Genome wide molecular evolution analysis of begomoviruses reveals unique diversification pattern in coat protein gene of Old World and New World viruses

Debayan Mondal¹ · Somnath Mandal¹ · Sandip Shil² · Nandita Sahana¹ · Goutam Kumar Pandit¹ · Ashok Choudhury³

Received: 26 December 2018 / Accepted: 15 March 2019 / Published online: 4 April 2019
© Indian Virological Society 2019

Abstract Begomoviruses (Family-*Geminiviridae*) are plant infecting single stranded DNA viruses known to evolve very fast. Here, we have analysed the DNA-A sequences of 302 begomoviruses reported as ‘type isolates’ from different countries following the list of International Committee on Taxonomy of Viruses till 2017. Phylogenetic analysis was performed which revealed two major evolutionarily distinct groups namely Old World (OW) and New World (NW) viruses. Our work present evidence that *cp* gene has varied degree of diversification among the viruses reported from NW and OW. The NW viruses are more conserved in their *cp* gene sequences than that of OW viruses irrespective of host plant families. Further analysis reveals that *cp* gene differs in its recombination pattern among OW and NW viruses whereas *rep* gene is highly recombination prone in both OW and NW viruses. The sequence conservation in *cp* gene in NW viruses is a result of meagre recombination and subsequent low substitution rate in comparison to OW viruses. Our results demonstrated that the *cp* gene in NW viruses is less likely to possess nuclear localisation sequences than OW *cp* gene.

Further we present evidence that the NW-*cp* is under the influence of strong purifying selection. We propose that the precoat protein (*pcp*) gene present exclusively in the 5’ of *cp* gene in OW viruses is highly diversified and strong positive selection working on *pcp* gene might be attributing largely to the diversity of OW-*cp* gene.

Keywords Begomoviruses · Coat protein · Evolution · NLS · Mean diversity

Introduction

Rapidly evolving plant viruses can cause extreme crop loss and affect socio-economic conditions adversely. Earlier studies have suggested that *geminiviruses* may evolve as quickly as the RNA viruses [11]. Among the seven genera of *geminiviruses*, begomovirus is the largest group and is considered as the most devastating plant virus infecting vegetables and legume crops. Major symptoms of the begomovirus infections are yellowing of the leaves, inward curling of leaves and stunting of the plants causing major yield loss. Begomoviruses have one or two small circular ssDNA molecules namely DNA-A and DNA-B [7, 18]. Typically, begomoviruses originating from New World (NW) are known to be bipartite having separately encapsidated DNA-A and DNA-B whereas the viruses from Old World (OW) are monopartite in majority and comprised either only DNA-A, or DNA-A occasionally associated with alpha or betasatellite molecules [5]. About 15% of OW viruses are reported to be bipartite [15]. Recently some monopartite viruses have also been reported from NW [32, 48, 53]. DNA-A is the most conserved genomic component of begomovirus and harbours the necessary genes for the replication, movement and infection process

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s13337-019-00524-7>) contains supplementary material, which is available to authorized users.

✉ Nandita Sahana
nanditasahana@gmail.com

¹ Department of Biochemistry, Uttar Banga Krishi Viswavidyalaya, Coochbehar, West Bengal, India

² Regional Research Centre, ICAR-CPCRI, Mohitnagar, Jalpaiguri, West Bengal 735102, India

³ Soil Microbiology Laboratory, Regional Research Station, Terai Zone, Uttar Banga Krishi Viswavidyalaya, Coochbehar, West Bengal, India

of the virus [6]. DNA-A encodes altogether six proteins; the virion sense strand of virus encodes coat protein (by *cp* gene or *v1* gene), required for encapsidation and movement in plants [16] and pre-coat protein (*pcp* gene or *v2* gene), a pathogenicity determinant involved in movement of virus in plants [47]. These two genes overlap with each other at the 5' of *cp* and 3' of *pcp* sharing about two hundred nucleotides. The *cp* and *pcp* genes produce two separate proteins comprising about three hundred fifty and one hundred twenty amino acids under different coding frame. The complementary sense strand encodes Replication associated protein (*rep* gene or *c1* gene) required for viral replication [35]. These three genes take away more than 70% of the DNA-A. The other genes on complementary strand, *c2*, *c3* and *c4* serve as the suppressors of RNA silencing mechanism [59], establish an environment conducive for virus replication [40], contribute towards pathogenicity determination and suppress the host PTGS pathway by binding to siRNAs [38] respectively.

Coat protein is a multifunctional protein which contributes to formation of shelling of the geminate particle, transmission of the virus through whitefly, self-binding, systemic movement of the virus in the plant, nuclear targeting and nuclear export of the viral genome [19]. The structural and functional roles displayed by the *cp* gene are thought to be the reasons behind its remarkable sequence conservation [19, 37]. On the other hand, begomoviral *rep* gene performs multiple functions like cleavage or ligation of viral DNA, oligomerization of domains of the protein [20, 22], provide binding sites for viral replication enhancer protein (REn) [55] and host factors [1, 9, 14]. Unlike *cp* gene, the increased variability of *rep* gene both in the nucleotide and protein level has been observed instead of its multifunctional role. Earlier studies suggested that the evolutions of begomoviruses are the result of frequent homologous and heterologous recombination and substitution mutations [43, 45]. But the impact of these events on the genomic components of NW and OW viruses with functional implication has not been addressed till date.

Recently small interfering RNA and transacting small interfering RNA based strategies have gained significant importance worldwide to develop resistance against these viruses [24, 41, 42, 56]. But rapid mutation rate and subsequent fast evolution of begomoviruses impose difficulty in strategizing Pathogen Derived Resistance (PDR). The only successful example of RNAi based resistance is against *Bean Golden Mosaic Virus* of common beans [2] which has been attributed to the low genetic diversity of the virus in that area. Identifying evolutionarily stable sequences of the virus is an important prerequisite to build RNAi based strategies. In the present study we have compared two major genomic component i.e. *cp* and *rep* genes in all the 'type isolates' of begomoviruses reported

from different countries. We present evidence that the evolutionary pattern of the *cp* gene from the OW and NW phylo groups were distinct and NW viruses were more conserved in their *cp* gene sequences. The OW viruses were diverse in their *cp* and *rep* genes to the same extent whereas the NW viruses have significant difference in diversity of these genes.

Materials and methods

Multiple sequence alignments and phylogenetic analysis

Three hundred and two Begomovirus 'type isolates' enlisted in ICTV till 2017 were considered for the analysis. We collected information regarding their reporting year, country and host plants (Supplementary material 1). The abbreviated names of the viruses are given following ICTV codes. The full sequences of DNA-A along with *cp*, *rep* and *pcp* genes of all the viruses were retrieved separately from the National Centre for Biotechnology Information (NCBI) GenBank Database (<https://www.ncbi.nlm.nih.gov/>).

Multiple sequence alignment was performed by Clustal Omega and the phylogenetic tree was constructed through *Molecular Evolutionary Genetics Analysis tool MEGA X* [26]. The evolutionary history was inferred using the Neighbor-Joining method. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. *Chicken anemia virus* (M55918.1) from the genus *Circovirus*, family *Circoviridae* was taken as an outgroup virus for the phylogenetic analysis. The mean evolutionary diversity (π) for the entire population was calculated by the equation

$$a_i = \frac{q}{q-1} \sum_{i=1}^q \sum_{j=1}^q x_i x_j d_{ij}$$

In MEGA X where x_i is the estimate of average frequency of the i th allele in the entire population, and q is the number of different sequences in the entire sample. All the calculated mean diversity was found statistically different using F test.

Pair-wise sequence diversity analysis

Sequence diversity of the *cp*, *rep*, *pcp* and *ac4* genes acquired from NCBI Database was analysed with Sequence Demarcation Tool (SDT v 1.2) (<http://web.cbio.uct.ac.za/~brejnev/>) following Muhire et al. [33]. Subsequently,

heat map was generated for each datasets with sequence having more than 60% homogeneity.

Recombination analysis

Identification of potential recombination events, contributing major and minor parental sequences and possible recombination breakpoint positions of the entire DNA-A as well as the *cp* (*av1*), *rep* (*ac1*), *pcp* (*av2*) and *ac4* genes from OW and NW viruses were carried out using the RDP, GENECONV, Bootscan (RecScan), MaxChi, Chimaera, SisScan, PhylPro, VisRD, LARD methods as implemented in Recombination Detection Program v.4.95 following Martin et al. [31]. The Bonferroni-adjustment (with p values cut off of 0.05) was considered to deal with the problem of multiplicity. The recombination events that were detected by at least three programmes were considered as significant events. To generate the break point distribution plot population scaled recombination rates (ρ) were calculated from recombination rate plot using CONVERT and INTERVAL programmes of LdHat package in the RDP version 4. The Minor Allele Frequency (MAF) cut off was kept 0.05 in the programme. Once the ρ value was generated for each data set then the break point distribution plot was obtained using the ρ value as an input to normalise the sequence data sets.

Detection of positive and negative selection at amino acid sites

dN/dS values were calculated based on the log ratio of the rate of non-synonymous substitutions to the rate of synonymous (silent) substitutions following the method described by Ho et al. [21], to identify whether the genes are under negative or positive selection. All related nucleic acid sequences were fetched, codons were generated and aligned in T-COFFEE software (tcoffee.crg.cat), and codon alignments were further submitted to Single Likelihood Ancestor Counting (SLAC), hosted by DataMonkey web server (www.datamonkey.org) [23, 34]. Negatively and positively selected sites in the *cp* and *rep* datasets were identified using three distinct methods (viz. SLAC, PARIS and FUBER). The dN/dS values were calculated from the downloaded csv format.

Nuclear localisation signal sequence identification

The Nuclear Localisation Signal sequences were identified and analysed using two different softwares. First we used cNLS Mapper tool (http://nls-mapper.iab.keio.ac.jp/cgi-bin/NLS_Mapper_form.cgi), where the cut off e-score was fixed for 4.00 during the initial run and 8.00 for increasing specificity of the prediction in two consecutive runs. The

same protein sequences were also submitted to Nuc-Pred [4] to generate nuc-pred scores. The higher scores indicate more probability of a given sequence to actually act as a nuclear localisation signal. The identified NLS regions were further analysed in WebLogo tool. Each logo consists of stacks of symbols, one stack for each position in the sequence. The overall height of the stack indicates the sequence conservation at that position, while the height of symbols within the stack indicates the relative frequency of each amino acid at that position.

Result and discussion

Differential rate of nucleotide diversity in *cp* and *rep* genes among members of OW and NW viruses

Phylogenetic analysis of three hundred and two begomoviruses was performed with the DNA-A sequences reported from different parts of the world (Supplementary material 1). The phylogenetic tree revealed that the evolutionary clusters could be grouped based on the geographical distribution of the virus (supplementary material 2). Viral sequences reported from same continents were found to be clustered together. The first major cluster constituted of one hundred and fifteen viruses from Southeast Asian countries whereas one hundred and nineteen viruses from Latin American, USA, and Caribbean countries formed the second major cluster (supplementary material 2, Supplementary material 3a and 3b). The Southeast Asian viruses were found to possess a longer DNA-A molecule (Avg. 2749 bases) than the Latin American viruses (Avg. 2624 bases). In the South Asian cluster 88 monopartite and 26 bipartite viruses were found suggesting that the monopartite viruses are predominant in this part of the world. The Latin American cluster was dominated by bipartite viruses. Except these major clusters, two small clusters consisted of sweepoviruses [30] and legumoviruses have been reported based on their host specificity [30, 44, 46]. Based on this analysis the present viruses were grouped into well accepted NW and OW viruses [6] for further analysis.

The nucleotide diversity of the *cp* and *rep* genes in both OW and NW viruses have been calculated. The diversity of the *cp* and *rep* gene in OW viruses were found to be similar but in NW viruses significant difference were observed in sequence diversity as evident in the obtained matrices (Fig. 1). For *rep* gene in both NW and OW viruses the predominance of blue colour in matrix suggests that *rep* gene is highly diversified during evolution. For OW viruses the range of pair wise identity in *rep* gene was found to be 61–92% with no sharp peak, whereas in NW-*rep* gene a blunt curve was observed from 60 to 90% (Supplementary

Fig. 1 Sequence diversity analysis of **a** Old World *cp* **b** Old World *rep* **c** New World *cp* and **d** New World *rep* genes using Sequence Demarcation Tool version 1.2 (SDTv1.2). Pair wise identity is shown from blue (minimum) to red (maximum) matrix which represents

percentage pair wise genome scores in nucleotide identity plot taking percentage pair wise identity in X-axis and proportion of pair wise identity in Y-axis. The name of viruses taken for the analysis is given in the supplementary material 1a and 1b (color figure online)

material 3c). The calculated mean diversity (π) of *rep* genes of OW and NW were found to be 0.287 and 0.344 respectively (Table 1). However, a discrete difference in pair wise identity matrix was observed in the case of *cp* gene. The OW-*cp* gene was found to be more diverse than the NW viruses. In the NW viruses *cp* gene showed a very sharp peak ranging from 73 to 93% whereas in the OW a blunt curve was found ranging from 61 to 93% (Supplementary material 3c). The calculated mean diversity for *cp* gene in the OW and NW viruses were 0.293, 0.208 respectively (Table 1).

Further the datasets were divided according to the host plant of the virus into small groups to analyse whether the diversity in these genes are specific to viruses infecting

particular family of plant. It was observed that except for viruses infecting Cucurbitaceae and Malvaceae all other families have a higher mean diversity of NW-*rep* gene than in the OW-*rep* gene (Table 1). In the case of *cp* gene sequences all the family-groups of viruses showed a higher mean diversity in OW compared to the NW. These results may suggest that the diversification of *cp* and *rep* genes in NW viruses were not parallel during evolution unlike OW viruses and the *cp* gene is much conserved across the bipartite viral species found in NW compared to viruses of OW irrespective of any particular host plant family.

The *cp* gene is reported to be the most conserved gene in the DNA-A of begomovirus [10, 28]. But a comparative diversity analysis of the *cp* genes from different

Table 1 Calculated mean nucleotide diversity (π) of Old World and New World *cp* and *rep* genes

Host plant family	<i>cp</i>		<i>rep</i>	
	OW	NW	OW	NW
All	0.2933 ^a	0.2082 ^b	0.2870 ^a	0.3449 ^b
Solanaceae	0.2882 ^a	0.2107 ^b	0.2786 ^a	0.3188 ^b
Cucurbitaceae	0.1603	0.1641	0.2867	0.2145
Euphorbiaceae	0.2997	0.1934	0.3098	0.3698
Malvaceae	0.2609	0.1716	0.2335	0.1349
Asteraceae	0.2804	–	0.2190	–
Fabaceae	0.2859	0.2156	0.2424	0.3559
Piperaceae	0.3188	0.2359	0.3152	0.5599

Different letters in superscript indicate statistically significant differences in the mean diversity values (by *F*-test with $p < 0.01$)

geographical origins are limited. Indication of *cp* gene specificity of the viruses from different geographical location has been given in the study by Harrison et al. [19] where it has been claimed that monoclonal antibodies reacted only with the viruses of the same continent, and gave negative results in viruses of different geographical location. Considering the fact that a very high rate of molecular evolution has been observed in the begomoviruses [11] it is of prime interest that viruses of NW were able to remain conserved at *cp* gene whereas the OW viruses have allowed significant diversity at the same gene.

Evidence of meagre recombination in NW *cp* gene compared to its OW counterpart

To analyse the plausible reason behind the uneven diversification of the *cp* and *rep* genes of OW and NW viruses; the recombination potential was analysed for the entire DNA-A sequence. The DNA-A of OW viruses showed a high number of significant recombination events (185 events) as compared to NW viruses (128 events) (Supplementary file 4a and 4b). The break point distribution plot indicates that in both OW and NW viruses a globally significant hot spot is present in *rep* gene spanning around 200 bp at the central position of the gene whereas a locally significant hotspot was detected in OW viruses at the end position of *cp* gene spanning around 100 bp which was absent in NW viruses (Supplementary file 4c). Further to study the differences in recombination pattern of the genes, *cp* and *rep* sequences were analysed separately. Well supported recombination events were high in *rep* genes (51 and 55 respectively) and comparatively low for *cp* genes (5 and 30 respectively) (Table 2, and Supplementary material 5a–e). Population scaled recombination rate (ρ) was

calculated for all datasets and was found to be in the same range for all the sequence data sets except for NW-*cp* gene. The recombination break point distribution plot was generated using the calculated ρ value as input to minimise the error. The recombination break point plot identified three globally significant recombination hotspots in the *rep* gene of both NW and OW viruses at both the ends of the gene and in the middle region (Fig. 2). The recombination breakpoint distribution plot for *cp* gene of OW viruses showed globally significant recombination hotspots in the 5' terminal ends around nucleotide position 194 spanning 80–100 nucleotides which is almost absent in the NW viruses (Fig. 2). This data suggests that the recombination pattern for *rep* gene is similar in case of NW and OW viruses but for *cp* genes an additional hotspot is present in OW viruses that may increase the total recombination events and subsequently the diversity of OW-*cp* gene.

The viral rolling cycle replication process uses recombination for replicating complementary strand which makes the genes on complementary strand, like *rep* more vulnerable to recombination [28]. For *cp* gene, the recombination events were found to be low in comparison to *rep* gene from viruses of both the worlds. Recombination hotspot at the end of *cp* gene and at the junction point of *cp-ac3* has been reported earlier in begomoviruses [28]. The present data suggest that this hotspot may exclusively present in OW viruses but absent in NW viruses. Another difference that has been noticed is the presence of an extra hotspot in the OW-*cp* gene at a position where 5' terminal of *cp* and 3' terminal of *pcp* gene overlaps (Fig. 2).

NLS sequence in NW *cp* gene

One of the major functions performed by begomoviral coat protein is nuclear targeting and nuclear export of the viral genome in the host plant cells which is crucial for the viral replication [19]. We tried to evaluate the functional implication of the difference observed in *cp* sequence of NW and OW viruses by investigating the presence of NLS sequences by two independent methods. In the first experiment Nuc-Pred score of the coat proteins were generated individually and the mean score were calculated (Table 3). The mean Nuc-Pred score was found significantly different between NW and OW coat protein. The Nuc-Pred score of NW coat protein was very low (0.15) compared to OW coat proteins (0.33). We repeated the analysis for individual groups of viruses infecting different host plant families and found that irrespective of the host plant families the score was lower in NW viruses than in OW viruses.

In the second method, we analysed the coat protein sequences with NLS mapper for the possible presence of NLS (Supplementary material 6a, 6b). We could detect

Table 2 Calculated mean nucleotide diversity (π), population scaled recombination rate (ρ), Waterson theta for mutation rates (θ), ρ/θ and non-synonymous to synonymous substitution rate (dN/dS) for Old World and New World *cp* and *rep* genes

	π	ρ	θ	ρ/θ	Recombination events	dN/dS	Total number of amino acid sites under positive selection
OW <i>cp</i>	0.293	3.696	0.04642	79.604	30	0.478	4
NW <i>cp</i>	0.2102	10.77	0.04501	87.634	5	0.369	0
OW <i>rep</i>	0.2866	3.582	0.04087	239.34	55	0.611	10
NW <i>rep</i>	0.344	3.786	0.01943	195.053	51	0.705	16

Different letters in superscript indicate statistically significant differences in the mean diversity values (by *F*-test with $p < 0.01$)

Fig. 2 The break point distribution plot of **a** Old World *cp*, **b** New World *cp*, **c** Old World *rep*, **d** New World *rep* genes. All estimated breakpoint positions are indicated by small vertical lines at the top of each plot. 100 nucleotide windows were moved along the sequence alignment and number of breakpoints detected within the region was

plotted. The horizontal lines at the top of the graph indicate 99% and 95% confidence threshold for globally significant breakpoint regions. The light and dark grey areas respectively indicate 99% and 95% confidence threshold for locally significant breakpoint regions. Red arrows indicate breakpoint hotspots (color figure online)

Table 3 Mean values of NucPred Scores based on nuclear localization signals (NLS) sequences found in Old World and New World coat proteins

Group name	OW	NW
All	0.334 ^a	0.150 ^b
Solanaceae	0.335 ^a	0.152 ^b
Cucurbitaceae	0.230	0.118
Euphorbiaceae	0.341	0.120
Malvaceae	0.313	0.167
Asteraceae	0.383	0.120
Fabaceae	0.320	0.146
Piperaceae	0.326	0.105

large number of NLS sequences in both OW and NW coat proteins where the e-score was kept < 4.00 , but with a higher specificity (e-score > 8.00) no NLS sequences were detected in NW coat protein. In OW viruses the NLS sequence were prevalent in N-terminal of the protein. The N-terminal NLS with high specificity (e-score > 9) starts with amino acid number one and is around twenty amino acid long. We generated web logo from the N-terminal of the protein sequences spanning the NLS (Fig. 3) and could detect significant difference in the basic R and K residues. In NW viruses most of the basic residues were found to be absent in the generated logo. The results suggest that probably the NLS sequences at N-terminal are gradually disappearing from NW coat proteins and the intracellular

Fig. 3 Comparison of NLS sequence of Old World and New World coat proteins. The predicted NLS sequences were collected from each coat protein and aligned. The motifs were generated in weblogo software (<https://weblogo.berkeley.edu/logo.cgi>). The location in alignments is given in the X axis. The height of the residue signifies the conservation of the residue in the dataset

movement of the NW viruses to nucleus is possibly not governed by the coat proteins any more.

Earlier report emphasized on presence of NLS sequences in N-terminal, C-terminal and Middle region of coat protein from different geminiviruses [3, 13, 25, 27, 57, 58]. The NLS motif in coat protein is demonstrated to help in nuclear targeting of the OW viruses like MYMV [13], TYLCV [25], and ACMV [58]. Ha et al. [15] have reported that the jute infection CoGMV and CoYVV lacks the second and third basic domain in the N-terminal region of coat protein like some other NW viruses. Earlier report opined that DNA-B genomic components of the bipartite NW virus are originated from DNA-A components followed by gain of some new functions [6, 47]. Our data suggests that during evolution NW *cp* gene might have lost its function as a nuclear movement protein. This function is probably performed by Nuclear Shuttle Protein (NSP) encoded by DNA-B of NW bipartite virus [6, 47, 57] in the present day which has been reported of having two NLS sequences [39, 54] and presumed to be a modified *cp* gene [6].

The NW *cp* genes under purifying selection

To investigate whether the genes are under positive or negative selection the dN/dS values were derived. The OW and NW-*rep* genes showed a higher dN/dS value of 0.61 and 0.70 respectively. The number of detected amino acid sites under positive selection for the *rep* genes was ten and sixteen respectively (Table 2). On the other hand the NW-

cp gene showed lower dN/dS value (0.37) followed by OW-*cp* (0.48) (Table 2). There was no positively selected site found in NW-*cp* gene whereas in OW viruses four positively selected sites were detected for *cp* gene. Taken together these data may suggest that the *rep* genes are majorly under the influence of positive selection and still diversifying. The values also justify the higher mean diversity level of begomoviral *rep* genes from both the worlds. The lower dN/dS values for *cp* genes may suggest that in both worlds *cp* genes are under high negative or purifying selection as reported earlier in different strain of begomoviruses [11, 29, 37] but NW-*cp* genes are under stronger purifying selection pressure during evolution compared to OW viruses.

The population scaled recombination rate (ρ) and contribution of recombination in mutation per site (ρ/θ) for OW and NW datasets were also calculated (Table 2). The ρ/θ rates were found to be higher in NW viruses and lower in OW viruses for both the genes which may be due to high rate of pseudo recombination in NW genomic components. But the detected actual recombination events in NW-*cp* gene were found to be very low. The number of recombination events showed positive but non-significant (0.805) correlation with mean sequence diversity (π) of the genes. But the mean sequence diversity (π) and the dN/dS values showed a significant positive correlation (0.90). Therefore the sequence diversity of the NW and OW viruses seems to be attributed by recombination as well as substitution but may be more influenced by the substitution rather than recombination.

Overlapping of *pcp* gene in OW viruses may be crucial for *cp* gene diversity

Since the *cp* and *rep* genes have overlapping sequences of *pcp* and *ac4* genes respectively it was pertinent to analyse the impact of these gene sequences on overall diversity of *cp* and *rep* gene. The mean sequence diversity of *pcp* and *ac4* genes (Supplementary material 7) were analysed. The mean diversity of OW-*ac4* gene was found to be similar (0.2883) with OW-*rep* gene, but the NW-*ac4* showed a higher mean diversity (0.4037) than NW-*rep* gene (Supplementary material 7). We speculate that this high rate of sequence diversity may attribute towards *rep* gene diversification in NW viruses. An extra recombination hotspot in the OW-*cp* genes was observed which was absent in NW-*cp* genes (Fig. 2). The *pcp* gene is present in the virion strand of only OW viruses known to partially share the DNA sequence (190–200 bases) with *cp* gene by an overlapping at 3' of *pcp* and 5' of *cp* genes. Mean genetic diversity of the *pcp* sequence (0.334) suggested that *pcp* genes are highly diversified (Supplementary material 7). The total number of well supported recombination events,

and the ρ/θ value of the *pcp* gene sequence suggested that *pcp* sequence is a recombination prone sequence (Table 2 and Supplementary material 8). Moreover the dN/dS value for the *pcp* sequence was found to be near unity (0.998) which suggests that the *pcp* gene is under strong positive selection pressure and still diversifying (Table 2). With only 120 amino acids long protein we found seventeen amino acid sites that are under diversifying selection. When the difference between the non-synonymous to synonymous substitution rates across the codons of the pre-coat protein was plotted, it was observed that the rate of non-synonymous substitution is higher in the C-terminal of the protein whereas the synonymous substitutions are prevalent in the N-terminal of the protein (Supplementary material 9). Interestingly, the codons at the C-terminal of the protein are constituted of the sequences that are shared by both *pcp* and *cp* genes.

The diversity in the *pcp* gene and the strong positive selection working at the C-terminal of the pre coat protein seems to have a destabilising effect on the nucleotide sequence that are shared by both the *pcp* and *cp* genes in the OW viruses which may significantly contribute to the diversity of *cp* gene in the OW viruses. Due to the inactivation of *pcp* gene in the NW viruses by deletion of its promoter region [21], this instability in the nucleotide sequences has been released, which may act as a purifying selection at the *cp* gene of NW viruses and may contribute towards the conservation of the NW-*cp* gene than that of the OW counterparts during evolution.

The study of the evolutionary history of OW and NW begomoviruses are important to understand the strategies applied by the virus to adapt new biochemical environment and shifting towards new host. In the present study we found that the sequence diversification of the genes residing in DNA-A component of NW viruses is uneven. As a result the *rep* genes in NW viruses have diversified immensely whereas the *cp* genes remained conserved. But in the case of OW viruses the sequence diversification of *cp* and *rep* gene has been found to be similar. Previous studies have reported that *cp* is conserved among geminiviruses, different begomoviruses and in the Indian strain of ToLCV [11, 19, 25, 28, 29]. In the present study comparative diversity analysis of the genes in OW and NW viruses has been investigated. We observed the stringent nature of the *cp* genes in NW viruses in all host plant families. We present evidence that the NW-*cp* genes are under strong purifying selection whereas *rep* gene from the viruses of both the world are under diversifying selection. Both recombination and substitution were found to be comparatively less in NW-*cp* genes than OW-*cp* genes whereas *rep* gene was found to be highly recombination prone in both cases. The recombination hotspot at the junction of *cp-ac3* was exclusively present in OW viruses.

Our analysis with the *pcp* genes of all OW viruses suggest that *pcp* gene is under strong diversifying selection and has high sequence variability. We propose that the presence of diversified *pcp* gene might play an important role in overall sequence variability of OW *cp* gene. The inactivation of the *pcp* gene in the NW viruses might be the cause for the stability of the nucleotide sequences in the NW *cp* genes during evolution.

Begomoviral epidemics have been seen as a result of favourable conjunction of plant, vector and virus creating new virus variants especially by recombination. The recombination may occur either in the host plants during mixed infection or in the vector whitefly. As the coat protein is crucial for transmission of the virus through whitefly the variation in this gene directly impact the adaptation of the virus to new host [17]. The NW viruses infect around twelve families of host species and in the OW the begomoviruses infects eighteen families of plants (supplementary material 3a and 3b). Evolutionary pressure of adjusting to new host is lower in NW viruses because of fewer families of host plants which may be one of the reasons for less diversification of the gene.

The importance of different conserved motif and sequences for establishment of viral infection in host plants has been discussed in several studies [51, 52]. In begomovirus the conserved NLS sequences are the key for shuttling of the viral DNA to nucleus and thus facilitate further replication of the virus. Our analysis suggests the NW-*cp* genes are very unlikely to possess NLS sequences. Earlier reports have suggested that the nuclear shuttling of the NW virus is performed by the DNA-B located *nsp* genes [6, 39, 47, 57]. Results of this study suggest that the NW *cp* gene may have lost some of its ancient roles. The impairment of *cp* genes to perform the nuclear movement of the virus due to absence of the NLS sequences may explain the inability of DNA-A component of NW viruses to establish a systemic infection [8, 12].

Monopartite and Bipartite begomoviruses were originated in the OW, where the diversity of the viruses is found to be more [36]. Monopartite viruses prevailed in OW with small genome comprised of only DNA-A seldom with alpha or betasatellite molecules. The functional roles of these satellite molecules are not well deciphered. Although one protein $\beta c1$ from betasatellite has been reported to have role in viral movement of OW viruses [50] but the DNA-A alone is capable of establishing systemic infections in those viruses. Bipartite viruses migrated to NW in recent past and the DNA-B component which is probably derived from original DNA-A [6, 47, 49], has gained functions like cell to cell movement and nuclear shuttling of the virus. We propose that the strengthening of DNA-B molecule with these functions may have relieved the NW-*cp* gene from certain roles it used to perform earlier

whereas in OW viruses the *cp* gene has continued to perform all these roles. The presence of *pcp* gene imposes a strong diversifying selection resulting a higher variability of the gene. The inactivation of the *pcp* gene by promoter deletion, lower pressure of new host adaptability and the loss of some ancient functions has allowed NW-*cp* gene to evolve under strong negative selection and thus has developed stringency in their nucleotide sequences.

Acknowledgements The authors would like to acknowledge Vice Chancellor of UBKV Dr. Chirantan Chattopadhyay, for his critical inputs during the course of study. We also acknowledge Dr. Sankalpa Ojha, Department of Agricultural Statistics; UBKV for helping us in the statistical analysis. Debayan Mondal has received University Merit Scholarship throughout his Master's study. The funding is exclusively sponsored by Institutional Project under Directorate of Research, UBKV (Project Code- UBKV/DR-191;04.26).

References

- Ach RA, Durfee T, Miller AB, Taranto P, Hanley-Bowdoin L, Zambriski PC, Gruissem W. RRB1 and RRB2 encode maize retinoblastoma-related proteins that interact with a plant D-type cyclin and geminivirus replication protein. *Mol Cell Biol*. 1997;17:5077–86.
- Aragao FJ, Faria JC. First transgenic geminivirus-resistant plant in the field. *Nat Biotechnol*. 2009;27(12):1086–8.
- Botcher B, Unseld S, Ceulemans H, Russell RB, Jeske H. Geminivirus structures of African cassava mosaic virus. *J Virol*. 2004;78(13):6758–65.
- Brameier M, Krings A, MacCallum RM. NucPred—predicting nuclear localization of proteins. *BMC Bioinform*. 2007;23(9):1159–60.
- Briddon RW, Stanley J. Sub-viral agents associated with plant-infecting single-stranded DNA viruses. *Virology*. 2006;344:198–210.
- Briddon RW, Patil BL, Bagewadi B, Nawaz-ul-Rehman MS, Fauquet CM. Distinct evolutionary histories of the DNA-A and DNA-B components of bipartite begomoviruses. *BMC Evol Biol*. 2010;10:97.
- Brown JK, Fauquet CM, Briddon RW, Zerbini FM, Moriones E, Navas-Castillo J. Family *Geminiviridae*. In: *Virus taxonomy 9th report of the international committee on taxonomy of viruses*. 2012. p. 351–73.
- Carvalho MF, Lazarowitz SG. Interaction of the movement protein NSP and the *Arabidopsis* Acetyltransferase AtNSI is necessary for cabbage leaf curl geminivirus infection and pathogenicity. *J Virol*. 2004;78(20):11161–71.
- Castillo AG. Interaction between a geminivirus replication protein and the plant sumoylation system. *J Virol*. 2004;78:2758–69.
- De Bruyn A, Harimalala M, Zinga I, Mabvakure BM, Hoareau M, Ravigné V, Walters M, Reynaud B, Varsani A, Harkins GW, Martin DP, Lett JM, Lefeuvre P. Divergent evolutionary and epidemiological dynamics of cassava mosaic *Geminiviruses* in Madagascar. *BMC Evol Biol*. 2016;16:182.
- Duffy S, Holmes EC. Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus tomato yellow leaf curl virus. *J Virol*. 2007;82(2):957–65.
- Gardiner WE, Sunter G, Bisaro DM. Identification of tomato golden mosaic virus-specific RNAs in infected plants. *J Virol*. 1989;170:243–50.
- Guerra Peraza O, Kirk D, Seltzer V, Veluthambi K, Scmit AC, Hohn T, Herzog E. Coat proteins of Rice tungro bacilliform virus and Mungbean yellow mosaic virus contain multiple nuclear-localization signals and interact with importin α . *J Gen Virol*. 2005;86:1815–26.
- Gutierrez C. DNA replication and cell cycle in plants: learning from *Geminiviruses*. *EMBO J*. 2000;19:792–9.
- Ha C, Coombs S, Revill P, Harding R, Vu M, Dale J. Molecular characterization of begomoviruses and DNA satellites from Vietnam: additional evidence that the New World *Geminiviruses* were present in the Old World prior to continental separation. *J Gen Virol*. 2008;89:312–26.
- Hanley-Bowdoin L, Settledge SB, Orozco BM, Nagar S, Robertson D. *Geminiviruses*: models for plant DNA replication, transcription and cell cycle regulation. *Crit Rev Biochem Mol Biol*. 1999;18:71–106.
- Harrison BD, Robinson DJ. Natural genomic and antigenic variation in whitefly-transmitted *Geminiviruses* (begomoviruses). *Annu Rev Phytopathol*. 1999;37:369–98.
- Harrison BD, Baker H, Bock KR, Guthrie EJ, Meredith G, Atkinson M. Plant viruses with circular single-stranded DNA. *Nature*. 1977;270:760–2.
- Harrison BD, Swanson MM, Fargette D. Begomovirus coat protein: serology, variation and functions. *Physiol Mol Plant Pathol*. 2002;60(5):257–71.
- Heyraud-Nitschke F, Schumacher S, Laufs J, Schaefer S, Schell J, Gronenborn B. Determination of the origin cleavage and joining domain of geminivirus rep proteins. *Nucl Acids Res*. 1995;23:910–6.
- Ho ES, Kuchie J, Duffy S. Bioinformatic analysis reveals genome size reduction and the emergence of tyrosine phosphorylation site in the movement protein of New World bipartite begomoviruses. *PLoS ONE*. 2014;9(11):e111957.
- Horvath MP, Schweiker VL, Bevilacqua JM, Ruggles JA, Schultz SC. Crystal structure of the *Oxytricha nova* telomere end binding protein complexed with single strand DNA. *Cell*. 1998;95:963–74.
- Kosakovsky Pond SL, Frost SD. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics*. 2005;21(10):2531–3.
- Kumar A, Sarin NB. RNAi: a promising approach to develop transgenic plants against *Geminiviruses* and insects. *J Plant Physiol Pathol*. 2013;1:1.
- Kumar SP, Patel SK, Kapopara RV, Jasrai YT, Pandya HA. Evolutionary and molecular aspects of indian tomato leaf curl virus coat protein. *Int J Plant Genomics*. 2012. <https://doi.org/10.1155/2012/417935/>.
- Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol*. 2018;10:1093.
- Kunik T, Salamon D, Zamir D, Navot N, Zeidan M, Michelson I, Gafni Y, Czosnek H. Transgenic tomato plants expressing the tomato yellow curl virus capsid protein are resistant to the virus. *Biotechnology (NY)*. 1994;12:500–4.
- Lefeuvre P, Harkins GW, Lett JM, Briddon RW, Chase MW, Moury B, Martin DP. Evolutionary time-scale of the begomoviruses: evidence from integrated sequences in the Nicotiana genome. *PLoS ONE*. 2011;6(5):e19193.
- Lima ATM, Silva JCF, Silva FN, Castillo-Urquiza GP, Silva FF, Seah YM, Mizubuti ESG, Duffy S, Zerbini FM. The diversification of begomovirus populations is predominantly driven by mutational dynamics. *J Virus Evol*. 2017;3(1):vex005.
- Lozano G, Trenado HP, Valverde RA, Navas-Castillo J. Novel begomovirus species of recombinant nature in sweet potato (*Ipomoea batatas*) and *Ipomoea indica*: taxonomic and phylogenetic implications. *J Gen Virol*. 2009;90:2550–62.

31. Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. RDP4: detection and analysis of recombination patterns in virus genomes. *Virus Evol.* 2015;1(1):1–5.
32. Melgarejo TA, Kon T, Rojas MR, Paz-Carrasco L, Zerbini FM, Gilbertson RL. Characterization of a New World monopartite begomovirus causing leaf curl disease of tomato in Ecuador and Peru reveals a new direction in geminivirus evolution. *Virology.* 2013;87:5397–413.
33. Muhire BM, Varsani A, Martin DP. SDT: a virus classification tool based on pairwise sequence alignment and identity calculation. *PLoS ONE.* 2014;9(9):e108277.
34. Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Pond SLK, Scheffler K. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Mol Biol Evol.* 2013;30(5):1196–205.
35. Nash TE, Dallas MB, Reyes MI, Buhrman GK, Ascencio-Ibañez JT, Hanley-Bowdoin L. Functional analysis of a novel motif conserved across *geminiviruses* Rep proteins. *J Virol.* 2011;85:1182–92.
36. Nawaz-ul-Rehman MS, Fauquet CM. Evolution of *geminiviruses* and their satellites. *FEBS Lett.* 2009;583:1825–32.
37. Padidam M, Beachy RN, Fauquet CM. Tomato leaf curl geminivirus from India has a bipartite genome and coat protein is not essential for infectivity. *J Gen Virol.* 1995;76:25–35.
38. Pandey SP, Shahi P, Gase K, Baldwin IT. Herbivory-induced changes in the small-RNA transcriptome and phytohormone signaling in *Nicotiana attenuata*. *Proc Natl Acad Sci USA.* 2009;105:4559–64.
39. Pascal E, Sanderfoot AA, Ward BM, Medville R, Turgeon R, Lazarowitz SG. The geminivirus BR1 movement protein binds single-stranded DNA and localizes to the cell nucleus. *Plant Cell.* 1994;6:995–1006.
40. Pedersen TJ, Hanley-Bowdoin L. Molecular characterization of the AL3 protein encoded by a bipartite geminivirus. *J Virol.* 1994;202:1070–5.
41. Praveen S, Kushwaha CM, Mishra AK, Singh V, Jain RK, Varma A. Engineering tomato for resistance to tomato leaf curl disease using viral rep gene sequences. *Plant Cell Tissue Organ Cult.* 2005;83(3):311–8.
42. Praveen S, Ramesh SV, Koundal V, Mishra AK, Jain RK. Small RNA mediated silencing to target tomato leaf curl virus. *Plant Interact.* 2007;2(4):213–8.
43. Preiss W, Jeske H. Multitasking in replication is common among *geminiviruses*. *J Virol.* 2003;77(5):2972–80.
44. Qazi J, Ilyas M, Briddon RW. Legume yellow mosaic viruses: genetically isolated begomoviruses. *Mol Plant Pathol.* 2007;8(4):343–8.
45. Raney JL, Delongchamp RR, Valentine CR. Spontaneous mutant frequency and mutation spectrum for gene A of WX174 grown in *E. coli*. *Environ Mol Mutagen.* 2004;44:119–27.
46. Rey MEC, Ndunguru J, Berrie LC, Paximadis M, Berry S, Cossa N, Nuaila VN, Mabasa KG, Abraham N, Rybicki EP, Martin D, Pietersen G, Esterhuizen LL. Diversity of dicotyledonous-infecting *geminiviruses* and their associated DNA molecules in Southern Africa, including the south-west Indian Ocean Islands. *Viruses.* 2012;4(9):1753–91.
47. Rojas MR, Hagen C, Lucas WJ, Gilbertson RL. Exploiting chinks in the plant's armor: evolution and emergence of *geminiviruses*. *Annu Rev Phytopathol.* 2005;43:361–94.
48. Rosario K, Seah YM, Marr C, Varsani A, Krabberger S, Stainton D, Moriones E, Ploston JE, Duffy S, Breitbart M. Vector-enabled metagenomic (VEM) surveys using whiteflies (Aleyrodidae) reveal novel begomovirus species in the New and Old Worlds. *Viruses.* 2015;7(10):5553–70.
49. Rybicki EP. A phylogenetic and evolutionary justification for three genera of *Geminiviridae*. *Arch Virol.* 1994;139:49–77.
50. Saeed M, Zafar Y, Randles JW, Rezaian MAA. Monopartite begomovirus-associated DNA beta satellite substitutes for the DNA B of a bipartite begomovirus to permit systemic infection. *J Gen Virol.* 2007;88:2881–9.
51. Sahana N, Kaur H, Basavaraj Tena F, Jain RK, Paulkatis P, Canto T, Praveen S. Inhibition of the host proteasome facilitates papaya ringspot virus accumulation and proteasomal catalytic activity is modulated by viral factor HcPro. *PLoS ONE.* 2012;7:e52546.
52. Sahana N, Kaur H, Palukaitis P, Canto T, Praveen S. The asparagine residue in the FRNK box of potyviral helper-component protease is critical for template function and subcellular localization. *J Gen Virol.* 2014;95:1167–77.
53. Sanchez-Campos S, Martínez-Ayala A, Márquez-Martín B, Aragón-Caballero L, Navas-Castillo J, Moriones E. Fulfilling Koch's postulates confirms the monopartite nature of tomato leaf deformation virus, a begomovirus native to the New World. *Virus Res.* 2013;173:286–93.
54. Sanderfoot AA, Lazarowitz SG. Cooperation in viral movement—the geminivirus BL1 movement protein interacts with BR1 and redirects it from the nucleus to the cell periphery. *Plant Cell.* 1995;7:1185–94.
55. Settlege SB, See RG, Hanley-Bowdoin L. Geminivirus C3 protein: replication enhancement and protein interactions. *J Virol.* 2005;79(15):9885–95.
56. Singh D, Gill JS, Gumber RK, Singh R, Singh S. Yield and fibre quality associated with cotton leaf curl disease of Bt-cotton in Punjab. *J Environ Biol.* 2013;34(1):113–6.
57. Sudarshana MR, Wang HL, Lucas WJ, Gilbertson RL. Dynamics of bean dwarf mosaic geminivirus cell-to-cell and long-distance movement in *Phaseolus vulgaris* revealed, using the green fluorescent protein. *J Mol Plant Microbe Interact.* 1998;11:277–91.
58. Unseld S, Hohnle M, Ringel M, Frischmuth T. Subcellular targeting of the coat protein of African cassava mosaic geminivirus. *Virology.* 2001;286:373–83.
59. Wang H, Buckley KJ, Yang XJ, Buchmann RC, Bisaro DM. Adenosine kinase inhibition and suppression of RNA silencing by geminivirus AL2 and L2 proteins. *J Virol.* 2005;79:7410–8.

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