of CLRDV in all three Alabama counties considered here probably represent the same genotype, a scenario that would be expected under a “single introduction hypothesis.”

Phylogenetic relationships. The ML phylogenetic tree reconstructed for the seven CLRDV complete genome sequences resolved three clades consisting of CLRDV-at, CLRDV-typical, and CLRDV-AL genotypes, with CLRDV-AL being basally positioned on the tree (Fig. 3A). However, based on the ORF0 nucleotide and P0 amino acid ML trees, the CLRDV-AL isolate and CLRDV-typical genotype were more closely related to each other than to the CLRDV-at genotype, an observation that was consistent with the results of the pairwise distance analysis (Table 1). In addition, the topologies of the CLRDV P0 amino acid and nucleotide trees were congruent (Fig. 3B and C).

Phylogenetic analysis of the ORF3 sequences for the U.S. (Alabama, Georgia, and Mississippi) and South American isolates indicated that the sequences reported in the United States grouped in

the same clade, separated from the South American viruses. However, bootstrap values for the major clades within the phylogeny were relatively low (<63%), and short branches overall are indicative of a close genetic relationship among all isolates (Fig. 3D).

Recombination analysis. Three unique recombination events were detected among CLRDV isolates. Recombination breakpoints were found in ORF0 (event 2; Table 2), for which CLRDV-AL was identified as the putative recombinant sequence, having CLRDV-typical as the putative minor parent and an unknown putative major parent. Inter-CLRDV genotype recombination was detected between the CLRDV-typical and CLRDV-at genotypes (events 1 and 3; Table 2), but none of them with recombination breakpoints located in ORF0. Although the intergenotype genetic divergence among CLRDV-at and CLRDV-typical (and CLRDV-AL) predicted intraspecies recombination, signals showed low statistical support (Table 2). The neighbor-net analysis, a second approach used for predicting recombination, was consistent with the results of RDP analysis by showing weak evidence of intergenotype recombination among the available CLRDV isolates (Supplementary Fig. S2). These results may be explained by evolutionary

Fig. 3. Phylogenetic relationships for cotton leafroll dwarf virus-Alabama (CLRDV-AL) and the six previously reported CLRDV isolates from Argentina (ARG), Brazil (BRA), and the United States (USA), based on maximum likelihood analysis (1,000 bootstrap iterations) of **A**, the full-length genome sequences, **B**, the nucleotide sequence of open reading frame 1 (ORF0), **C**, the amino acid sequences of the ORF0-encoded protein zero, and **D**, the nucleotide sequences of ORF3, including South American isolates and sequences from Georgia (GA) and Mississippi (MS).

mechanisms other than recombination primarily driving the CLRDV genome diversification. The long branches characteristic of all three sister clades, especially CLRDV-AL, were interpreted as evidence that the accumulation of genomic mutations could be an important force shaping the evolution and diversification of these seven known CLRDV genotypes.

Discussion

CLRDV has become an important constraint on cotton production where outbreaks have occurred in Africa, Asia, South America, and most recently the southern United States and represents a great risk to unaffected cotton-producing areas worldwide. Despite the increasing economic importance, relatively little information is available about the extent and geographic distribution of CLRDV genomic diversity, and until this report, only six genome sequences have been determined.

CLRDV, an aphid-transmitted polerovirus of cotton, exotic to the United States, has posed a new threat to the southern cotton-producing U.S. states since its emergence in Alabama during 2017 (Avelar et al. 2019) and subsequently in Georgia and Mississippi during 2017 to 2018 (Aboughanem-Sabanadzovic et al. 2019; Tabassum et al. 2019). In this study, the complete genome sequence was determined for a previously unknown genotype of CLRDV found widespread in cotton fields in Alabama during 2017 to 2018. Additional information is needed to determine the origin and route of spread of the viral isolates persistent in South America and the recent putative introduced virus in Alabama (CLRDV-AL). Of particular interest is a comparison of viral genomic variability and viral-encoded protein sequences that may provide clues about biological differences or similarities among isolates/genotypes occurring in the Austral-Pacific region and on the Asian continent versus those extant in the United States, the newest site of invasion in North America. To accomplish this, the complete genome sequences must be determined for additional isolates in North America, for representatives from Asia/surrounding areas, and for isolates that may exist in West Africa, the suspected origin of CLRDV (Cauquil and Vaissayre 1971).

The working criterion for the demarcation of polerovirus species based on sequence data for members of the family *Luteoviridae* utilized a >10% divergence at the amino acid level in any viral protein as the species cutoff (Domier 2012). However, the taxonomy for the *Luteoviridae* is presently under revision (International Committee on Taxonomy of Viruses 2018) (to note, the 2018 release was ratified in February 2019 under MSL no. 34; Dr. A. Miller, *Luteoviridae* Study Group Chair, *personal communication*). Specifically, for CLRDV, it has been proposed that isolates sharing <90% amino acid identity for the P0 coding region, but ≥90% amino acid identity in the remaining proteins, represent different genotypes of a single viral species (Agrofoglio et al. 2017; da Silva et al. 2015). Although the P0 amino acid sequence of CLRDV-AL is approximately 13% divergent from the other six known CLRDV isolates (South America) at the amino acid sequence level, viral P1 through P5 do not exceed the 10% pairwise distance cutoff. Based on this criterion, the CLRDV-AL isolate is a previously undescribed genotype of CLRDV. In addition, the majority of intraspecific genomic variability occurs in the CLRDV

ORF0 locus, an observation that is consistent with previous findings for well-studied poleroviruses (Delfosse et al. 2014; Pazhouhandeh et al. 2006).

An F-box-like motif located in P0 was previously shown to be associated with VSR activity among poleroviruses, and amino acid substitutions have been found to inhibit the suppressor activity (Pazhouhandeh et al. 2006). Based on the alignment of the CLRDV F-box-like motifs ($n = 7$), CLRDV-AL harbored a single amino acid substitution for which the Ile was replaced by a Val at amino acid residue 72. The amino acid substitution is identical to that previously identified for P0 encoded by the CLRDV-at genotype (Agrofoglio et al. 2017). However, functional analysis for which the valine and isoleucine residues were exchanged between the CLRDV-typical and CLRDV-at genotypes showed that this substitution alone was not consistent with differences in resultant symptom phenotypes observed among CLRDV isolates, suggesting that additional mutations or perhaps amino acid substitutions leading to modifications in the P0 secondary structure may be responsible for distinct CBD symptoms associated with the CLRDV genotypes (Agrofoglio et al. 2019).

Among the P0 proteins for the three CLRDV genotypes, the amino acid substitutions were dispersed throughout and indicated the potential for imparting changes to the predicted P0 secondary structure. Most striking were two predicted alpha helices resulting from the Lys→Glu substitution at amino acid residue 107 in the P0 of CLRDV-AL. A comparison of the P0 sequences for the CLRDV-AL isolates collected from three counties (Barbour, Elmore, and Macon, respectively) showed that they shared >99% nucleotide identity and 99% amino acid similarity, indicating that the three CLRDV-AL isolates were representatives of one virus genotype. This evidence, in turn, points to the likelihood that all isolates in Alabama have resulted from the same single introduction. Sequencing of additional genomes (and ORF0) sequences for isolates found in other cotton-growing areas in the United States is required to substantiate this hypothesis for other CLRDV isolates discovered in the Gulf Coast states. Although the origin of the CLRDV-AL isolate is not known, at least one predicted recombination event was identified in the CLRDV-AL genome. The breakpoints for this predicted event were located within ORF0. The CLRDV-typical genotype (GU167940) was implicated as the putative minor parent, which could at least in part explain the close phylogenetic relationships apparent between ORF0/P0 of CLRDV-typical and CLRDV-AL. Notably, the CLRDV-AL P0 F-box-like motif harbors the Ile→Val substitution at amino acid residue 72, like that previously reported for CLRDV-at genotypes from Brazil and Argentina (Agrofoglio et al. 2019; da Silva et al. 2015). These patterns suggest that the ORF0 features of CLRDV-AL may have been contributed by a putative CLRDV-at-like parent, whereas the remainder of the CLRDV-AL genome, which is highly reminiscent of extant CLRDV-typical genomes, could have been contributed by a putative CLRDV-typical parent. Although the recombination analyses did not predict statistically significant events, the availability of additional CLRDV genome sequences may increase signal detection. Nonetheless, whether the CLRDV-AL P0 evolved through an accumulation of

Table 2. Predicted recombination events detected for cotton leafroll dwarf virus-Alabama (CLRDV-AL) and CLRDV genome sequences from Argentina (ARG) and Brazil (BRA)

Event ^a	Breakpoint ^b		Recombinant	Parents		Methods ^c	P value ^d
	Begin	End		Minor	Major		
1	3,279 (ORFs 1-2)	5,865 (?)	KF906260_BRA11 ^e KF906261_BRA06	GU167940_ARG06	KF359946_ARG10	R, G, B, M, C, S, 3	4.61E-10
2	65 (?)	188 (ORF0)	Macon_AL_USA18 ^e	GU167940_ARG06	Unknown	R, B, M, S	4.95E-19
3	3,482 (?)	3,900 (ORFs 3-5)	KF359947_ARG10 ^e	GU167940_ARG06	KF359946_ARG10	B, M, C, 3	3.48E-04

^a It is possible that this apparent recombination signal could have been caused by an evolutionary process other than recombination.

^b Numbering starts at the 5' end of the minus-strand primer-binding site and increases clockwise. A question mark indicates that a breakpoint could not be precisely pinpointed. ORF = open reading frame.

^c R = RDP, G = GeneConv, B = Bootscan, M = MaxChi, C = Chimaera, S = SisScan, and 3 = 3SEQ.

^d The reported P values are for the methods indicated in bold, and they are the lowest P values calculated for the region in question.

^e The recombinant sequence may have been misidentified (one of the identified parents might be the recombinant).

convergent mutations, primarily through recombination, or both mechanisms cannot yet be determined. Elucidating the evolutionary interactions among and between ancestral and extant CLRDV genomes will require additional sequencing of representative isolates from locales where blue disease/CLRDV is known to occur, including in Africa, the putative origin of CLRDV.

Previous efforts to breed cotton for CBD resistance in South America have relied on germplasm lines available in Africa where CBD was first reported (Fang et al. 2010). This was necessary because cotton varieties developed in Argentina and Brazil were susceptible to the CLRDV-typical genotype. None of the cotton varieties available for production in the U.S. Gulf Coast states and exposed to infection in Alabama during 2017 to 2018 have shown resistance to CLRDV-AL (K. Lawrence and R. L. Nichols, *personal observations*). To develop resistance in varieties adapted to the various U.S. cotton-growing regions, identification and introgression of genetic resistance to CLRDV will be the best option for virus management.

Because the P0 sequence of CLRDV-AL is >10% divergent from previously reported CLRDV P0 sequences (namely, CLRDV-at and CLRDV-typical genotypes), whether the disease can be managed by the genetic resistance available in the CBD-tolerant varieties developed in South America remains to be determined. Whether CLRDV-AL has a resistance-breaking capacity similar to the CLRDV-at isolates is also not known. Thus, the relationship between the virulence of CLRDV-AL, specifically with respect to its P0 VSR activity, in U.S. varieties compared with CBD-resistant cotton developed in Argentina and Brazil requires immediate attention.

P0 of CLRDV-AL is highly divergent from P0 of typical and atypical CLRDV isolates, making CLRDV-AL a new, previously undescribed genotype of CLRDV. In light of the P0-VSR activity and its recognition as an “avirulence determinant” in CBD resistance breaking (Agrofoglio et al. 2019), breeding programs developing CLRDV-resistant germplasm will now need to consider at least the three known CLRDV genotypes, their associated biological and genetic attributes, and the predicted P0 structure for known and emergent CLRDV genotypes during screening and selection of cotton germplasm for resistance.

Acknowledgment

The authors thank Professor Tatsuya Nagata for providing primer sequences used for the CLRDV genome amplification.

Literature Cited

- Aboughanem-Sabanadzovic, N., Allen, T. W., Wilkerson, T. H., Conner, K. N., Sikora, E. J., Nichols, R. L., and Sabanadzovic, S. 2019. First report of cotton leafroll dwarf virus in upland cotton (*Gossypium hirsutum*) in Mississippi. *Plant Dis.* 103:1798.
- Agrofoglio, Y. C., Delfosse, V. C., Casse, M. F., Hopp, H. E., Bonacic Kresic, I., Ziegler-Graff, V., and Distéfano, A. J. 2019. P0 protein of cotton leafroll dwarf virus-atypical isolate is a weak RNA silencing suppressor and the avirulence determinant that breaks the cotton *Cbd* gene-based resistance. *Plant Pathol.* 68:1059-1071.
- Agrofoglio, Y. C., Delfosse, V. C., Casse, M. F., Hopp, H. E., Kresic, I. B., and Distéfano, A. J. 2017. Identification of a new cotton disease by an atypical cotton leafroll dwarf virus in Argentina. *Phytopathology* 107:369-376.
- Avelar, S., Schrimsher, D. W., Lawrence, K., and Brown, J. K. 2019. First report of cotton leafroll dwarf virus associated with cotton blue disease symptoms in Alabama. *Plant Dis.* 103:592.
- Bruyère, A., Brault, V., Ziegler-Graff, V., Simonis, M.-T., Van den Heuvel, J. F. J. M., Richards, K., Guilley, H., Jonard, G., and Herrbach, E. 1997. Effects of mutations in the beet western yellows virus readthrough protein on its expression and packaging and on virus accumulation, symptoms, and aphid transmission. *Virology* 230:323-334.
- Cascardo, R. S., Arantes, I. L. G., Silva, T. F., Sachetto-Martins, G., Vaslin, M. F. S., and Corrêa, R. L. 2015. Function and diversity of P0 proteins among cotton leafroll dwarf virus isolates. *Virology* 12:123.
- Cauquil, J. 1977. Etudes sur une maladie d'origine virale du cotonnier: La maladie bleue. *Cot Fib Trop.* 32:259-278.
- Cauquil, J., and Vaissayre, M. 1971. La “maladie bleue” du cotonnier en Afrique: Transmission de cotonnier à cotonnier par *Aphis gossypii* glover. *Cot Fib Trop.* 26:463-466.
- Corrêa, R. L., Silva, T. F., Simões-Araújo, J. L., Barroso, P. A. V., Vidal, M. S., and Vaslin, M. F. S. 2005. Molecular characterization of a virus from the family *Luteoviridae* associated with cotton blue disease. *Arch. Virol.* 150:1357-1367.
- da Silva, A. K. F., Romanel, E., Silva, T. da F., Castilhos, Y., Schrago, C. G., Galbieri, R., Be'lot, J., and Vaslin, M. F. S. 2015. Complete genome sequences of two new virus isolates associated with cotton blue disease resistance breaking in Brazil. *Arch. Virol.* 160:1371-1374.
- Delfosse, V. C., Agrofoglio, Y. C., Casse, M. F., Kresic, I. B., Hopp, H. E., Ziegler-Graff, V., and Distéfano, A. J. 2014. The P0 protein encoded by cotton leafroll dwarf virus (CLRDV) inhibits local but not systemic RNA silencing. *Virus Res.* 180:70-75.
- Distéfano, A. J., Bonacic Kresic, I., and Hopp, H. E. 2010. The complete genome sequence of a virus associated with cotton blue disease, cotton leafroll dwarf virus, confirms that it is a new member of the genus *Polerovirus*. *Arch. Virol.* 155:1849-1854.
- Domier, L. L. 2012. Family *Luteoviridae*. Pages 1045-1053 in: *Virus Taxonomy, Ninth Report of the International Committee on Taxonomy of Viruses*. A. M. Q. King, M. J. Adams, E. B. Carstens, and E. J. Lefkowitz, eds. Elsevier/Academic Press, London, UK.
- Edgar, R. C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792-1797.
- Fang, D. D., Xiao, J., Canci, P. C., and Cantrell, R. G. 2010. A new SNP haplotype associated with blue disease resistance gene in cotton (*Gossypium hirsutum* L.). *Theor. Appl. Genet.* 120:943-953.
- Huson, D. H., and Bryant, D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23:254-267.
- International Committee on Taxonomy of Viruses. 2018. Virus taxonomy: 2018b release EC 50. <https://talk.ictvonline.org/taxonomy/>
- Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. 2018. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35:1547-1549.
- Langmead, B., and Salzberg, S. L. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9:357-359.
- Lyons-Weiler, J., Hoelzer, G. A., and Tausch, R. J. 1998. Optimal outgroup analysis. *Biol. J. Linn. Soc.* 64:493-511.
- Martin, D. P., Murrell, B., Golden, M., Khoosal, A., and Muhire, B. 2015. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol.* 1:1-5.
- Mayo, M. A., and Ziegler-Graff, V. 1996. Molecular biology of luteoviruses. *Adv. Virus Res.* 46:413-460.
- Muhire, B. M., Varsani, A., and Martin, D. P. 2014. SDT: A virus classification tool based on pairwise sequence alignment and identity calculation. *PLoS One* 9:e108277.
- Pazhouhandeh, M., Dieterle, M., Marrocco, K., Lechner, E., Berry, B., Brault, V., Hemmer, O., Kretsch, T., Richards, K. E., Genschik, P., and Ziegler-Graff, V. 2006. F-box-like domain in the polerovirus protein P0 is required for silencing suppressor function. *Proc. Natl. Acad. Sci.* 103:1994-1999.
- Rice, P., Longden, I., and Bleasby, A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* 16:276-277.
- Rott, M. E., and Jelkmann, W. 2001. Characterization and detection of several filamentous viruses of cherry: Adaptation of an alternative cloning method (DOP-PCR), and modification of an RNA extraction protocol. *Eur. J. Plant Pathol.* 107:411-420.
- Sharman, M., Lapbanjob, S., Seburuang, P., Belot, J.-L., Galbieri, R., Giband, M., and Suassuna, N. 2015. First report of cotton leafroll dwarf virus in Thailand using a species-specific PCR validated with isolates from Brazil. *Australas. Plant Dis. Notes* 10:24.
- Smirnova, E., Firth, A. E., Miller, W. A., Scheidecker, D., Brault, V., Reinbold, C., Rakotondrafara, A., Chung, B., and Ziegler-Graff, V. 2015. Discovery of a small non-AUG-initiated ORF in poleroviruses and luteoviruses that is required for long-distance movement. *PLoS Pathog* 11:e1004868.
- Smith, A. B. 1994. Rooting molecular trees: Problems and strategies. *Biol. J. Linn. Soc.* 51:279-292.
- Tabassum, A., Bag, S., Roberts, P., Suassuna, N., Chee, P., Whitaker, J. R., Conner, K. N., Brown, J., Nichols, R. L., and Kemeraite, R. C. 2019. First report of cotton leafroll dwarf virus infecting cotton in Georgia, U.S.A. *Plant Dis.* 103:1803.
- U. S. Department of Agriculture, National Agricultural Statistics Service. 2018. State Agriculture Overview. Available from: https://www.nass.usda.gov/Quick_Stats/Ag_Overview/stateOverview.php?state=ALABAMA. Accessed on May, 2019.
- Zhuo, T., Li, Y.-Y., Xiang, H.-Y., Wu, Z.-Y., Wang, X. B., Wang, Y., Zhang, Y. L., Li, D. W., Yu, J. L., and Han, C. G. 2014. Amino acid sequence motifs essential for P0-mediated suppression of RNA silencing in an isolate of potato leafroll virus from inner Mongolia. *Mol. Plant-Microbe Interact.* 27:515-527.