scientific reports

OPEN

Positive selection and intrinsic disorder are associated with multifunctional C4(AC4) proteins and geminivirus diversification

Carl Michael Deom, Marin Talbot Brewer & Paul M. Severns

Viruses within the *Geminiviridae* family cause extensive agricultural losses. Members of four genera of geminiviruses contain a *C4* gene (*AC4* in geminiviruses with bipartite genomes). *C4(AC4)* genes are entirely overprinted on the *C1(AC1)* genes, which encode the replication-associated proteins. The C4(AC4) proteins exhibit diverse functions that may be important for geminivirus diversification. In this study, the influence of natural selection on the evolutionary diversity of 211 *C4(AC4)* genes relative to the *C1(AC1)* sequences they overlap was determined from isolates of the *Begomovirus* and *Curtovirus* genera. The ratio of nonsynonymous (d_N) to synonymous (d_S) nucleotide substitutions indicated that *C4(AC4)* genes are under positive selection, while the overlapped *C1(AC1)* sequences are under purifying selection. Ninety-one of 200 *Begomovirus C4(AC4)* genes encode elongated proteins with the extended regions being under neutral selection. *C4(AC4)* genes from begomoviruses isolated from tomato from native versus exotic regions were under similar levels of positive selection. Analysis of protein structure suggests that C4(AC4) proteins are entirely intrinsically disordered. Our data suggest that non-synonymous mutations and mutations that increase the length of C4(AC4) drive protein diversity that is intrinsically disordered, which could explain C4/AC4 functional variation and contribute to both geminivirus diversification and host jumping.

Geminiviridae is the largest family of plant viruses and contains more species than any other virus family[1]. Geminiviruses cause diseases in a wide range of cultivated and wild hosts worldwide[2]. In cultivated crops these diseases result in significant economic losses and routinely jeopardize food security[3]. Why this virus family, particularly the genus *Begomovirus,* is so diverse is not well-understood. Geminiviruses have small mono- or bipartite circular ssDNA genomes (~2.5–5.2 kb) and encode for 4–8 proteins[4]. Four of the nine genera, *Curtovirus* (three species), *Topocuvirus* (one species), *Turncurtovirus* (three species) and *Begomovirus* (>420 species) contain a small gene, designated *C4* (*AC4* in bipartite begomoviruses), that is nested within the *C1* gene (*AC1* in bipartite begomoviruses)[4]. *C4(AC4)* genes encode for, or have the potential to encode for, a protein of approximately 10 kDa[4]. *C4(AC4)* likely originated by the mechanism of overprinting within *C1(AC1)* (Fig. 1a), which occurs when mutations result in the generation of a de novo gene that overlaps an existing gene. Generally, overprinting occurs through a +1 frameshift[5,6], which is the case with *C4(AC4)* (Fig. 1b).

Gene overprinting allows for viruses to increase coding capacity, and subsequently genetic variability, while maintaining small, compact genomes. In viruses, overlapping genes may experience different forms of selection, including being under evolutionary constraint[7,8], being under differing evolutionary constraints with the de novo gene usually being under positive or weak purifying selection and the ancestral gene being under purifying selection[6], or being under independent adaptive evolution[9]. The viral proteins encoded by many overprinted de novo genes are predicted to be intrinsically disordered and are commonly accessory proteins affecting pathogenicity or viral movement rather than virus replication or structure[5], which is in line with the presently known functions of C4(AC4).

The *C1(AC1)* gene is conserved in genome location and function throughout the *Geminiviridae* family[4], whereas *C4(AC4)* genes are present in only 4 geminivirus genera, so we assume for further discussion that *C1(AC1)* represents the ancestral gene. *C1(AC1)* encodes a multifunctional replication-associated protein (Rep),

Department of Plant Pathology, University of Georgia, Athens, GA 30602, USA. email: deom@uga.edu

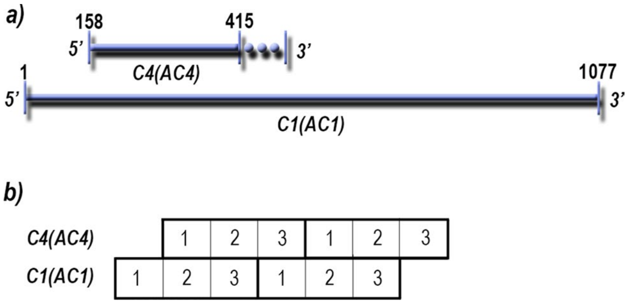

Figure 1. *C4(AC4)* overprinted on *C1(AC1)*. (**a**) *C4(AC4)* gene overprinted on the *C1(AC1)* gene. Nucleotide positions refer to nucleotide positions of isolate AF379637 (Beet curly top virus, BCTV-CA/Logan[US:Log:76]) where the BCTV *C4* gene encodes for a protein of 85 amino acids. The region following nucleotide position 415 indicates that some geminiviruses have *C4(AC4)* genes that encode for proteins that have lengths greater than 85 amino acids. (**b**) Schematic showing + 1 frameshift of *C4(AC4)* relative to the *C1(AC1)*.

which is the only geminivirus protein essential for replication and, therefore, likely to be under strong purifying selection. In addition, C1(AC1) proteins, which do not have polymerase activity, restore DNA replication competency to terminally differentiated host cells establishing an environment for virus replication[10]. The *C4(AC4)* gene is overprinted on a portion of the 5′-half of the *C1(AC1)* gene that encodes for domains involved in DNA binding and cleavage/ligation, oligomerization, and in interacting with multiple proteins[11].

Available information on the role of the C4(AC4) proteins from isolates of the *Curtovirus* and *Begomovirus* genera suggests a variety of functions, which likely indicate the evolving nature of the proteins. Known functions within the curtovirus C4 proteins include the induction of hyperplasia[12–16] and a role in systemic movement[17]. Functions within monopartite begomovirus C4 proteins include compromising the hypersensitive response[18], enhancing drought tolerance[19], enhancing virus accumulation[20], induction of hyperplasia[20], movement[21–23], suppression of the salicylic acid defense response[24], and suppression of RNA silencing[20,25–30]. Functions within bipartite begomovirus C4 proteins include induction of hyperplasia[31], maintenance of a betasatellite[32], and suppression of RNA silencing[33–36]. In some cases, bipartite begomovirus C4(AC4) proteins have no detectable functions[37–40]. Many of the C4(AC4) functions likely arise from multiple C4(AC4)-host protein interactions, including Shaggy-like protein kinases[16,31,41], Clavata1-type plasma membrane receptor-like kinases[25,33,42], S-adenosyl Methionine Synthetase[26], and Hypersensitive Induced Reaction 1[18]. Some AC4 related functions result from the protein binding to miRNAs and siRNAs and acting as a suppressor of RNA-silencing[28,36]. The function of the C4 proteins encoded in the *Topocuvirus* and *Turncurtovirus* genera have not been studied to date.

Limited information is available on genetic diversity within the *C4(AC4)* genes. Previous studies were restricted to natural populations of single geminivirus species and were limited in their inferences. An analysis of natural populations on a local scale of Tobacco leaf curl virus (21 isolates; this virus is likely referred to now as Eupatorium yellow vein virus) from *Eupatorium makinoi*[43] or begomoviruses causing cotton leaf curl disease (14 isolates) from cotton[7] suggest that *C4* is evolutionarily constrained and conserved through purifying selection. A study of Tomato yellow leaf curl China virus (TYLCCNV) within a naturally infected tomato (*Solanum lycopersicum*) plant, or *Nicotiana benthamiana* and tomato plants inoculated with a DNA clone of TYLCCNV, reported that a higher mutational rate occurs in the *AC1-AC4* overlapping region than the upstream non-overlapping region of *AC1*[44]. More recently, research on the genetic diversity of the *C4* and *C1* overlapping sequences from 11 isolates of Tomato leaf deformation virus[45] indicated that the *C4* gene sequence was under positive selection and corresponding *C1* gene sequence was under purifying selection.

To gain a broader and more extensive understanding of the evolution of a diverse group of geminivirus *C4(AC4)* gene sequences relative to the *C1(AC1)* gene sequences that they overlap, the ratios of the rates of nonsynonymous (d_N) to synonymous (d_S) nucleotide substitutions ($\omega = d_N/d_S$) were determined for each coding region from isolates representing 200 species of begomoviruses and from 11 isolates of a single curtovirus species. We also evaluated *C4(AC4)* diversity within a group of begomoviruses isolated from tomato in native and exotic locations of the host. Last, we compare the intrinsically disordered nature of the geminivirus C4(AC4) proteins to the amino acid sequence from the overprinted C1(AC1) region to evaluate the potential for structural-disorder derived functional diversify in C4(AC4) that might explain how the protein can obtain varying roles. Our analyses suggest that begomovirus and curtovirus C4(AC4) proteins are rapidly evolving through strong positive selection and that the intrinsically disordered nature of C4/AC4 proteins could generate multifunctional characteristics.

Results

Selection of geminivirus *C4(AC4)* and *C1(AC1)* gene sequences.

N-terminal myristoylation of C4(AC4) proteins from the *Curtovirus* and *Begomovirus* genera has been shown to be necessary for function[16,31,34]. Therefore, all *C4(AC4)* gene sequences analyzed begin with the ATG codon that immediately precedes a *N*-terminal myristoylation motif. All *C4(AC4)* genes are in a + 1-frameshift relative to the *C1(AC1)* reading frame, so the *C1(AC1)* coding sequences analyzed begin at the third nucleotide of the *C4(AC4)* gene, to retain the *C1(AC1)* reading frame. The *Begomovirus* genus is composed of > 420 species[4]. We randomly selected

200 species for analysis, using the exemplary isolate from each species[4]. Isolates analyzed from the *Begomovirus* genus fall into six groups, C4(AC4) proteins composed of 85, 90, 94, 96, 97 or 100 amino acids in length (Supplementary Table S1). All *C4* genes in isolates from the *Curtovirus, Topocuvirus* and *Turncurtovirus* genera encode proteins of 85 amino acids in length (255 nucleotides). The *Curtovirus* genus is composed of three species, with only the Beet curly top virus (BCTV) species having a *N*-terminal myristoylation motif in most of its isolates. Eleven isolates from the BCTV species were analyzed. Two *Curtovirus* species, Horseradish curly top virus and Spinach severe curly top virus, have single isolates that do not have a *N*-terminal myristoylation motif and were not included in the analysis. From the *Turncurtovirus* genus, four isolates of the Turnip curly top virus species and the single isolate of the Turnip leaf roll virus species were included in the maximum likelihood (ML) phylogenetic analysis and the intrinsically disordered protein analysis, as was, Tomato pseudo-curly top virus, the single isolate in the *Topocuvirus* genus. BCTV coat protein (CP) gene sequences (762 nucleotides), which are conserved[7,46,47], were used as a control for evaluating purifying selection. Stop codons were not included in the analyses.

C4(AC4) genes are under positive selection.

We initially looked at the evolutionary relationships among *C4(AC4)* genes, as well as the overprinted *C1(AC1)* gene sequences. A phylogenetic tree (Fig. 2) was inferred using maximum likelihood in MEGAX[48]. The inferred phylogeny of *Geminiviridae* members based on *C4(AC4)* and *C1(AC1)* sequences had little to no strong statistical support as many of the terminal nodes and nearly all of the basal nodes had uniformly low statistical support (bootstrap support values < 0.5). While the phylogeny is poorly resolved, the lack of resolution and lack of strong node support suggests that the *C4(AC4)* and *C1(AC1)* genes accrue a high number of mutations relatively quickly, which obscures the phylogenetic signal from lineages with shared descent. Interestingly, 1 BCTV isolate (M24597, BCTV/CA/Logan) did not group with the other 10 BCTV isolates, suggesting this isolate has undergone recombination.

To determine if *C4(AC4)* is under positive selection we analyzed nucleotide substitutions in *C4(AC4)* compared to substitutions in the same portion of the *C1(AC1)* sequence that *C4(AC4)* overlaps for a diverse group of *Begomovirus* and *Curtovirus* isolates (Fig. 2). For the 200 *Begomovirus C4(AC4)* gene sequences, ω (the ratio of the number of nonsynonymous substitutions per nonsynonymous site to the number of synonymous substitutions per synonymous site) was 1.528 for the initial 255 nucleotides representing 85 amino acids, indicating that *C4(AC4)* is under strong positive selection (Table 1). Likelihood ratio tests (LRT) applied to the three pairs of models in CODEML indicated that ω varied across codons (M0 vs. M3) and that *C4(AC4)* was under positive selection (M2a vs. M1a and M8 vs. M7) (Supplementary Table S2). The Bayes empirical Bayes (BEB) analysis showed that 23 codon sites were under positive selection with a posterior probability (PP) of ≥ 95% (Table 1), which represents approximately 27% of the amino acid residues encoded by the *C4(AC4)* gene. Additionally, the codons under positive selection were not restricted to specific regions of the gene, but were spread across the entire length of *C4(AC4)*, suggesting that mutations across the entire gene sequence generates functional protein diversity (*e.g.* absence of premature stop codons). In contrast, the length of the *C1* sequences that the *C4(AC4)* is overprinted on were under purifying selection (ω = 0.127; Table 1). Although ω varied across codons (M0 vs. M3), the overprinted region of *C1* was not under positive or neutral selection (M2a vs. M1a and M8 vs. M7) (Supplementary Table S2), nor were any individual codons suggested to be under positive selection (BEB) (Table 1).

It was previously suggested that coat protein genes from New World (NW) begomoviruses are under stronger purifying selection than those from Old World (OW) begomoviruses[49]. To determine if NW *C4(AC4)* genes are under lower positive selection than OW *C4(AC4)* genes, we compared the ω-values of *C4(AC4)* from NW versus OW begomoviruses. Of the 200 begomoviuses we analyzed above, 79 were NW viruses and 121 were OW viruses (Fig. 2). The ω-values were 1.545 and 1.550 for the initial 255 nucleotides of the 79 NW *C4(AC4)* genes and the 121 OW *C4(AC4)* genes, respectively. This suggests that *C4(AC4)* genes from NW and OW viruses are under near identical levels of strong positive selection. In contrast, the ω-values of the *C1(AC1)* gene sequences from NW and OW begomoviruses that the *C4(AC4)* genes are overprinted on were under near identical levels of purifying selection (ω = 0.132 and ω = 0.131, respectively).

Ninety-one of the 200 begomoviruses analyzed have C4(AC4) proteins of 96–100 amino acids in length (Supplementary Table S1). An analysis of the extended regions (nucleotides 256–288, representing 11 amino acids) indicated that they were under neutral selection (ω = 0.935) (Table 1, Supplementary Table S2). BEB analysis indicated that 1 of the 11 amino acids were under positive selection with a posterior probability (PP) of ≥ 95%. In contrast, the extended regions represented by the *C1(AC1)* sequences were under purifying selection (ω = 0.321). (Table 1). Of the 91 begomovirus, 83 were OW, suggesting that extended *C4(AC4)* genes are much more common in OW begomoviruses than in NW begomoviruses.

When we analyzed the begomovirus *C4* sequences from 11 isolates from the BCTV species (Fig. 2), the *C4* sequences were also under positive selection (ω = 1.143), while the overlapped *C1* sequences were under purifying selection (ω = 0.194) (Table 1 and Supplementary Table S2). BEB analysis indicated only 3 codons were under strong positive selection (PP ≥ 95%) in the BCTV *C4* genes analyzed (Table 1). As an additional test of purifying selection at a region outside of the *C1* sequences, we determined ω for the coat protein (CP) genes of the 11 BCTV isolates. As expected, the BCTV CP genes, which are highly conserved, were under strong purifying selection (ω = 0.041; Table 1), considerably more so than the *C1* sequences.

To look at variation within site positions that lead to amino acid changes, the relative nucleotide substitution rates within the first, second and third codon positions were determined (Fig. 1b and Table 2). The most variable codon position (highest relative substitution rate) for all overlapping sequences analyzed was the second position of *C4,* which corresponds to the third position of *C1* [C4(AC4)-2/*C1*(AC1)-3]. To estimate the uniformity of variation at codons, the α parameter was determined as a function of the discrete gamma distribution for each overlapping region[50] (Table 2). Begomovirus codon position C4(AC4)-2/C1(AC1)-3 (in nucleotides 1–255)

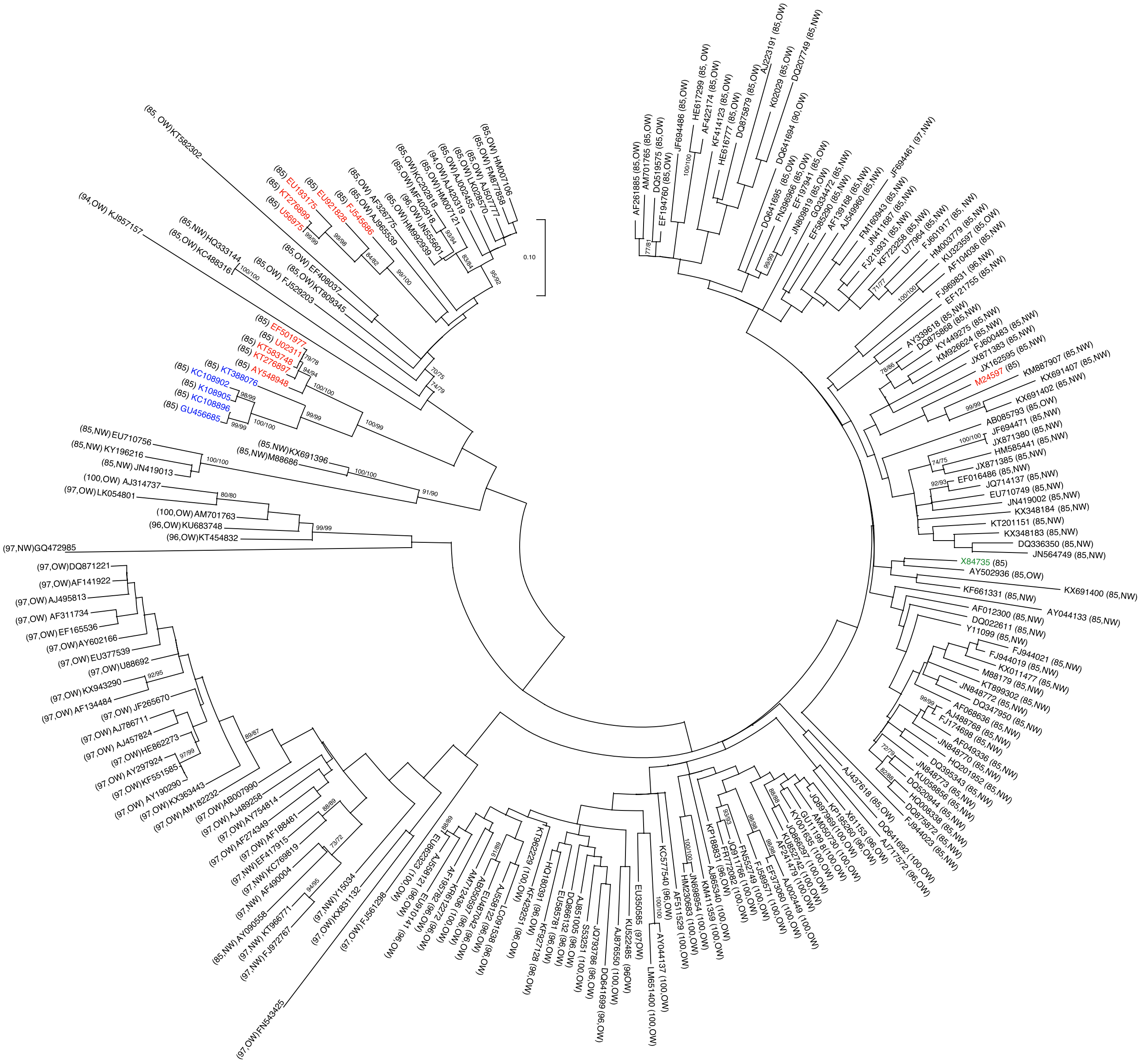

Figure 2. Maximum likelihood (ML) phylogenetic tree of *C4(AC4)* sequences of diverse geminivirus isolates. ML tree based on a Hasegawa, Kishino, and Yano model of evolution assuming a gamma distribution and invariable sites (HKY + G + I) showing genetic relationships among *C4(AC4)* sequences (nucleotides 1–255) from four genera of the family *Geminiviridae*; exemplary isolates from 200 species of the *Begomovirus* genus, 11 isolates from the Beet curly top virus species of the *Curtovirus* genus, four isolates of the Turnip curly top virus species and the single isolate of the Turnip leaf roll virus species from the *Turncurtovirus* genus, and Tomato pseudo-curly top virus, the single isolate in the *Topocuvirus* genus. The ML tree for *C1(AC1)* sequences (255 nucleotides) that the *C4(AC4)* genes overlap is highly identical and not shown. Gene sequences are taken from full-length genomic sequences and the GenBank accession number of each virus is indicated. Preceding or following the GenBank accession numbers in parentheses are the amino acid length of the respective C4(AC4) proteins and the designation for New World (NW) or Old World (OW). Numbers at branches represent bootstrap values (≥ 70% are shown) determined from 500 replicates. Since the ML trees for *C4(AC4)* and *C1(AC1)* are nearly identical, bootstrap values for *C4(AC4)* and *C1(AC1)* are separated by a slash. The bar near the top of the cladogram represents nucleotide substitutions per site for the *C4(AC4)* sequences. Virus isolate(s) in black text are begomoviruses, red text are curtoviruses, blue text are turncurtoviruses, green text is a topocuvirus.

displayed an α-value of 1.57. An α > 1 with a high relative substitution rate compared to the other positions indicates that C4(AC4)-2/C1(AC1)-3 positions have intermediate substitution rates, while a few codons have very high or very low relative substitution rates[50]. The high substitution rate and infinitely large α-value for the

Coding region	Nucleotides in coding region	Amino acids	ω[a]	Positively selected sites[b]
Gene sequences of 200 begomoviruses				
C4/AC4	255	85	1.528	30 (**23**)[c] **3S**, 6S, **8C**, **9L**, **10F**, 11S, **14E**, **16T**, **17T**, **19K**, **21N**, **27Y**, **29Q**, **30P**, **31G**, **37Q**, **43Q**, **44A**, 52T, **57P**, 58L, **62N**, **67E**, 73A, **74A**, **75R**, **76T**, 80L, 82Q, **84P**
C1/AC1	255	85	0.127	–
Gene sequences of 91 begomovirus extensions				
C4/AC4	33	11	0.935	3 (**1**)[d] 4A, 6S, **7L**
C1/AC1	33	11	0.321	–
Gene sequences of 11 curtovirus isolates				
C4	255	85	1.143	23 (**3**)[e] 4L, 9C, 16F, 17R, 19Q, 20I, 23Y, 26W, 27Y, **29Q**, 30P, **31G**, 33H, 34I, 37Q, 45A, 50P, **53T**, 68V, 77Q, 81H, 82M, 85R
C1	255	85	0.194	–
CP	762	254	0.041	–

Table 1. Positive selection analyses of *C4(AC4)* or *C1(AC1)* gene sequences from begomovirus and curtovirus species. [a]Mean ratio of the nonsynonymous (d_N) and the synonymous substitution rates (d_S). ω < 1 represents purifying selection; ω = 1 neutral selection; and ω > 1 positive selection. [b]Number of amino acid sites under positive selection based on Bayes empirical Bayes analysis. Amino acid residue number and single letter amino acid designation having a posterior probability of P > 90% in regular font and of P ≥ 95% in bold font. [c]Amino acid positions and amino acids refer to HE616777 (African cassava mosaic Burkina Faso virus; ACMBFV). [d]Amino acid positions and amino acids refer to AJ851005 (Ageratum leaf curl virus; ALCuV). [e]Amino acid positions and amino acids refer to M24597 (Beet curly top virus; BCTV/CA/Logan).

Length (nucleotides)	Codon position	Relative rate	α parameter
Gene sequences from 200 begomovirus species			
255	C4(AC4)-1/C1(AC1)-2	1[a]	0.49
255	C4(AC4)-2/C1(AC1)-3	2.41	1.57
255	C4(AC4)-3/C1(AC1)-1	1.19	0.48
Gene extension sequences from 91 begomovirus species			
33	C4(AC4)-1/C1(AC1)-2	1[a]	0.30
33	C4(AC4)-2/C1(AC1)-3	6.17	0.67
33	C4(AC4)-3/C1(AC1)-1	2.67	0.65
Gene sequences from 11 curtovirus isolates			
255	C4-1/C1-2	1[a]	0.36
255	C4-2/C1-3	2.75	∞
255	C4-3/C1-1	1.66	0.72

Table 2. Codon-site nucleotide variation in *C4(AC4)* and *C1(AC1)* overlapping regions. [a]The estimated substitution rates are determined relative to this position. For codon position, refer to Fig. 1b.

curtovirus codon positions C4-2/C1-3 indicates a constant substitution rate for all sites. In contrast, codon positions C4(AC4)-2/C1(AC1)-3 for the begomovirus 33 nucleotide extension, had a very high relative substitution rate compared to the other codon positions, but had an α-value of 0.67. An α ≤ 1 indicates that most of the codon positions have very low substitution rates with some sites being hotspots with very high substitution rates[50].

Selective pressure on begomoviruses from tomato from native and exotic locations.

The movement of crops around the world results in plants encountering new pathogens. This 'new encounter phenomenon' occurs when a crop has been introduced into a new geographical region and pests and/or pathogens that evolved with native host species in the introduced region infect and cause disease in the newly introduced exotic crop species[51]. In this study, the native range of a plant is defined as the geographic region in which the species originated as well as the surrounding area where indigenous people may have moved and domesticated the species. Exotic is defined as regions far enough outside the native range of the plant that indigenous people were not likely involved in moving the species, rather it was introduced by movement over long distances. When we tallied whether the begomovirus isolates of > 420 species originated from a host plant within its native range or outside of its native range (exotic) we found that 52.5% of the *Begomovirus* species were isolated from a host plant outside of its native range (Supplementary Table S3).

A large number of *Begomovirus* species (112 of 420) have been isolated from tomato with approximately 70% of the species being isolated from tomato grown outside of western South America and Central America, the

Coding region	Nucleotides in coding region	Amino acids	ω[a]	Positively selected sites[b]
Gene sequences of 29 species from native locations				
C4/AC4	255	85	1.717	29 (**19**)[c] **8F**, **9S**, **10S**, 11N, **14G**, 15S, 16T, **17S**, **19R**, **21T**, **27S**, 31G, 37R, 39Y, **43P**, **44A**, **48S**, 49P, **57Q**, **58L**, 62N, **67A**, 70L, **74S**, **75N**, **76L**, 77L, **82P**, **84R**
C1/AC1	255	85	0.134	2 (**1**) 57V, **73Q**
Gene sequences of 63 species from non-native locations				
C4/AC4	255	85	1.359	28 (**23**)[d] **3L**, **6S**, **8P**, **9S**, 10S, **11S**, **14V**, **16P**, **17S**, 18S, **19E**, **21P**, **26S**, **27L**, **30I**, **31T**, 37Q, **44P**, **45A**, **49S**, 51T, **53R**, **58T**, **68E**, **74V**, **76R**, 81Q, **85P**
C1/AC1	255	85	0.157	3(0) 2S, 16Q, 18R

Table 3. Positive selection analyses of *C4(AC4)* or *C1(AC1)* gene sequences from begomovirus species isolated from *Solanum lycopersicum* from native and non-native locations. [a]Mean ratio of the nonsynonymous (d_N) and the synonymous substitution rates (d_S). $\omega < 1$ represents purifying selection; $\omega = 1$ neutral selection; and $\omega > 1$ positive selection. [b]Number of amino acid sites under positive selection base on Bayes empirical Bayes analysis[47]. Amino acid residue number and single letter amino acid designation having a posterior probability of $P > 90\%$ in regular font and of $P \geq 95\%$ in bold font. [c]Amino acid positions and amino acids refer to AF101476 (Chino del tomate virus; CdTV). [d]Amino acid positions and amino acids refer to JN135234 (Pepper leaf curl Lahore virus; PepLCLaV).

native range of tomato and areas where it was likely spread by indigenous people. Ninety-two of the 112 species (29 of the 34 native and 63 of the 78 exotic) had *C4(AC4)* genes with a N-terminal myristoylation motif and a length of 85–100 amino acids. If begomovirus isolates travelled with tomato plants or seed when they were moved from the New World, we might expect to see evidence of purifying selection in *C4(AC4)* if a host-specific interaction would be essential for begomoviruses on tomato outside of its native range (*e.g.* no reservoir of congeners for viruses to move from onto tomato). *C4(AC4)* gene sequences from begomovirus isolates from tomato in native (29 begomoviruses) and exotic (63 begomoviruses) locations (Supplementary Table S4) were under strong positive selection with ω-values of 1.717 and 1.359, respectively, for the initial 255 nucleotides representing 85 amino acids (Table 3). Likelihood ratio tests (LRT) applied to the three pairs of models in CODEML indicated that ω varied across codons (M0 vs. M3) and that *C4(AC4)* was under positive selection (M2a vs. M1a and M8 vs. M7) (Supplementary Table S5). The codons under positive selection were located throughout the length of the *C4(AC4)* reference sequence, suggesting changes across the entire gene sequence may generate protein diversity. As expected, the *C1(AC1)* sequences were under purifying selection (Table 3 and Supplementary S5). Therefore, there is no evidence of either purifying or neutral selection in *C4(AC4)* genes from tomato-associated isolates taken from native or exotic regions.

Geminivirus *C4(AC4)* genes encode intrinsically disordered proteins. Most overprinted de novo virus genes are predicted to encode intrinsically disordered proteins (IDPs)[5]. The inherent flexability with IDPs is a major factor in the interactions with multiple partners affecting cellular interactions regulating development, metabolic and signaling pathways and stress responses[52,53]. To determine if this was the case with the C4(AC4) proteins, 2 protein disorder prediction programs were utilized. Both disorder predictors gave similar results with all 217 C4(AC4) proteins predicted to be fully disordered (Fig. 3a–c; Supplementary Table S6). The same analysis predicts that overlapping *C1(AC1)* coding regions are primarily ordered. This is especially true for the overlapping *C1(AC1)* regions of begomoviruses *C4(AC4)* gene sequences encoding for proteins of 100 amino acids in length (Fig. 3), which would suggest that within the intact C1(AC1) proteins this region is highly ordered. As a control, the disorder predictors were used to analyze the structure of the curtovirus CP. Since the CPs of icosahedral viruses have a canonical ß-barrel jelly roll structure[54], we expect the BCTV CPs to be mostly ordered. The N-termini of the BCTV CPs are disordered (~ 18% of the proteins), while the C-terminal (82% of the protein) is predicted to be ordered (Fig. 3d; Supplementary Table S6). While the ordered region includes the ß-barrel jelly roll structure of the CPs, the disordered N-terminal region has been shown to bind DNA and take on order that results in capsid assembly[55,56]. Taken together, the results suggest that C4(AC4) proteins are IDPs while C1(AC1) proteins are mostly ordered.

Discussion

Overprinted de novo genes in virus genomes are thought to be an evolutionary mechanism to maximize coding potential and to provide genetic novelty. Frequently, young de novo genes are rapidly evolving relative to the ancestral gene, allowing for the evolution and development of diverse novel genetic functions, while older de novo genes may be under purifying selection and evolve more slowly[6]. Although nucleotide variation occurs and is generally assumed to be selected for in the overprinted gene, at least in the + 1 frameshift, the amino acid sequence of the protein encoded by the ancestral gene is conserved due to redundancy in the genetic code. Most of the nucleotide variation occurs in the third position (wobble position, genetic degeneracy allows for a higher probability of a synonymous change) of the codons of the ancestral gene, which is the second codon position in the overprinted gene (+ 1 frameshift) (Fig. 1b). Mutations in codon position 2 that do not result in stop codons are

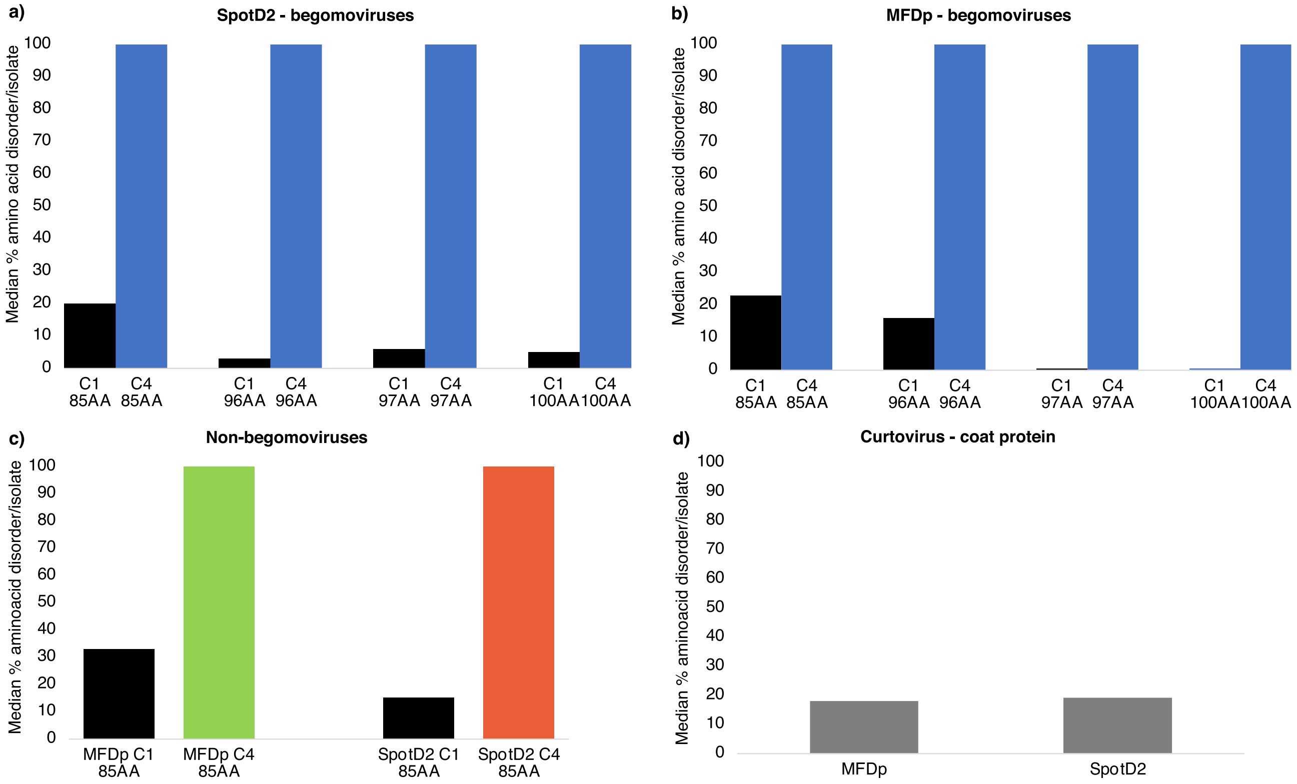

Figure 3. Intrinsically disordered protein analysis. (**a**) Medium percentage amino acid intrinsic disorder/begomovirus isolate determined by SPOT-Disorder2 for C4(AC4) proteins and the overlapped C1(AC1) protein regions. (**b**) Medium percentage amino acid intrinsic disorder/begomovirus isolate determined by MFDp2 for C4(AC4) proteins and the overlapped C1(AC1) protein regions. For (**a**) and (**b**), 106 begomovirus isolates encoded C4(AC4) proteins and the corresponding C1(AC1) regions of 85 amino acids in length, 26 isolates of 96 amino acids in length, 37 isolates of 97 amino acids in length and 28 isolates of 100 amino acids in length (Supplementary Table S6). (**c**) SPOT-Disorder2 and MFDp2 analysis of encoded C4(AC4) proteins and the overlapped C1(AC1) protein regions of non-begomovirus isolates, 85 amino acids in length, that included the isolates from the *Curtovirus*, *Turncurtovirus* and *Topocuvirus* genera (Supplementary Table S6). (**d**) SPOT-Disorder2 and MFDp2 analysis of curtovirus coat proteins, 254 amino acids in length (Supplementary Table S6).

always non-synonomous mutations that lead to an amino acid change. Indeed, the α parameter value indicated that position 3 of *C1(AC1)* codons and position 2 of *C4(AC4)* codons had the highest relative rate of nucleotide substitution and were the most variable. Thus, the overprinted gene, *C4(AC4)*, undergoes positive selection while the ancestral gene, *C1(AC1)*, remains under purifying selection.

The overall mutation rate in the overlapping genes appears to be rapid enough to obscure a clear phylogenetic signal as we were unable to generate a majority resolved phylogeny with statistical confidence. However, purifying selection on *C1(AC1)* likely favors the accumulation of synonymous substitutions while positive selection on *C4(AC4)* favors the relatively rapid accrual of non-synonymous substitutions, and presumably C4(AC4) protein diversity.

Since the *C4(AC4)* genes that have been studied have diverse functions, we determined the type of selective pressure the genes are under relative to the *C1(AC1)* sequences the *C4(AC4)* genes overlap. As a group, the ω-values for the *C4(AC4)* genes of the begomoviruses ($\omega = 1.528$) were under greater positive selection than the *C4* of the curtovirus isolates ($\omega = 1.143$) within nucleotides 1–255. This difference is likely because the curtovirus isolates are from 1 species, while the begomovirus isolates represent 200 species, reflecting greater genetic and taxonomic diversification.

Relative to the positive selection observed for nucleotides 1–255 in begomovirus *C4(AC4)* genes, near neutral selection ($\omega = 0.935$) occurred in the extended region of the *C4(AC4)* genes having protein lengths of 96–100 amino acids. While the extended region is not evolving as rapidly as the region of *C4(AC4)* representing nucleotides 1–255, diversity in the extended region is occurring more rapidly than the overlapped *C1(AC1)* region, which is still under purifying selection ($\omega = 0.321$). The elimination of a stop codon and increase in length of C4(AC4) proteins could provide a domain with novel function or the ability to modulate existing C4(AC4) function. It may be that the extended regions of the begomovirus *C4(AC4)* genes, especially in OW begomoviruses, represent a more recent event that results in more rapid evolution than that observed in the core *C4(AC4)* genes representing the initial 1–255 nucleotides. The generation of a de novo gene by overprinting and protein extension by elimination of stop codons has been proposed for the creation of the ALTO/MT gene in polyomviruses[57]. In the case of the begomoviruses, our study suggests that the entire length of the *C4(AC4)* (nucleotides 1–255) is available for generating protein diversity through strong positive selection as well as the elongation of *C4(AC4)* genes, albeit the extended regions evolve less rapidly. These data support the conclusion that positive selection of *C4(AC4)* genes could explain how the resulting proteins have evolved diversified and diverged functions. In contrast, purifying selection of *C1(AC1)* is consistent with its essential role in virus replication.

All C4(AC4) proteins analyzed were predicted to be IDPs. An important characteristic of IDPs is their ability to be promiscuous, binding multiple partners (i.e., proteins or nucleic acid)[53,58]. While the compact nature of virus genomes restricts the number of virus-encoded proteins, rapidly evolving IDPs would give the virus an advantage in the ability of the IDPs to interact with multiple host targets, resulting in multiple functions. Depending on the environment and binding partners, IDPs take on different conformations and different functions. Inducibility of more ordered structure may also occur through post-translation modifications[53], such as regulation of BRASSINISTEROID INSENSITIVE I KINASE INHIBITOR in the brassinosteroid pathway by phosphorylation[59,60]. Indeed, phosphorylation of Ser49 in the BCTV C4 protein is required for the protein to bind to and inhibit the function of AtSKs and negatively regulate the brassinosteroid pathway[16]. Therefore, phosphorylation of Ser49 may induce and stabilize a C4 conformation required for function. Since C4(AC4) proteins bind a number of targets, including various host proteins and mi/siRNAs, the intrinsically disordered nature of C4(AC4) may allow the proteins to be multifunctional and the high rate of non-synonymous substitutions would allow the proteins to abruptly modify existing functions or to take on new functions. Because of the high diversity within C4(AC4) proteins, each C4(AC4) might be expected to bind a specific set of host partners and members of these sets might overlap between different C4(AC4) proteins, allowing for some redundancy as well as unique specificities.

Our results indicate that curtovirus and begomovirus *C4(AC4)* genes encode for rapidly evolving IDPs. Typically, overprinted de novo genes encode for accessory proteins that contribute to viral pathogenicity, but are not structural proteins and are unnecessary for virus replication. Begomovirus *C4(AC4)* genes have been implicated in modulating numerous host defense reactions[18,20,24–30,33–36]. It is noteworthy that positive selection has been shown to drive rapid evolution of antiviral RNAi genes in *Drosophila* and multiple invertebrates[61,62]. Therefore, it might be inferred that rapid evolution of begomovirus *C4(AC4)* genes could be reflective of the proteins adapting to rapidly evolving antiviral RNAi host genes in plants. Indeed, plant virus suppressors of RNAi have been shown to be subject to positive selection that could be attributed to frequent shifts between host species[63].

Begomovirus is the largest virus genus with diversification primarily driven by mutational dynamics and substitution rates similar to RNA viruses[47,64]. Over half of the recorded *Begomovirus* species were isolated from plants that occurred in biogeographic regions outside of their native range (Table Supplementary S3). Tomato-associated begomoviruses represent a large group of viruses isolated from the host's native and exotic locations. Positive selection and the gene-wide distribution of positively selected codon positions on the *C4(AC4)* gene was comparable in tomato-associated *Begomovirus* regardless of geographic range of the tomato host. Furthermore, these same patterns of positive selection on *C4(AC4)* were observed throughout all the begomoviruses in this study. Because *C4(AC4)* has a diversity of known functions, including a role in host infection[65,66], it is tempting to speculate that positive, diversifying selection over the whole *C4(AC4)* gene, coupled with intrinsically disordered nature of C4(AC4) could provide begomoviruses the innate flexibility to more rapidly adopt to new hosts and radiate where they are found anywhere in the world.

Lastly, it is worth noting that geminivirus sequences in GenBank are biased towards crop plant hosts. Therefore, it is not unreasonable to predict that *C4(AC4)* variability may be greater, considering the number of begomovirus species is likely larger than those presently characterized when also considering geminiviruses yet to be identified from non-crop species worldwide. Future experiments will provide additional insights and clarity into the mechanism(s) by which C4(AC4) proteins function and will aid in determining the extent to which positive selective pressure and intrinsic disorder diversifies C4(AC4) function.

Methods and materials

Selection of geminivirus gene sequences.

All sequences analyzed were obtained from full-length geminivirus genomic sequences available in GenBank (GenBank accession numbers are given in Fig. 2). All begomovirus sequences used to determine the ratios of the rates of non-synonymous (d_N) to synonomous (d_S) nucleotide substitutions ($\omega = d_N/d_S$) were randomly selected from listed species available in the International Committee on Taxonomy of Viruses (ICTV) Report on the taxonomy of the *Geminiviridae*[4] (Available online at https://www.ictv.global/report/geminiviridae).

Non-synonymous-synonymous substitution rate and codon site nucleotide variation.

Phylogenetic trees were inferred using maximum likelihood (ML) in MEGAX[48]. Evolutionary models of nucleotide substitution were determined based on ML in MEGAX. Initially, ω was calculated based on one ratio (M0) with the maximum likelihood method by CODEML in PAML v4.8 package[67] contingent on nucleotide alignments and neighbor-joining trees based on the Jukes-Cantor model constructed in Geneious v6[68]. To detect the statistical power of positive selection and the individual amino acid residues under positive selection, likelihood ratio tests were applied on three pairs of models in the program CODEML to test for departure from neutral models: one ratio (M0) vs. discrete (M3), nearly neutral (M1a) vs. positive selection (M2a), and β (M7) vs. β & ω (M8). To determine if non-neutral models best fit the data, we compared twice the difference of the log-likelihood values of each pair of models using a χ^2 distribution. Amino acid sites under positive selection were estimated by Bayes empirical Bayes analysis based on the M8 model[69]. Codon substitution rates and α parameters were determined using a maximum likelihood approach in the module BASEML within PAML[67–69].

Predicting intrinsically disordered. There are numerous computational software tools available for predicting intrinsically disordered protein regions. Based on different methodologies available, SPOT-Disorder2 and MFDp2 were used to determine intrinsic disorder in C4(AC4) and C1(AC1) amino acid sequences from 217 geminivirus isolates. SPOT-disorder2 is an updated version of a highly rated disorder predictor that uses an ensemble of deep bidirectional long short-term memory and inception-residual squeeze-and excitation convolutional neural network[70–72]. MFPp2 is a highly rated meta-disorder predictor, which combines the methodology of a number of predictors[72,73].

Received: 11 December 2020; Accepted: 13 May 2021
Published online: 27 May 2021

References

1. Zhao, L., Rosario, K., Breitbart, M. & Duffy, S. Eukaryotic circular Rep-encoding single-stranded DNA (CRESS DNA) viruses: Ubiquitous viruses with small genomes and a diverse host range. *Adv. Virus Res.* **103**, 71–132 (2019).
2. García-Arenal, F. & Zerbini, F. M. Life on the edge: Geminiviruses at the interface between crops and wild plant hosts. *Annu. Rev. Virol.* **6**, 411–433 (2019).
3. Beam, K. & Ascencio, J. T. Geminivirus resistance: A minireview. *Front. Plant Sci.* **11**, 1131 (2020).
4. Zerbini, F. M. *et al.* ICTV Report Consortium. ICTV virus taxonomy profile: *Geminiviridae*. *J. Gen. Virol.* **98**, 131–133 (2017).
5. Rancurel, C., Khosravi, M., Dunker, A. K., Romero, P. R. & Karlin, D. Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. *J. Virol.* **83**, 10719–10736 (2009).
6. Sabath, N., Wagner, A. & Karlin, D. Evolution of viral proteins originated de novo by overprinting. *Mol. Biol. Evol.* **29**, 3767–3780 (2012).
7. Sanz, A. I., Fraile, A., Gallego, J. M., Malpica, J. M. & Garca-Arenal, F. Genetic variability of natural populations of cotton leaf curl geminivirus, a single-stranded DNA virus. *J. Mol. Evol.* **49**, 672–681 (1999).
8. Mizokami, M. *et al.* Constrained evolution with respect to gene overlap of hepatitis B virus. *J. Mol. Evol.* **44**(Suppl), S83–S90 (1997).
9. Zaaijer, H. L., van Hemert, F. J., Koppelman, M. H. & Lukashov, V. V. Independent evolution of overlapping polymerase and surface protein genes of hepatitis B virus. *J. Gen. Virol.* **88**, 2137–2143 (2007).
10. Hanley-Bowdoin, L., Bejarano, E. R., Robertson, D. & Mansoor, S. Geminiviruses: Masters at redirecting and programming plant processes. *Nat. Rev. Microbiol.* **11**, 777–788 (2013).
11. Fondong, V. N. Geminivirus protein structure and function. *Mol. Plant Pathol.* **14**, 635–649 (2013).
12. Latham, J. R., Saunders, K., Pinner, M. S. & Stanley, J. Induction of plant cell division by beet curly top virus gene C4. *Plant J.* **11**, 1273–1283 (1997).
13. Lai, J. *et al.* RKP, a RING finger E3 ligase induced by BSCTV C4 protein, affects geminivirus infection by regulation of the plant cell cycle. *Plant J.* **74**, 905–917 (2009).
14. Mills-Lujan, K. & Deom, C. M. Geminivirus C4 protein alters *Arabidopsis* development. *Protoplasma* **239**, 95–110 (2010).
15. Park, J. *et al.* C4 protein of Beet severe curly top virus is a pathomorphogenetic factor in *Arabidopsis*. *Plant Cell Rep.* **29**, 1377–1389 (2010).
16. Mills-Lujan, K., Andrews, D. L., Chou, C. & Deom, C. M. The roles of phosphorylation and SHAGGY-like kinases in Geminivirus C4 protein induced hyperplasia. *PLoS ONE* **10**, e0122356 (2015).
17. Teng, K. *et al.* Involvement of C4 protein of Beet severe curly top virus (Family *Geminiviridae*) in virus movement. *PLoS ONE* **5**, e11280 (2010).
18. Mei, Y., Ma, Z., Wang, Y. & Zhou, X. Geminivirus C4 antagonizes the HIR-mediated hypersensitive response by inhibiting the HIR1 self-interaction and promoting degradation of the protein. *New Phytol.* **225**, 1311–1326 (2019).
19. Corrales-Gutierrez, M. *et al.* The C4 protein from the geminivirus *Tomato yellow leaf curl virus* confers drought tolerance in Arabidopsis through an ABA-independent mechanism. *Plant Biotechnol. J.* **18**, 1121–1123 (2019).
20. Jing, C. *et al.* The Malvastrum yellow vein virus C4 protein promotes disease symptom development and enhances virus accumulation in plants. *Front. Microbiol.* **10**, 2425 (2019).
21. Jupin, I., Dekouchkovsky, F., Jouanneau, F. & Gronenborn, B. Movement of Tomato yellow leaf curl geminivirus (TYLCV)—Involvement of the protein encoded by ORF C4. *Virology* **204**, 82–90 (1994).
22. Rojas, M. R. *et al.* Functional analysis of proteins involved in movement of the monopartite Begomovirus, *Tomato yellow leaf curl virus*. *Virology* **291**, 110–125 (2001).
23. Tomás, D. M., Cañizares, M. C., Abad, J., Fernández-Muñoz, R. & Moriones, E. Resistance to *Tomato yellow leaf curl virus* accumulation in the tomato wild relative *Solanum habrochaites* associated with the C4 viral protein. *Mol. Plant-Microbe Interact.* **24**, 849–861 (2011).
24. Medina-Puche, L. *et al.* A defense pathway linking plasma membrane and chloroplasts and co-opted by pathogens. *Cell* **182**, 1109–1124 (2020).
25. Rosas-Diaz, T. *et al.* A virus-targeted plant receptor-like kinase promotes cell-to-cell spread of RNAi. *Proc. Natl. Acad. Sci. USA* **115**, 1388–1393 (2018).
26. Ismayil, A. *et al.* *Cotton leaf curl Multan virus* C4 protein suppresses both transcriptional and post-transcriptional gene silencing by interacting with SAM synthetase. *PLoS Pathog.* **14**, e1007282 (2018).
27. Kon, T., Sharma, P. & Ikegami, M. Suppressor of RNA silencing encoded by the monopartite tomato leaf curl Java begomovirus. *Arch. Virol.* **152**, 1273–1282 (2007).
28. Amin, I. *et al.* Suppressors of RNA silencing encoded by the components of the Cotton leaf curl begomovirus-betasatellite complex. *Mol. Plant-Microbe Interact.* **8**, 73–83 (2011).
29. Luna, A. P., Morilla, G., Voinnet, O. & Bejarano, E. R. Functional analysis of gene-silencing suppressors from Tomato yellow leaf curl disease viruses. *Mol. Plant-Microbe Interact.* **25**, 1294–1306 (2012).
30. Wang, B. *et al.* V2 of tomato yellow leaf curl virus can suppress methylation-mediated transcriptional gene silencing. *J. Gen. Virol.* **95**, 225–230 (2014).
31. Mei, Y., Yang, X., Huang, C., Zhang, X. & Zhou, X. Tomato leaf curl Yunnan virus-encoded C4 induces cell division through enhancing stability of Cyclin D 1.1 via impairing NbSKη-mediated phosphorylation in *Nicotiana benthamiana*. *PLoS Pathog.* **14**, e1006789 (2018).
32. Iqbal, Z., Shafiq, M., Ali, I., Mansoor, S. & Briddon, R. W. Maintenance of Cotton leaf curl Multan betasatellite by Tomato leaf curl New Dehli virus—Analysis by mutation. *Front. Plant Sci.* **8**, 2208 (2017).
33. Carluccio, A. V., Prigigallo, M. I., Rosas-Diaz, T., Lozano-Duran, R. & Stavolone, L. S-acylation mediates Mungbean yellow mosaic virus AC4 localization to the plasma membrane and in turns gene silencing suppression. *PLoS Pathog.* **14**, e1007207 (2018).

34. Fondong, V. N. *et al.* The consensus N-myristoylation motif of a geminivirus AC4 protein is required for membrane binding and pathogenicity. *Mol. Plant-Microbe Interact.* **20**, 380–391 (2007).
35. Vanitharan, R., Chellappan, P., Pita, J. S. & Fauquet, C. M. Differential roles of AC2 and AC4 of cassava geminiviruses in mediating synergism and suppression of posttranscriptional gene silencing. *J. Virol.* **78**, 9487–9498 (2004).
36. Chellappan, P., Vanitharani, R. & Fauquet, C. M. Short interfering RNA accumulation correlates with host recovery in DNA virus-infected hosts, and gene silencing targets specific viral sequences. *J. Virol.* **78**, 7465–7477 (2004).
37. Elmer, J. S. *et al.* Genetic analysis of the tomato golden mosaic virus. II. The product of the ALl coding sequence is required for replication. *Nucleic Acids Res.* **16**, 7043–7060 (1988).
38. Sung, Y. K. & Coutts, R. H. A. Mutational analysis of potato yellow mosaic geminivirus. *J. Gen. Virol.* **76**, 1773–1780 (1995).
39. Pooma, W. & Petty, I. T. D. Tomato golden mosaic virus open reading frame *AL4* is genetically distinct from its *C4* analogue in monopartite geminiviruses. *J. Gen. Virol.* **77**, 1947–1951 (1996).
40. Hoogstraten, R. A., Hanson, S. F. & Maxwell, D. P. Mutational analysis of the putative nicking motif in the replication-associated protein (AC1) of bean golden mosaic geminivirus. *Mol. Plant-Microbe Interact.* **7**, 594–599 (1996).
41. Bi, H., Fan, W. & Zhang, P. C4 Protein of Sweet potato leaf curl virus regulates brassinosteroid signaling pathway through interaction with AtBIN2 and affects male sterility in *Arabidopsis*. *Front. Plant Sci.* **8**, 1689 (2017).
42. Li, Z. *et al.* C4, the pathogenis determinant of *Tomato leaf curl Guangdong virus*, may suppress post-translational gene silencing by interacting with BAM1 protein. *Front. Microbiol.* **11**, 851 (2020).
43. Yahara, T., Ooi, K., Oshita, S., Ishii, I. & Ikegami, M. Molecular evolution of a host-range gene in geminiviruses infecting asexual populations of *Eupatorium makinoi*. *Genes Genet. Syst.* **73**, 137–141 (1998).
44. Ge, L., Zhang, J., Zhou, X. & Li, H. Genetic structure and population variability of Tomato yellow leaf curl China virus. *J. Virol.* **81**, 5902–5907 (2007).
45. Melgarejo, T. A. *et al.* Characterization of a New World monopartite begomovirus causing leaf curl disease of tomato in Ecuador and Peru reveal a new division in geminivirus evolution. *J. Virol.* **8**, 5397–5413 (2013).
46. Wyatt, S. D. & Brown, J. K. Detection of geminiviruses in aqueous extracts by polymerase chain reaction. *Phytopathology* **86**, 1288–1293 (1996).
47. Duffy, S. & Holmes, E. C. Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus *Tomato yellow leaf curl virus*. *J. Virol.* **82**, 957–965 (2008).
48. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
49. Mondal, D. *et al.* Genome wide molecular evolution analysis of begomoviruses reveals unique diversification pattern in coat protein gene of Old World and New World viruses. *VirusDis.* **30**, 74–83 (2019).
50. Yang, Z. & Kumar, S. Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. *Mol. Biol. Evol.* **13**, 650–659 (1996).
51. Buddenhagen, I. W. & de Ponti, O. M. B. Crop improvement to minimize future losses to disease and pests in the tropics. In *Breeding for Durable Disease and Pest Resistance*. 23–47 (FAO Plant Prod. Prot. Pap. 55. 1984).
52. Covarrubias, A. A., Romero-Perez, P. S., Cuevas-Velazquez, C. L. & Rendon-Luna, D. F. The functional diversity of structural disorder in plant proteins. *Arch. Biochem. Biophys.* **680**, 108229 (2020).
53. Mishra, P. M., Verma, N. C., Rao, C., Uversky, V. N. & Nandi, C. K. Intrinsically disordered proteins of viruses: Involvement in the mechanism of cell regulation and pathogenesis. *Prog. Mol. Biol. Transl. Sci.* **174**, 1–78. https://doi.org/10.1016/bs.pmbts.2020.03.001 (2020).
54. Cheng, S. & Brooks, C. L. Virus coat proteins are segregated in structural fold space. *PLoS Comp.* **9**, e1002905. https://doi.org/10.1371/journal.pcbi.1002905 (2013).
55. Hipp, K., Grimm, C., Jeske, H. & Böttcher, B. Near-atomic resolution structure of a plant geminivirus determined by electron cryomicroscopy. *Structure* **25**, 1303–1309 (2017).
56. Hesketh, E. L. *et al.* The 3.3 Å structure of a plant geminivirus using cryo-EM. *Nat. Commun.* **9**, 2369 (2018).
57. Carter, J. J. *et al.* Identification of an overprinting gene in Merkel cell polyomavirus provides evolutionary insight into the birth of viral genes. *Proc. Natl. Acad. Sci. USA.* **110**, 12744–12749 (2013).
58. Wright, P. E. & Dyson, H. J. Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293**, 321–331 (1999).
59. Wang, J. *et al.* Structural insights into the negative regulation of BRI1 signaling by BRI1-interacting protein BKI1. *Cell Res.* **24**, 1328–1341 (2014).
60. Jiang, J. *et al.* The intrinsically disordered protein BKI1 is essential for inhibiting BRI1 signaling in plants. *Mol. Plant* **8**, 1675–1678 (2015).
61. Obbard, D. J., Jiggins, F. M., Halligan, D. L. & Little, T. J. Natural selection drives extremely rapid evolution in antiviral RNAi genes. *Curr. Biol.* **16**, 580–585 (2006).
62. Obbard, D. J., Gordon, K. H. J., Buck, A. H. & Jiggins, F. M. The evolution of RNAi as a defense against viruses and transposable elements. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **364**, 99–115 (2009).
63. Murray, G. G. R., Kosakovsky, S. L. & Obbard, D. J. Suppressors of RNAi from plant viruses are subject to episodic positive selection. *Proc. R. Soc. B.* **280**, 20130965. https://doi.org/10.1098/rspb.2013.0965 (2013).
64. Lima, A. T. M. *et al.* The diversification of begomoviruse populations is predominantly driven by mutational dynamics. *Virus Evol.* **3**, vex005 (2017).
65. Hipp, K., Rau, P., Schafer, B., Pfannstiel, J. & Jeske, H. Translation, modification, and cellular distribution of two AC4 variants of African cassava mosaic virus in yeast and their pathogenic potential in plants. *Virology* **498**, 136–148 (2016).
66. Chen, K., Khatabi, B. & Fondong, V. N. The AC4 protein of a cassava geminivirus is required for virus infection. *Mol. Plant Microbe Interact.* **32**, 865–875 (2019).
67. Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
68. Kearse, M. *et al.* Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
69. Yang, Z., Wong, W. S. W. & Nielson, R. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* **22**, 1107–1118 (2005).
70. Hanson, J., Paliwal, K. K., Litfin, T. & Zhou, Y. SPOT-Disorder 2: Improved protein intrinsic disorder prediction by ensembled deep learning. *Genomics Proteomics Bioinform.* **17**, 645–656 (2019).
71. Katuwawala, A., Oldfield, C. J. & Kurgan, L. Accuracy of protein-level disorder predictions. *Brief. Bioinform.* **21**, 1509–1522 (2019).
72. Nielson, J. T. & Mulder, F. A. A. Quality and bias of protein disorder predictors. *Sci. Rep.* **9**, 5137 (2019).
73. Mizianty, M. J., Peng, Z. & Kurgan, L. A. MFDp2-accurate predictor of disorder in proteins by fusion of disordered probabilities, content and profiles. *Intrinsically Disord. Proteins* **1**, e24428 (2013).

Acknowledgements

Dr. Kisha Shelton for expertise in Photoshop.

Author contributions

C.M.D., P.S. and M.T.B. designed the study, performed the analyses and analyzed the data. C.M.D. and P.S. wrote the manuscript. All authors reviewed the manuscript.

Funding

This research was supported by funding from the Georgia Agricultural Experiment Station.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-90557-0.

Correspondence and requests for materials should be addressed to C.M.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.