Characterization of the complete genome and ORF0 protein for a previously unreported strain of *Cotton leafroll dwarf virus*, an introduced polerovirus in the USA

Sofia Avelar  
School of Plant Sciences, The University of Arizona  
Tucson, AZ 85721

Roberto Ramos-Sobrinho  
School of Plant Sciences, The University of Arizona  
Tucson, AZ 85721

Kassie Conner  
Alabama Cooperative Extension System  
Plant Diagnostic Lab, Auburn University  
Auburn, AL 36849

Robert L. Nichols  
Cotton Incorporated  
Cary, NC 27513

Kathy Lawrence  
Department of Entomology and Plant Pathology, Auburn University  
Auburn, AL 36849

Judith K. Brown  
School of Plant Sciences, The University of Arizona  
Tucson, AZ 85721

**Corresponding author:**  
J.K. Brown  
Email: jbrown@ag.arizona.edu
ABSTRACT

During 2018, virus-like disease symptoms were observed in upland cotton fields in Alabama (AL) and other cotton-growing states in the southern U.S. To identify the suspect causal virus, symptomatic cotton samples were collected from commercial fields in Macon County, AL and subjected to Illumina RNA sequencing (RNAseq). The BLASTn analysis of the Illumina virus-like contig of 5,771 nucleotides (nt) against sequences available in the GenBank database showed highest nt identity (~95%) with Cotton leafroll dwarf virus (CLRDV) (genus, Polerovirus; family, Luteoviridae) from Argentina (AR) and Brazil (BR).

The almost full-length viral genome sequence, determined by RNAseq, was verified by reverse transcription (RT)-PCR amplification, cloning, and Sanger sequencing. A complete CLRDV genome of 5,865 nt in length was obtained, which shared 94.8-95.2% identity with previously reported CLRDV isolates. The ‘CLRDV-AL’ genome has seven predicted open reading frames (ORFs). Although the viral ORFs 1-5 share 91.9-99.5% amino acid (aa) identity with CLRDV isolates from AR and BR, the ORF0, which encodes a suppressor of host gene-silencing, was divergent and shared only 82.4-88.5% aa identity with the AR and BR isolates. Phylogenetic analysis of seven full-length CLRDV genomes resolved three sister clades representing CLRDV-AL, CLRDV ‘typical’, and CLRDV ‘atypical’, respectively. Three recombination events were detected by RDP among the seven CLRDV isolates with breakpoints occurring along the genome. Pairwise nt identity comparisons of ORF0 nt sequences for the three CLRDV-AL field isolates indicated they were >99% identical, suggesting a single introduction of a previously unidentified CLRDV strain.

S. Avelar
**Keywords**: aphid vector, cotton blue disease, emerging plant virus, ssRNA plant virus

**INTRODUCTION**

Cotton is among the most important fiber crops worldwide, with the United States being the third largest cotton producer and leading cotton exporter. In Alabama, upland cotton *Gossypium hirsutum* (L.) is valued at ~$335 million dollars (USDA, 2018). During 2017 and 2018 cotton-growing season, viral disease symptoms were observed in commercial cotton fields in Alabama, Florida, Georgia, and Mississippi (Aboughanem-Sabanadzovic et al. 2019; Avelar et al., 2019; ACES, 2019; Tabassum et al. 2019). Symptoms consisted in curling and downward cupping of leaves exhibiting a bluish-green color, petiole and vein reddening, shortened internodes and dwarfing of plants, and reduced boll set. Yield loss was estimated 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 AL resulted in a viral partial sequence that was most closely related to *Cotton leafroll dwarf virus* (CLRDV) (genus, *Polerovirus*; family, *Luteoviridae*) isolates from Brazil (BR) and Argentina (AR) (Avelar et al., 2019). The presence of CLRDV-like sequences has also been confirmed in cotton plants from Georgia and Mississippi exhibiting similar symptoms (Tabassum et al. 2019; Aboughanem-Sabanadzovic et al. 2019).

The cotton blue disease (CBD) was first reported in aphid infested cotton fields in the Central African Republic during 1949. Disease symptoms were leaf rolling and vein-yellowing, shortened internodes, and stunting of plants. Also, the leaves developed an
intense green to blue color (Cauquil and Vaissayre 1971; Cauquil 1977). 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 of CBD was not identified until 2005, when a partial genome sequence was determined from cotton plants exhibiting CBD symptoms in BR, which shared its closest sequence homology with other well-characterized poleroviruses (Corrêa et al. 2005). The first CLRDV complete genome sequence was later sequenced and characterized, by Distéfano et al. (2010), as a positive-sense, single stranded RNA polerovirus of 5.8 kilobases (kb) in length.

In BR, the disease has been controlled through the use of 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 CBD’ (ACBD) (da Silva et al. 2015) to distinguish it from the ‘typical’ blue disease symptoms observed initially in AR and BR caused by ‘typical’ CLRDV, hereafter, CLRDV-typical strain. In 2009, an outbreak occurred in cotton in AR, where blue disease had been similarly managed. The previously CBD-resistant cotton plants exhibited foliar reddening and withering, and leaves were held at abnormally upright position (Agrofoglio et al. 2017). A comparative sequence analysis of the ACBD-inducing genome, hereafter, CLRDV-atypical strain, with previously reported CLRDV-typical isolates indicated that CLRDV-atypical isolates in AR and BR were very similar to one another and therefore represented a unique CLRDV strain.

The poleroviruses encode for seven open reading frames (ORFs) (Domier 2012; Smirnova et al. 2015). For CLRDV, the *P0* encoded by the ORF0 of the ‘typical’ and
‘atypical’ strains have been shown to act as viral suppressor of RNA silencing (VSR), possibly by degrading Argonaute1 (AGO1) (Cascardo et al. 2015; Agrofoglio et al. 2019).

The CLRDV P0 proteins harbor a F-box-like motif, LPxx(L/I)x_{10-13}P, previously identified as essential for VSR activity in polerovirus (Pazhouhandeh et al. 2006; Zhuo et al. 2014).

Compared to the ‘typical’ CLRDV F-box-like motif, ‘atypical’ CLRDV isolates have a single aa substitution, in which the isoleucine (I) at position 72 has been replaced by valine (V) (Agrofoglio et al. 2017). Although the CLRDV-P0 has been implicated as an avirulence (Avr) determinant associated with resistance-breaking ‘atypical’ strains of CLRDV, the I-72→V-72 substitution does not solely explain the range of phenotypic variability associated with all atypical CLRDV isolates (Agrofoglio et al. 2019). Thus, the contribution of CLRDV-P0 to CLRDV disease resistance in cotton requires additional characterization to dissect differential viral-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, from where CLRDV partial sequence has been recently reported (Avelar et al. 2019). In addition, the P0 ORF sequence was determined and compared at the nucleotide (nt) and aa levels for field isolates collected from three Alabama counties, Barbour, Elmore, and Macon. The ORF0 coding region of CLRDV is of particular interest due its apparent association with resistance-breaking in cotton and because of the extensive divergence in P0 observed between ‘typical’ and ‘atypical’ isolates from South America and most recently, the CLRDV-AL isolates reported causing outbreaks in the U.S. Results reported here also
suggest that differences in CLRDV ORF0 sequences for isolates characterized so far are taxonomically informative at the level of viral strain.

MATERIALS AND METHODS

Plant samples

Cotton plants exhibiting downward cupping, reddening of veins and petioles, and dwarfing symptoms were collected in Barbour, Elmore, and Macon County, AL, during the 2018 cotton-growing season. The leaf, petiole and stem tissues were immediately frozen in liquid nitrogen and shipped the following day on dry ice to the School of Plant Sciences, The University of Arizona, Tucson, AZ, USA. The frozen plant samples were stored at -80°C until processing.

Purification of RNA and Illumina RNA-sequencing

Total RNA was isolated and purified from symptomatic petioles 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-polymerase chain reaction (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 (RIN, >7.0) was submitted to the ASU
Genomics Core facility for RNA-seq library preparation and Illumina shotgun sequencing (NextSeq v2 2x75 High output, paired-end reads, 75 bp).

Adapter sequences were removed, and the reads were trimmed using the BBBDuk Adapter/Quality Trimming Version 37.64 implemented in Geneious Prime v. 2019.0.4 (http://www.geneious.com). The trimmed reads were mapped to the CLRDV reference sequence with which they shared the highest nt homology (GenBank Accession No. HQ827780) using Bowtie2 (Langmead and Salzberg 2012). Mapped reads were subjected to reference-guided assembly using the Geneious Prime v. 2019.0.4 software, (http://www.geneious.com).

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

Total RNA from selected field samples was obtained as described above. First strand cDNA synthesis was carried out using SuperScript™ IV First-Strand Synthesis System (Invitrogen, Carlsbad, CA, US), according to the manufacturer’s instructions, with random hexamer and virus-specific primers (Primer P4; Table S1). Amplification of the different CLRDV-AL genomic regions was performed using 1X JumpStart REDTaq ReadyMix (Sigma-Aldrich, St. Louis, MO, US), or LongAmp® Taq 1X Master Mix (New
England Biolabs, Ipswich, MA, US), 0.2 μM of each primer, 1 μL of cDNA, and nuclease-free water, in a 25 μL reaction volume. Cycling conditions were: initial denaturation at 94 °C for 2 min; 40 cycles of denaturation at 94 °C for 30 s, annealing (see 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 1X GelRed (Biotium, Hayward, CA, USA), in 1X Tris-acetate EDTA buffer pH 8.0, (TAE).

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

**Genome characterization**

The CLRDV-AL ORFs were predicted in Geneious Prime 2019.1.3 (http://www.geneious.com) 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.)
9

2000), implemented in Geneious Prime 2019.1.3 (http://www.geneious.com), was used to predict the secondary aa structure for the seven P0 proteins associated with the seven available CLRDV complete genome sequence.

203 Pairwise distances and phylogenetic analyses

204 The complete CLRDV-AL genome sequence, and six full-length genomes of CLRDV downloaded from GenBank (accessed on April 2019), were used to confirm species demarcation. Nucleotide (nt) and aa sequences for ORFs 0-5 were obtained from the complete genome data set. Pairwise nt sequence comparisons were performed for the full-length genome and ORFs 0-5 data sets using Sequence Demarcation Tool (SDT) v.1.2 (Muhire et al. 2014). Amino acid sequences (ORFs 0-5) were aligned separately using MUSCLE (Edgar 2004), and pairwise aa identities were calculated in SDT v.1.2 (Muhire et al. 2014).

212 For the phylogenetic analysis, a multiple sequence alignment was made for the seven available CLRDV genome sequences, and for the nt and aa sequences of the CLRDV-AL ORF0 coding region using MUSCLE (Edgar 2004), implemented in Geneious Prime 2019.1.3 (http://www.geneious.com). The most optimal nt and aa substitution models were determined for the aligned sequences using MEGAX (Kumar et al. 2018) and the Bayesian Information Criterion (BIC). The Maximum likelihood (ML) trees were reconstructed in MEGAX (Kumar et al. 2018) with 1000 bootstrap iterations. The trees were edited using FigTree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/) and Inkscape (https://inkscape.org/pt/).
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 recombination detection program (RDP v.4.9, Martin et al. 2015), including RDP, Geneconv, Boot-scan, Maximum Chi Square, Chimaera, SisterScan, and 3Seq, using default parameters, except that sequences were considered linear. A $P$-value lower than a Bonferroni-corrected cut-off 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. Recombination analysis was carried out using a second approach, the Neighbor-Net method, implemented in SplitsTree4 v.4.14.8 (Huson and Bryant 2006), with the default parameters.

RESULTS

CLRDV-AL complete genome characterization

The CLRDV-AL RNA-seq Illumina reads (43,255) were mapped to the most closely related sequence available in GenBank, a ‘typical’ CLRDV isolate (Accession no. HQ827780) from BR and assembled into a consensus sequence of 5,771 nt in length. When aligned with the complete CLRDV sequence (HQ827780) of 5,865 nt, the non-coding 3’-end region of the Illumina-based CLRDV-AL lacked the last 94 nucleotides. To determine the complete genome, the CLRDV-AL sequence was amplified by RT-PCR, cloned, and sequenced by primer walking. An apparently full-length genome of 5,865 nts
in length was obtained, based on the 94 additional nucleotides at the 3'-end. Although
the non-coding 3'-end 94 nt sequence was not represented at a depth in Illumina reads
sufficient to permit its assembly into the complete genome, it was obtained by RT-PCR
amplification, yielding a full-length genome. The CLRDV-AL complete genome for the
Macon County, AL isolate was 5,865 nt in length, and the sequence was deposited in
GenBank as the Accession no. MN071395.

The CLRDV-AL encodes seven predicted viral open reading frames ORFs 0-5 (Fig.
1) and has a genome organization characteristic of previously reported CLRDV isolates
(Agrofoglio et al. 2017; Distéfano et al. 2010; Silva et al. 2015). The CLRDV-AL was
identified as a single stranded, positive sense RNA virus, and member of the genus,
Polerovirus (family, Luteoviridae). The closest known relatives of CLRDV-AL genome
sequence are six CLRDV isolates previously described from AR and BR, the only CLRDV
genomes available so far, with which it shared 94.8 to 95.2% nt identity.

Pairwise identity analysis of six viral coding regions, ORFs 0 and 1-5, indicated
that the predicted ORF0 of CLRDV-AL was the most divergent, at 90.7-92.4% nt identity
(82.8-86.6% aa identity), whereas, the predicted 22.4 kDa viral coat protein (P3), encoded
by ORF3, was the most conserved, at 97.0-98.0% nt identity (98.5-99.5% aa). The ORF1,
encoding the predicted 70.1 kDa P1 protein, was the second most divergent region, at
92.9-94.2% shared nt identity (91.9-95.2% aa). The predicted 118.7 kDa viral RNA-
dependent RNA polymerase (RdRp), a fusion protein of viral proteins P1 and P2, encoded
by ORFs 1 and 2, shared 94.5-95.1% nt identity (95.2-96.4% aa). ORF4, which encodes
the predicted 19.4 kDa movement (P4) protein, shared 97.0-98.5% nt identity (92.5-96.1%
aa). ORF3-5, encoding the predicted 77.2 kDa P3-P5 protein, involved in aphid

S. Avelar
transmission and virus accumulation, shared 95.3-96.0% nt identity (96.0-97.3% aa) (Table 1). Finally, the CLRDV-AL genome contained the non-AUG start codon CUG, used to initiate translation of ORF3a, described for both the ‘atypical’ and ‘typical’ CLRDV isolates that encode the predicted protein 3a (P3a), a putative long-distance movement protein of 5.2 kDa in size (Bruyère et al. 1997; Domier 2012; Smirnova et al. 2015).

The aa sequence of the viral P0, of 28.9kDa in size, reported to have VSR activity among other poleroviruses and for CLRDV (Agrofoglio et al. 2019; Cascardo et al. 2015; Pazhouhandeh et al. 2006), varied at a number of AA residues among the CLRDV-atypical, CLRDV-typical, and CLRDV-AL ORFs. Fifteen AA substitutions within P0 were uniquely present in the CLRDV-AL isolate, compared to the two previously described ‘atypical’ and ‘typical’ CLRDV strains for which complete genome sequences are available (Fig. 2). Given these 15 unique AA substitutions, the predicted secondary structure for CLRDV-AL P0 differed from that of both the ‘atypical’ and ‘typical’ CLRDV strains in South America, which each encode a unique P0. In particular, the replacement of lysine (Lys) with a glutamic acid (Glu) residue at AA residue 107 in CLRDV-AL yielded a unique, predicted secondary structure harboring two additional alpha helices (Fig. 2).

Also, based on aa alignment of the F-box-like motifs previously identified in the P0 ORF of the ‘typical’, ‘atypical ‘ and ‘AL’ CLRDV strains (Pazhouhandeh et al. 2006), the CLRDV-AL F-box-like motif had a mutation in which the isoleucine (Ile) was replaced with a valine (Val) at AA residue 72. An identical replacement has been previously reported in the F-box-like motif of CLRDV-atypical strain (da Silva et al. 2015; Agrofoglio et al. 2017). Based on the latter mutation and the overall P0 aa sequence, CLRDV ‘atypical’ strains and CLRDV-AL shared the greatest homology among the three strains (Fig. 2).
The complete nt sequence for the ORF0 from Barbour and Elmore counties were submitted to GenBank under the accession numbers MN046205 and MN046206, respectively. A comparison of the P0 aa sequence among the three CLRDV-AL isolates, representing Barbour, Elmore, and Macon County, indicated they were identical except at the residues 153 and 247, at which AA substitutions occurred in Elmore and Macon field isolates, respectively (SFig. 1). This observation, taken together with the similar symptoms observed in cotton in Alabama during 2017-2018, suggest that the isolates of CLRDV in all three of the Alabama counties considered here represent the same strain. Further, this scenario would be expected if a single introduction has given rise to the emergence of CLRDV in the U.S.

**Phylogenetic relationships**

The ML phylogenetic tree reconstructed for the seven CLRDV complete genome sequences showed three clades with ‘typical’ and ‘atypical’ CLRDV strains each forming a sister clade to CLRDV-AL, which was is the basal position on the tree (Fig. 3A). However, based on the ORF0 nt and aa ML trees, the CLRDV-AL isolate and CLRDV ‘typical’ strain were more closely related to each other than to CLRDV-atypical strain, an observation that was consistent with the results of the pairwise distance analysis (Table 1). Even so, the topologies of the CLRDV P0 aa and nt trees, respectively, were congruent (Fig. 3B and 3C).

**Recombination analysis**
Three unique recombination events were detected among CLRDV isolates. Recombination breakpoints were found in the ORF0 (event 2; Table 2), for which CLRDV-AL was identified as the recombinant sequence, having CLRDV-typical as putative minor parent, and unknown putative major parent. Inter-CLRDV strain recombination was detected between the CRLDV-'typical' and '-atypical' strains (events 1 and 3; Table 2), but none of them with recombination breakpoints located in the ORF0. Although the inter-strain genetic divergence among the ‘typical’, and ‘atypical’ CLRDV, and CLRDV-AL, predicted intra-species recombination signals show 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 inter-strain recombination among the available CLRDV isolates (SFig. 2). These results may be explained by evolutionary mechanisms other than recombination primarily driving the CLRDV genome diversification. The long branches characteristic of all three sister clades, specially CLRDV-AL, were interpreted as evidence that the accumulation of genomic mutations being an important force shaping the evolution and diversification of these seven known CLRDV genotypes.

**DISCUSSION**

Cotton blue disease is an important constraint on cotton production where outbreaks have occurred in Africa, Asia, South America, and most recently the southern U.S., 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.

The CLRDV, an aphid-transmitted polerovirus of cotton, exotic to the U.S., poses a new threat to the southern cotton-producing states of the U.S., since its emergence in Alabama during 2017 (Avelar et al. 2019), and subsequently, in Florida, Georgia, and Mississippi during 2017-2018 (Aboughanem-Sabanadzovic et al. 2019; Tabassum et al. 2019; ACES, 2019). In this study, the complete genome sequence has been determined for a previously unknown strain of CLRDV found widespread in cotton fields in Alabama during 2017-2018. The virus is thought to have been introduced either on infected plant materials or by viruliferous aphids carried on hurricane winds. However, to determine the origin and route of spread of the isolates persisting in South America and now CLRDV-AL, in the southern U.S. cotton-growing states, additional information is required with respect to the extent of genomic variability among previously unstudied isolates/strains present on three other continents, in addition to the newest site of invasion in North America.

The working criterion for the demarcation of polerovirus species based on sequence data for members of the family, Luteoviridae has utilized a >10% divergence at the aa level in any viral ORF, as the species cutoff (Domier 2012), however, the taxonomy for the Luteoviridae is presently under revision (https://talk.ictvonline.org/taxonomy/ Virus Taxonomy: 2018b Release EC 50, Washington, DC, July 2018; Ratification February 2019 (MSL #34; personal communication, Dr. A. Miller; Chair, Luteoviridae Study Group). Specifically, for CLRDV, it has been proposed that isolates sharing < 90% aa identity for the P0 coding region, but ≥ 90% aa identity to the remaining ORFs, represent different
strains of a single viral species (Agrofoglio et al. 2017; da Silva et al. 2015). Although the P0 aa sequence of CLRDV-AL is ~13% divergent with the other six known CLRDV isolates (South America), at the aa sequence level, viral ORFs 1 through 5 do not exceed the 10% pairwise distance cutoff. Based on this criterion, the CLRDV-AL isolate is a previously undescribed strain of CLRDV, herein named CLRDV-AL. Also, the majority of intraspecific genomic variability occurs in the CLRDV ORF0 locus, an observation that is consistent with previous findings for well-studied poleroviruses (Pazhouhandeh et al. 2006; Delfosse et al. 2014).

An F-box-like motif located in the P0 has been previously shown to be associated with VSR activity among poleroviruses, and aa 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 aa substitution for which the Ile has been replaced by a Val at aa residue 72. The aa substitution is like that previously been identified for the CLRDV-atypical strain ORF0 (Agrofoglio et al. 2017). However, functional analysis for which the valine and isoleucine residues were exchanged between the CLRDV-typical and -atypical strains showed that this substitution alone was not consistent with differences in resultant symptom phenotypes observed among CLRDV isolates, suggesting additional mutations, or perhaps, aa substitutions leading to modifications in the P0 secondary structure may be responsible for distinct CBD symptoms associated with the CLRDV strains (Agrofoglio et al. 2019).

Among the P0 proteins for the three CLRDV strains, the aa 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 aa residue 107 in CLRDV-AL. A comparison of the P0 sequences for the CLRDV-AL isolates collected from three counties, Barbour, Elmore, and Macon County, respectively, showed that they shared >99% nt identity and 99% aa similarity, indicating that the three CLRDV-AL isolates were representatives of one virus strain. 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 U.S. 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 the ORF0. The CLRDV-typical strain (GU167940) was implicated as the putative minor parent, which could at least in part explain the close phylogenetic relationships apparent between the ORF0/P0 between CRLDV-typical and CLRDV-AL. Notably, the CLRDV-AL P0 F-box-like motif harbors the Ile→Val substitution at aa residue 72, like that previously reported for CLRDV-atypical strains from BR and AR (da Silva et al.2015; Agrofoglio et al. 2019). These patterns suggest that the ORF0 features of CLRDV-AL may have been contributed by a putative CLRDV-atypical-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-AL 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 convergent mutations or 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 AR and BR were susceptible to blue disease-associated CLRDV-typical strain. Similarly, cotton varieties developed for the U.S. and exposed to CLRDV-AL during 2017-2018 showed no resistance to CLRDV-AL. Whether cotton varieties with resistance to the CLRDV-atypical strain will protect against CLRDV-AL infection is not known because this strain has not been reported to occur anywhere except in the southern U.S. cotton-growing states (Avelar et al. 2019; Aboughanem-Sabanadzovic et al. 2019; ACES, 2019; Tabassum et al. 2019). Thus, identification and introgression of genetic resistance to blue disease into cotton varieties for the U.S. southern states has become essential. Because whether CLRDV-AL will spread to other cotton-growing areas of the U.S. is not known, measures should be taken to identify resistance to varieties adapted to the desert Southwest and midwestern cotton growing states, e.g. Arizona, California, Oklahoma, and Texas.

Because the P0 sequence of CLRDV-AL is >10% divergent from previously reported CLRDV P0 sequences, namely ‘typical’ and ‘atypical’ CLRDV strains, whether the disease can be managed by the genetic resistance available in the CBD-tolerant varieties developed in South America remains to be determined. Also, whether CLRDV-AL has a resistance-breaking capacity similar to the CLRDV-atypical isolates is not known.
Thus, the relationship between the virulence of CLRDV-AL, specifically with respect to its P0-VSR activity, in U.S. varieties compared to CBD-resistant cotton developed in AR and BR requires immediate attention. However, the similarities between the disease symptoms associated with CLRDV-AL-infected cotton in Alabama and CLRDV-atypical isolates from South America, suggest that CLRDV-AL has VSR activity similar to that associated with the CLRDV-atypical strain.

The P0 protein of CLRDV-AL is highly divergent from P0 of ‘typical’ and ‘atypical’ CLRDV isolates, making CLRDV-AL a new, previously undescribed strain 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 strains, during screening and selection of cotton germplasm for resistance.

ACKNOWLEDGMENTS

The authors wish to thank Prof. Tatsuya Nagata for providing primer sequences used for the CLRDV genome amplification. This research was supported by Cotton Incorporated grant # 06-829.

LITERATURE CITED


S. Avelar


Table 1. Results of the pairwise nucleotide and amino acid distance analyses of the viral open reading frames for *Cotton leafroll dwarf virus* (CLRDV)-Alabama, and the CLRDV-typical (CBD) and CLRDV-atypical (ACBD) strains using Sequence Demarcation Tool software v.1.2.

<table>
<thead>
<tr>
<th>Isolate</th>
<th>Sequence</th>
<th>Nucleotide and amino acid % identity (nt/aa)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ORF0 / P0</td>
<td>ORF1 / P1</td>
</tr>
<tr>
<td>CBD</td>
<td>GU167940_AR06</td>
<td>92.1 / 86.6</td>
</tr>
<tr>
<td>CBD</td>
<td>HQ827780_BR07</td>
<td>92.4 / 86.2</td>
</tr>
<tr>
<td>ACBD</td>
<td>KF906260_BR11</td>
<td>90.7 / 82.8</td>
</tr>
<tr>
<td>ACBD</td>
<td>KF906261_BR06</td>
<td>91.1 / 83.5</td>
</tr>
<tr>
<td>ACBD</td>
<td>KF359946_AR10</td>
<td>91.6 / 85.8</td>
</tr>
<tr>
<td>ACBD</td>
<td>KF359947_AR10</td>
<td>91.3 / 84.7</td>
</tr>
</tbody>
</table>
Table 2. Predicted recombination events detected for *Cotton leafroll dwarf virus* (CLRDV)-Alabama, and the cotton blue disease (CBD) and atypical cotton blue disease (ACBD)-associated complete CLRDV genome sequences.

<table>
<thead>
<tr>
<th>Event</th>
<th>Breakpoints*</th>
<th>Parents</th>
<th>Recombinant</th>
<th>Methods†</th>
<th>P value‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>1°</td>
<td>3279 (ORF1-ORF2) 5865 (?)</td>
<td>^KF906260_BR11 GU167940_AR06 KF359946_AR10 RGBMCS3</td>
<td>KF906261_BR06</td>
<td>4.61E-10</td>
<td></td>
</tr>
<tr>
<td>2°</td>
<td>65 (?) 188 (ORF0)</td>
<td>^Macon_AL_US18 GU167940_AR06 Unknown</td>
<td>RBMS</td>
<td>4.95E-19</td>
<td></td>
</tr>
<tr>
<td>3°</td>
<td>3482 (?) 3900 (ORFs 3-5)</td>
<td>^KF359947_AR10 GU167940_AR06 KF359946_AR10</td>
<td>BMC3</td>
<td>3.48E-04</td>
<td></td>
</tr>
</tbody>
</table>

* Numbering starts at the 5' end of the minus-strand primer-binding site and increases clockwise. (?), breakpoint could not be precisely pinpointed.

† R, RDP; G, GeneConv; B, Bootscan; M, MaxChi; C, Chimera; S, SisScan; 3, 3SEQ.

‡ The reported P values are for the methods indicated in red, and they are the lowest P values calculated for the region in question.

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

(?) The actual breakpoint position is undetermined.

^ The recombinant sequence may have been misidentified (one of the identified parents might be the recombinant).
Supplemental Table S1. Reverse-transcription polymerase chain reaction primers used for amplification of *Cotton leafroll dwarf virus* isolate from Alabama.

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence (5'-3')</th>
<th>Coordinates</th>
<th>Amplicon size</th>
<th>Annealing temperature</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLRDV3675F</td>
<td>CCACGTAGRCGCAACAGCGT</td>
<td>3676</td>
<td>309</td>
<td>62 °C for 20s, 56 °C for 10s</td>
<td>Sharman et al. 2015</td>
</tr>
<tr>
<td>Pol3982R</td>
<td>CGAGGCGCTGGAGATGAACT</td>
<td>3984</td>
<td>3516</td>
<td>55</td>
<td>Nagata T. Unpublished</td>
</tr>
<tr>
<td>P20F</td>
<td>ACAAAAAGACATAGGGGTTGT</td>
<td>1</td>
<td>56</td>
<td>62 °C for 20s, 56 °C for 10s</td>
<td>Sharman et al. 2015</td>
</tr>
<tr>
<td>P22R</td>
<td>CGAGGCTGGAGATGAACT</td>
<td>3516</td>
<td>857</td>
<td>62 °C for 20s, 56 °C for 10s</td>
<td>Modified from Sharman et al. 2015</td>
</tr>
<tr>
<td>AL_3130F</td>
<td>CCTTAATCTGGAGACCAGGA</td>
<td>3128</td>
<td>3616</td>
<td>55</td>
<td>Nagata T. Unpublished</td>
</tr>
<tr>
<td>AL_Pol3982R</td>
<td>AGAGGCGCTGGAGATGAACT</td>
<td>3984</td>
<td>4242</td>
<td>52</td>
<td>Correa et al. 2005</td>
</tr>
<tr>
<td>PL4F</td>
<td>TGCGACAAATAGTTAATGGAATACGGT</td>
<td>3616</td>
<td>3891</td>
<td>52</td>
<td>Correa et al. 2005</td>
</tr>
<tr>
<td>o3R</td>
<td>GTCTACCTATTTBGGRTTNTGGAA</td>
<td>4242</td>
<td>4815</td>
<td>52</td>
<td>Designed for this study</td>
</tr>
<tr>
<td>AL_3891F</td>
<td>ATCAGATTGCCGGCGGATTCA</td>
<td>3891</td>
<td>4815</td>
<td>52</td>
<td>Designed for this study</td>
</tr>
<tr>
<td>AL_4815R</td>
<td>GGGAGGGTAGATTCGCCAAG</td>
<td>4815</td>
<td>4659</td>
<td>52</td>
<td>Designed for this study</td>
</tr>
<tr>
<td>AL_4659F</td>
<td>CAACTACAGTCTAACAGAG</td>
<td>4659</td>
<td>5483</td>
<td>52</td>
<td>Designed for this study</td>
</tr>
<tr>
<td>AL_5483R</td>
<td>CCATGGCGGAGAATAGTTGC</td>
<td>5483</td>
<td>5299</td>
<td>52</td>
<td>Designed for this study</td>
</tr>
<tr>
<td>AL_5299F</td>
<td>TCTGACCTTGGAGTGCTGCAAG</td>
<td>5299</td>
<td>5865</td>
<td>52</td>
<td>Designed for this study</td>
</tr>
<tr>
<td>P4</td>
<td>ACACCGAAACCCAGGGAG</td>
<td>5865</td>
<td>5865</td>
<td>52</td>
<td>Designed for this study</td>
</tr>
<tr>
<td>P0_51F</td>
<td>TGCTTGAGAATTTGCTCCTGT</td>
<td>51</td>
<td>51</td>
<td>52</td>
<td>Designed for this study</td>
</tr>
<tr>
<td>P0_916R</td>
<td>TTGACCCAGCAACACACTGG</td>
<td>918</td>
<td>918</td>
<td>52</td>
<td>Designed for this study</td>
</tr>
</tbody>
</table>

Coordinates relative to CLRDV-AL isolate (Accession no. MN071395)
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 (kDa) were determined, using Geneious Prime software, 2019.1.3 http://www.geneious.com/.

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

Fig. 3. Phylogenetic relationships for the Cotton leafroll dwarf virus (CLRDV)-AL and the six previously reported CLRDV isolates based on Maximum likelihood analysis (1000 bootstrap iterations). A, Phylogenetic tree showing the relationships between the CLRDV isolates based on the full-length genome sequence. B, Phylogenetic tree showing relationships between the CLRDV isolates based on the nucleotide sequence of open reading frame 0 (ORF0). C, Phylogenetic tree showing the relationships
between the CLRDV isolates based on the amino acid sequences of the ORF0-encoded protein zero.

SFig.1. Alignment of the amino acid sequence for the Cotton leafroll dwarf virus (CLRDV) open reading frame 0-encoded protein, P0, determined for CLRDV isolates collected in three Alabama counties, Barbour, Elmore, and Macon. Discrepancies identified in the consensus sequence are highlighted.

SFig.2. Neighbor-net tree based on the complete nucleotide sequence for the seven full-length genome sequences of Cotton leafroll dwarf virus, showing three CLRDV groups, corresponding to the atypical, typical, and Alabama strains.
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 (kDa) were determined, using Geneious Prime software, 2019.1.3 http://www.geneious.com/.

298x69mm (300 x 300 DPI)
Fig. 2. The predicted amino acid (aa) secondary structure for the open reading frame encoding the protein 0 (P0) for the seven available Cotton leafroll dwarf virus (CLRDV) isolates. The black circles indicate the aa residues unique to CLRDV-AL at residues that are conserved for the CLRDV-typical and CLRDV-atypical strains. The pink structure represents the predicted alpha helices, while the blue symbols represent turns, the yellow arrows indicate beta sheets, and the gray symbols indicate coil-coil secondary structure. The P0 F-box-like motif is indicated by a grey rectangle, and the red circles indicate the conserved aa residues within the motif, according to Agrofoglio et al. (2019). The black asterisk indicates AA residue 72, where the isoleucine residue in CLRDV-typical P0 has been substituted for a valine in the P0 of both the CLRDV-atypical strain and CLRDV-AL.

231x186mm (300 x 300 DPI)
Fig. 3. Phylogenetic relationships for the Cotton leafroll dwarf virus (CLRDV)-AL and the six previously reported CLRDV isolates based on Maximum likelihood analysis (1000 bootstrap iterations). A, Phylogenetic tree showing the relationships between the CLRDV isolates based on the full-length genome sequence. B, Phylogenetic tree showing relationships between the CLRDV isolates based on the nucleotide sequence of open reading frame 0 (ORF0). C, Phylogenetic tree showing the relationships between the CLRDV isolates based on the amino acid sequences of the ORF0-encoded protein zero.

587x148mm (300 x 300 DPI)
SFig.1. Alignment of the amino acid sequence for the Cotton leafroll dwarf virus (CLRDV) open reading frame 0-encoded protein, P0, determined for CLRDV isolates collected in three Alabama counties, Barbour, Elmore, and Macon. Discrepancies identified in the consensus sequence are highlighted.

276x64mm (300 x 300 DPI)
SFig. 2. Neighbor-net tree based on the complete nucleotide sequence for the seven full-length genome sequences of Cotton leafroll dwarf virus, showing three CLRDV groups, corresponding to the atypical, typical, and Alabama strains.

641x454mm (600 x 600 DPI)