Population differentiation and selective constraints in Pelargonium line pattern virus

Aurora Castaño, Leticia Ruiz, Santiago F. Elena, Carmen Hernández*

Instituto de Biología Molecular y Celular de Plantas (Consejo Superior de Investigaciones Científicas-Universidad Politécnica de Valencia). CPI, Ed. 8E. Camino de Vera s/n, 46022 Valencia, Spain

* Corresponding author. Mailing address: Carmen Hernández, Instituto de Biología Molecular y Celular de Plantas, (Consejo Superior de Investigaciones Científicas-Universidad Politécnica de Valencia). CPI, Ed. 8E. Camino de Vera s/n, 46022 Valencia, Spain
Phone: +34 963 877 869 Fax: +34 963 877 859 E-mail: cahernan@ibmcp.upv.es
ABSTRACT

The genomic structure of Pelargonium line pattern virus (PLPV), a tentative member of a proposed new genus within the family Tombusviridae, has been recently determined. However, little is known about the genetic variability and population structure of this pathogen. Here, we have investigated the heterogeneity of PLPV isolates from different origins by sequence analysis of a 1817 nt fragment encompassing the movement (p7 and p9.7) and coat protein genes as well as flanking segments including the complete 3′ untranslated region. We have evaluated the selective pressures operating on both viral proteins and RNA genome in order to assess the relative functional and/or structural relevance of different amino acid or nucleotide sites. The results of the study have revealed that distinct protein domains are under different selective constraints and that maintenance of certain primary and/or secondary structures in RNA regulatory sequences might be an important factor limiting viral heterogeneity. We have also performed covariation analyses to uncover potential dependencies among amino acid sites of the same protein or of different proteins. The detection of linked amino acid substitutions has permitted to draw a putative network of intra- and interprotein interactions that are likely required to accomplish the different steps of the infection cycle. Finally, we have obtained phylogenetic trees that support geographical segregation of PLPV sequences.

Keywords: Pelargonium line pattern virus, family Tombusviridae, genetic variability, selective constraints, covariation analysis
1. Introduction

In general, RNA viruses are known to generate high levels of genetic variation that allow them to evolve rapidly facilitating their successful adaptation to new environments. The low fidelity of the viral encoded RNA dependent-RNA polymerases (RdRps), that may lack proofreading functions, has been proposed as the main underlying source for most variation (Agol, 2006; Castro et al., 2005; Holland et al., 1982; Sanjuán et al. 2010; Steinhauer and Holland, 1986). Nevertheless, the molecular composition of viral populations is not the direct result of the error rate of viral RdRps. In the case of plant RNA viruses, other factors play a major role in structuring genetic diversity of pathogen populations, including selection and genetic bottlenecks as those that occur during both systemic infection (French and Stenger, 2003; Li and Roossinck, 2004; Sacristán et al., 2003) and horizontal transmission by vectors (Ali et al., 2006; Betancourt et al., 2008). These other factors may lead to a considerable genetic stability as it has been reported for many different plant RNA viruses that appear more genetically stable than their animal counterparts (García-Arenal et al., 2001, 2003). This could be due to a combination of intrinsically lower rates of mutation, as suggested by recent and more accurate estimates (Malpica et al., 2002 Sanjuán et al., 2009; Tromas and Elena, 2010), and a reduced fixation rate of advantageous non-synonymous mutations because of weaker immune selection (García-Arenal et al., 2001). The identification and manipulation of factors that regulate the composition of the viral populations may offer a new set of tools to predict or control emerging diseases.

Pelargonium line pattern virus (PLPV) is a major geranium (Pelargonium spp.) pathogen in Spain, where prevalence rates above 50% have been reported (Alonso and Borja, 2005). The virus has also been detected in distinct European countries and it likely has a worldwide distribution (Bouwen and Maat, 1992; Franck and Loebenstein, 1994;
Stone, 1980). The frequent symptomless condition of PLPV infections (Alonso and Borja, 2005) compromise regulatory inspections and might have contributed to the spread of the virus. The factors that influence the appearance of symptoms, characterized by yellow-green spots and line patterns on the leaves, remain unclear but they are likely a combination of the viral isolate, the environmental conditions and the geranium cultivar.

PLPV virions are isometric in shape and hold a single stranded RNA molecule. Cloning and sequencing of genomic RNA (gRNA) together with reverse genetic experiments have recently allowed determination of the genome organization of PLPV (Castaño and Hernández, 2005, 2007; Castaño et al., 2009). The gRNA comprises 3883 nt and contains five open reading frames (ORFs) flanked by an unusually short untranslated region (UTR) of 6 nt at the 5´ end and by a 246 nt long UTR at the 3´ end. The two 5´-proximal ORFs encode proteins essential for replication, p27 and its read-through product p87 (the viral RdRp). Two small overlapping ORFs, located at the middle of the genome, encode proteins involved in viral movement (p7 and p9.7), while the 3´-proximal ORF encodes the coat protein (p37 or CP). The two replication proteins are translated directly from the gRNA whereas the movement and encapsidation proteins are translated from the unique PLPV subgenomic RNA (sgRNA) of 1.7 kb detected in infected tissue (Castaño et al., 2009).

PLPV taxonomic status has not been fully clarified yet. It was formerly considered as a tentative member of the genus Carmovirus but recent results supported its inclusion into a prospective new genus (Pelarspovirus) in the family Tombusviridae (Castaño and Hernández, 2005; Castaño et al., 2009; Stuart et al., 2006). Other tentative species of the prospective genus would be Pelargonium ringspot virus (PelRSV), Pelargonium chlorotic ring pattern virus (PCRPV) and Elderberry latent virus (ELV), that, as PLPV, produce only one sgRNA (Kinard and Jordan, 2002) in contrast with typical carmoviruses that generate two (Lommel et al., 2005). The distribution and importance of the two PLPV-
related, pelargonium-infecting viruses, PelRSV and PCRPV, is unknown as detection
surveys for these pathogens are lacking. Nevertheless, the presence of PCRPV in at least
two European countries has been recorded (Lisa et al., 1996; Ruiz et al., 2008) and
PelRSV was reported to cause serious problems to geraniums in Germany (Lesemann
and Adam, 1994).

The extent of PLPV variability remains to be ascertained as, so far, only the complete
sequence of a German isolate, that from which the genomic organization of the virus was
deduced (Castaño and Hernández, 2005; Castaño et al., 2009), and a partial sequence of
an American isolate, corresponding to the CP gene (Accession No. AY038067), have
been reported. Sequence information for other tentative members of the proposed genus
Pelarspovirus is even scarcer. Indeed, only the primary structure of PCRPV genome has
been fully determined while just the CP sequences of PelRSV and ELV are available.

In this work, we have studied the genetic variability among ten PLPV isolates
recovered from naturally infected geranium plants which were collected in four countries
at different times. We have obtained data that have allowed inferring selective constraints
acting on PLPV genome and/or encoded products and that suggest geographical
segregation of PLPV sequences. Additionally, covariation analyses have unveiled
potential protein interactions and have pointed to particular amino acids as candidates to
be involved in intra- and/or interprotein contacts.

2. Materials and methods

2.1. Viral isolates

Field PLPV isolates were obtained from geranium plants collected at distinct
geographical locations over four years (2000-2004). Sap from the original plants was
used to pass the virus into the experimental host *C. quinoa* by mechanical inoculation and
the viral population was recovered from this infected material. The isolates were
designed with the first letters of the country of origin followed by a number to distinguish
isolates from the same country (Table 1). PLPV isolate PV-0193, obtained from the
German Collection of Microorganisms and Cell Cultures (DSMZ, Braunschweig,
Germany) and characterized previously (Castaño and Hernández, 2005, 2007), was
included in the sequence analyses for comparison purposes.

2.2. Reverse transcription, PCR amplification, cloning and sequencing

Total RNA preparations were obtained from infected *C. quinoa* leaves by phenol
extraction and lithium precipitation (Verwoerd *et al.*, 1989) and used as templates for
reverse transcription (RT) reactions with Superscript II-RT (Invitrogen) and primer CH60
(5´-CCGGATCCGGGCGAGATCGGGGGTGTTAC-3´), complementary to the 3´
terminus of the viral sequence (nt 3859–3883) with a SmaI site (underlined) and a
BamHI site (in italics) at the 5´ terminus. RT products were PCR amplified with the
Expand High Fidelity PCR System (Roche) and primers CH60 and CH17 (5´-
GAAAATGGCCTCTACGGGGAC-3´), homologous to nt 2067-2088 of the PLPV
genome. After an initial denaturation step at 94 ºC for 2 min, PCR was performed for 35
cycles each of 30 sec at 95ºC, 30 sec at 60ºC and 3 min at 68ºC, followed by an extension
step of 10 min at 68 ºC. The resulting RT-PCR products were separated by
electrophoresis in 1% agarose gels, eluted and cloned into the pGEM-T easy vector
(Promega) or the plasmid pTZ19R (Fermentas). Two clones for each DNA fragment were
selected for sequencing with an ABI PRISM DNA sequencer 377 (Perkin-Elmer).

2.3. Sequence analysis
Multiple sequence alignments were constructed using MUSCLE (Edgard, 2004). For coding regions, translated amino acid sequences were first aligned and used as guide to built protein-coding nucleotide sequence alignments by concatenating codons using PAL2NAL (Suyama et al., 2006). The best-fitting model of nucleotide substitution was identified by MODELTEST (Posada and Crandall, 1998) as the general reversible GTR + $\Gamma_4$ model, with the frequency of each substitution type and the gamma distribution of among-site rate variation with four rate categories estimated from the empirical data. Recombination was ruled out as a potential confounding factor by using GARD (Kosakovsky Pond et al., 2006) and RDP (Martin et al., 2005) algorithms. A maximum likelihood tree was constructed using the above model of nucleotide substitution and its statistical significance was evaluated by bootstrap (upon 10,000 pseudoreplicates) using PHYML (Guindon and Gascuel, 2003). MEGA4 (Tamura et al., 2007) was used for computing within- and among-population nucleotide diversities (standard errors were computed by the bootstrap method based on 1000 pseudoreplicates) as well as for performing Tajima’s relative rates test (Tajima, 1993).

2.3.1. Identification of adaptive evolution in PLPV genomes

It is generally assumed that synonymous substitutions accumulate neutrally, at rate $d_S$ per synonymous site, because they have no effect on the amino acid composition of proteins and, henceforth, may not affect protein folding and functioning. In contrast, nonsynonymous substitutions, occurring at rate $d_N$ per nonsynonymous site, involve amino acid replacements and are more likely to affect, for bad or for good, the folding and function of proteins. The intensity of selection, $\omega$, can thus be evaluated as the ratio $\omega = d_N/d_S$. Values of $\omega < 1$ indicate purifying (i.e., negative) selection that results in
elimination of detrimental mutations from virus populations. A value of $\omega = 1$ represents selective neutrality, i.e., mutations stay in the population at frequencies which are only governed by genetic drift. Finally, values of $\omega > 1$ are indicative of directional (i.e., positive) selection, resulting in fixation of advantageous mutations (Sharp, 1997). Here, we have used the several maximum likelihood Bayesian methods available in the HYPHY packaged (Kosakovsky Pond and Frost, 2005) as implemented in the www.datamonkey.org server. Each ORF was separately analyzed. Only sites identified by at least half of the six methods available will be reported.

2.3.2. Molecular covariation within and among proteins

Selection not necessarily acts on single amino acids but it may operate on groups of amino acids in a concerted manner. This being the case, changes in one amino acid should appear associated to changes in the other members of the interaction group. Two different methods were used to evaluate the presence of coevolving amino acids within any given protein. First the Bayesian graphical model implemented in SPIDERMONKEY (Poon et al., 2007) and available online in the datamonkey server. Second, the mutual information content (MIC) approach described in Codoñer et al. (2006). For this second approach, significance $P$-values were computed, based on a million permutations, as the fraction of shuffles with a MIC value greater than or equal to the observed value. To minimize the number of false positives, the FDR method was applied (Benjamini and Hochberg, 1995). Only sites predicted to covary by both methods will be reported. Intermolecular covariation was only evaluated using the second methodology.

2.3.3. RNA secondary structure predictions
The secondary structure of selected regions of PFBV genome was predicted using MFO\textsc{fold} version 3.1 (www.bioinfo.rpi.edu/applications/mfold) (Mathews \textit{et al.}, 1999; Zucker, 2003).

3. Results

3.1. Genetic diversity in coding and non-coding regions of PLPV genome

The primary structure of a total of 18 PLPV cDNAs, 1817 nt in length and derived from nine isolates (Table 1), was determined, and the resulting sequences were combined with the two additional ones previously characterized from a German isolate (Castaño and Hernández, 2005, 2007), producing a total dataset of 20 sequences which encompassed part of the RdRp gene, the complete p7, p9.7 and CP genes as well as the 3′ UTR.

Genetic distances between each pair of sequences ranged from 0.001 to 0.081. The maximum values were found between pairs of sequences from isolates sampled during different years at different countries (e.g., pair SPA3-USA1 or SPA3-ITA4), and the minimal values were detected for pairs of sequences of a given isolate, as they could differ by just 1 nt (e.g., sequences from isolate ITA2). The nucleotide diversity for the whole population was 0.055±0.007 (Table 2), which was similar to nucleotide diversity estimates of populations of other plant viruses (García-Arenal \textit{et al.}, 2001). Nucleotide diversity calculated independently for each of the coding and non-coding regions included in the analysis ranged from 0.035±0.008 to 0.062±0.009, with the highest diversity corresponding to the CP gene and the lowest to the 3′ UTR (Table 2). Remarkably, the sequence of two genomic segments were strictly conserved in all
isolates, one encompassing nt 3642-3707 and corresponding to a 5′-portion of the 3′ UTR, and the other comprising nt 2240-2279 and matching the leader sequence of the PLPV sgRNA plus short flanking stretches (Castaño and Hernández, 2005). Consistent with the principle that transitions are biochemically more likely than transversions, transition mutations were much more frequent than transversions, with the maximum composite likelihood estimate of the overall transitions to transversions rates ratio being 3.116. This excess also occurs when purines (5.616) or pyrimidines (7.239) are considered separately. A similar observation has been previously made for many other viruses (Liang et al., 2002; Mansky and Temin, 1995; Rico et al., 2006; Schneider and Roossinck, 2001; Tromas and Elena, 2010; Vartanian et al., 1997).

Nucleotide diversity was also estimated between and within PLPV subpopulations, considering a subpopulation as the group of isolates that were originally collected from a given country (Germany, Spain, USA and Italy). Between subpopulation diversity values considering either the complete 1817 nt region, individual ORFs or the 3′ UTR were greater than within subpopulation diversity values (Table 2), suggesting that there is significant differentiation of population according to the country from which the isolates were sampled.

To gain a better insight into the relationships between all PLPV isolates, a maximum likelihood phylogenetic tree was constructed from the nucleotide sequences included in the study. The results revealed two major groups of PLPV sequences: group I included all Spanish sequences and group II embraced sequences from Italy, Germany and USA which were divided into three clusters according with their geographical distributions (Fig. 1). When the analysis was performed with amino acid sequences deduced from any of the individual genes, the same phylogenetic groups were defined though the statistical significance of the internal nodes was, in general, lower than that obtained using the complete nucleotide sequences (data not shown).
3.2. Selective constraints on coding regions

The direction and intensity of selective constraints operating in each coding region was evaluated using the $\omega$ rates ratio statistic. The estimated average $\omega$ values were 0.109 (95% IC: 0.057-0.187) for the partial RdRp gene, 0.188 (95% IC: 0.112-0.292) for the p7 gene, 0.201 (95% IC: 0.131-0.293) for the p9.7 gene, and 0.096 (95% IC:0.073-0.123) for the CP gene. Thus the $\omega$ ratio was significantly below one for the PLPV ORFs included in the analysis, indicating that all of them are under purifying selection. We sought next identifying which particular amino acid sites were under purifying or directional selective constraints. To do so, $\omega$ was estimated for each position in the alignments. For the partial RdRp, four sites were detected under negative selection (T711, L732, N760, I763). In the case of p7, one site was detected to be under negative selection (S23) and another one under positive selection (S5). Curiously, the latter site overlapped with one of those found under negative selection in the RdRp (N760). Concerning p9.7, 10 amino acid sites were found to be under negative selection and 27 sites were found in the case of CP. Such sites were mainly concentrated in the central and N-terminal region of p9.7 and CP, respectively (Fig. 2 and data not shown).

3.3. Structural conservation of potential RNA regulatory sequences

The PLPV genomic region under study (nt 2067-3883) contains at least two segments that are presumed to play a key role in regulation of viral replication/transcription: the 3´ UTR and the promoter for synthesis of the PLPV sgRNA (in the minus strand). Though the latter one has not been experimentally defined, it is expected to embrace an stretch of about 100 nt preceding the initiation site of PLPV sgRNA (nt 2251; Castaño and
Hernández, 2005), in line with that reported for subgenomic promoters of related viruses (Li and Wong, 2006; Wang and Simon, 1997; Wang et al., 1999). Such promoters may fold into hairpin-like structure and this type of conformation seems to be critical for the mechanism of transcription of sgRNAs (Li and Wong, 2006; Wang et al., 1999).

Secondary structure predictions showed that the putative PLPV subgenomic promoter might also adopt a hairpin-like conformation with a small lateral branch. Interestingly, the sequence variation detected in this segment when comparing isolates, essentially maintained the predicted folding since most mutations were located in single stranded regions or, when affecting double stranded regions, they were compensatory or located at the base of loops or stems (Fig. 3A).

Concerning the 3′ UTR, it is expected to contain structural elements critical for viral replication and, most probably, also for translation on the basis of that found in other members of family Tombusviridae (Batten et al., 2006; Fabian and White, 2006; Fabian et al., 2003; Pogany et al., 2003; Sarawaneeyaruk et al., 2009; Stupina et al., 2008; Turner and Buck, 1999; Wang and Wong, 2004). The key role of the 3′ UTR during the infectious cycle implies that strong constraints may operate on the region to preserve its functionality that will likely depend on certain primary, secondary and/or tertiary RNA structures. In line with this view, the 3′ UTR showed a value of genetic diversity that, remarkably, was lower than those calculated for ORFs and, moreover, the nucleotide sequence of a 5′-proximal segment of this region was strictly conserved in all isolates as indicated above. In silico analysis showed that the 3′ UTR may fold into a series of stem-loops that was basically conserved in the different variants (Fig. 3B). Remarkably, a 5′-proximal stem-loop was formed by the conserved segment whereas the 3′-adjacent stem-loop concentrated most of the heterogeneity found in the non-coding region (Fig. 3B).

Collectively, the results suggested that conservation of specific conformations in regulatory sequences confer selective advantages to the viral RNA.
3.4. Variability in PLPV proteins

The sequence heterogeneity was unevenly distributed in the PLPV proteins. The C-terminal portion of the RdRp inferred from the amplified genomic region was not taken in consideration to analyze variability distribution as it represented only ~1/6 of the complete replication molecule. In the case of p7, 6 out of the 10 polymorphic positions detected were located in the first third of the protein (N-terminal 20 amino acids) despite the corresponding coding sequence overlapped in part with that of RdRp gene (data not shown). An amino acid replacement mapped at the putative RNA binding domain of p7 (V30I) but it did not affect basic residues which have been found essential for RNA binding in other carmoviruses (Marcos et al., 1999; Navarro et al., 2006). The variability of p9.7 also concentrated at the N-proximal half of the molecule as 12 out of the 17 polymorphic positions were detected within the N-terminal 40 amino acid residues of the protein (Fig. 2A), although most of the corresponding coding sequence overlaps with p7 gene. As reported for other members of the family Tombusviridae (Lommel et al., 2005), three different structural domains can be distinguished in the PLPV CP: (i) R, the N-terminal internal domain which contains many positively charged residues and must interact with RNA, (ii) S, the shell domain which forms a barrel structure made up of β strands and constitutes the capsid backbone and, (iii) P, the protruding C-terminal domain. Analysis of the distribution of the heterogeneity in PLPV CP revealed that the percentage of polymorphic positions in the P domain (18.09%) was higher than in the R (10.95%) or S (13.12 %) domains. A stretch within the R domain was absolutely conserved in all isolates (from A14 to N46) probably because structural and/or functional constraints. Supporting the existence of such constraints, a high proportion of the sites predicted to be under negative selection were located in this stretch (Fig. 2B).
Next, as an additional test for the effect of selection, we analyzed the possible existence of covariation groups within and between proteins (Fig. 4). Firstly, we focused on covariations within-proteins. Regarding the partial RdRp, three amino acid residues showed significant covariation, S704N-E713A-S745A. In the case of p7, two covariation groups were detected S8T-V30I and S11I-L43I, whereas three covariation groups were observed for p9.7, Y3C-V6A, S16L-S39L and N24S-G88R (Fig. 4). The covariation that affected amino acids at positions 16 and 39 distinguished sequences from German isolate PV-0193, which showed the combination S16, S39, from those of the remaining isolates, that exhibited L residues at both positions. Up to eight covariation groups were detected for the CP, prominent among which was I10L-T64M, that differentiated the Spanish sequences (bearing the combination L10, M64) from those with other geographical origins (combination I10, T64). The analysis was extended to detect covarying positions between proteins. Remarkably, amino acid residues of p7 covaried with amino acid residues of the other three proteins included in the study. Thus, the p7 covariation S11I-L43I was significantly linked to RdRp substitution N760S and this linkage distinguished Italian isolates (with the combination RdRp S760, p7 I11, I43) from the remaining ones (with the combination RdRp N760, p7 S11, L43). In addition, the amino acid replacement V30I in p7 was linked to the amino acid replacement D32S in p9.7 and to the covarying group I10L-T64M in CP, and the corresponding combinations (p7 I30, p9.7 S32, CP L10, M64 versus p7 V30, p9.7 D32, CP I10, T64) segregated the Spanish isolate from the rest, further highlighting geographical distinctions between isolates. Finally, the covarying group S16L-S39L of p9.7 was linked to the amino acid substitution T236N of CP according to the programs employed, though visual inspection of alignments allowed to detect a strict association also with S213N/K (Fig. 2B). Indeed, the combination of S16, S39 in p9.7 and S213, T236 in CP was specific for the PV-0193
sequences whereas the remaining isolates showed the combination L16, L39 in p9.7 and N/K213, N236 in CP (Fig. 2B).

4. Discussion

In this work, the extent and structure of genetic diversity in PLPV have been explored by sequence analysis of an 1817 nt fragment (representing about 50% of the complete viral genome) of ten viral isolates sampled from four distinct geographical areas. Nucleotide diversity of the whole population was relatively low but in the range of those estimated for other plant viruses (Fraile et al., 1996, 1997; García-Arenal et al., 2001; Moya et al., 1993; Rodríguez-Cerezo et al., 1991; Rubio et al., 2001; Font et al., 2007). Diversity values among isolates collected in the same country were lower than those found among isolates of different countries suggesting a significant clustering of isolates by country of origin, a postulate supported by phylogenetic analysis. As the sample size of the present study is relatively small, characterization of new PLPV isolates is required to confirm whether genotype distribution into phylogroups related with geographical areas certainly reflects the structure of the viral population at global scale.

The relatively low genetic diversity found for PLPV suggested that negative selection is restricting the number of molecular variants. Consistently, \( \omega \) values estimated for coding regions pointed to purifying selection as the predominant evolutionary pressure operating on such genome segments likely to preserve the encoded amino acid sequences. The smallest \( \omega \) value was recorded for the CP, though it was similar to those of other plant viruses (García-Arenal et al., 2001; Font et al., 2007). CP in related viruses has been involved in other functions besides genome encapsidation, such as virus movement and suppression of RNA silencing (Genovés et al., 2006; Martínez-Turiño and Hernández, 2009; Thomas et al., 2003; Turina et al., 2000), which could explain the
strong negative selection observed. Examination of selective constraints on particular
amino acids showed, as expected, a strong bias among the number of sites under negative
and positive selection (42 versus 1). Only sites under purifying selection were detected in
the portion of the RdRp included in the analysis, despite such portion did not comprise
any of the eight motifs conserved in the RdRps (Koonin and Dolja, 1993). Regarding p7,
the unique negatively selected site was located in the putative RNA-binding motif of the
protein but it did not correspond to any of the basic residues that are presumably critical
for RNA-binding capability (Marcos et al., 1999; Navarro et al., 2006). In the case of
p9.7, 9 out of the 10 negatively selected sites concentrated in the central part of the
molecule (codons 29 to 54) which essentially matched the region that connects the two
hydrophobic domains that, according to that reported for related proteins (Navarro et al.,
2006; Saurí et al., 2005; Vilar et al., 2002), must be involved in membrane association.
Concerning the CP, almost 2/3 of the negatively selected sites (17/27) were located
within the N-terminal 66 residues that constitute the R domain; however, only two of
them corresponded to basic residues which are likely critical for RNA-binding capability
suggesting that, as likely occurs in the case of p7, selection is acting on the preservation
of the proper conformation of the RNA-binding motif. Moreover, the S and P domains
showed identical number of negatively selected sites (5 each) despite the general trend to
conservation of the former one in family Tombusviridae (Lommel et al., 2005).
On the other hand, the variability patterns found in regulatory sequences of the viral
RNA, such as the putative subgenomic promoter or the 3’ UTR, support the existence of
structural constraints that prevent the loss of their functionality. In the case of the
subgenomic promoter, besides its predicted role in transcriptional regulation, it
completely overlaps the 3’-portion of the RdRp gene (in the minus strand) and thus the
same stretch is expected to have a dual function as coding and as regulatory sequence,
which should considerably restrict heterogeneity. Accordingly, the mean of nucleotide
diversity in this region was lower than in other coding regions (Table 2) and, moreover, the nucleotide substitutions did not disrupt the predicted hairpin-like structure that is presumably required for the promoter function (Fig. 3A), suggesting that conservation of this conformation significantly influences the profile of naturally occurring mutations.

Regarding the 3′ UTR, the variability data support that maintenance of a specific folding composed by a series of stem-loops might limit its sequence heterogeneity. Different members of family Tombusviridae have been reported to contain in this region cis elements critical for viral replication, such as promoters and repressors for minus strand synthesis (Na and White, 2006; Pogany et al., 2003; Stupina and Simon, 1997; Zhang et al., 2004a, 2004b), and others relevant for gene expression, such as translational enhancers that promote cap independent translation, since the viruses of this family are characterized by non-blocked 5′-termini (reviewed by Kneller et al., 2006). An element of this type has been proposed to be present in the 3′ UTR of PLPV RNAs (Fabian and White, 2006) which, remarkably, would be embedded in the 5′-proximal stem-loop that is strictly conserved in all isolates (Fig. 3B) providing indirect evidence for its functional significance.

Population diversity studies with other members of the family Tombusviridae have also highlighted conservation of structural motifs in both regulatory RNA sequences and encoded proteins. Specifically, analysis of the genetic heterogeneity of Carnation mottle virus (CarMV) and Pelargonium flower break virus (PFBV), two members of the genus Carmovirus, has revealed that the pattern of natural variability preserves the conformation of putative replication cis-acting signals (Cañizares et al., 2001; Rico et al., 2006). Regarding proteins, different degrees of variation have been found when comparing equivalent products of PFBV, CarMV and PLPV though some common tendencies are noticeable. Among them, it is noteworthy mentioning the high conservation of the RNA binding motif of the small movement protein (or of the protein
itself in the case of PFBV) or the considerable sequence flexibility in the N-terminal region of the large movement protein.

Correlated amino acid mutation analysis has been widely used to infer functional interactions between different sites in a protein or between distinct proteins (e.g.: Altschuh et al., 1987; Codoñer et al., 2006; Hoffman et al., 2003; Larson et al., 2000; Thomas et al., 1996). The study of PLPV genetic variability has allowed identification of groups of amino acids that covary both within and between PLPV proteins, revealing a putative network of interactions that is likely needed for maintenance of proper protein folding and for driving the viral RNA from replication to cell-to-cell/systemic translocation. Though the exact mechanism that account for inter-cellular transport of carmo-like viruses is not yet known, it is not unlikely to require a physical interaction among CP and movement proteins as reported for other plant viruses (Akamatsu et al., 2007; Kim et al., 2004; Liu et al., 2001; Sánchez-Navarro and Bol, 2001; Sánchez-Navarro et al., 2006; Takeda et al., 2004). Consistently with this view, we have detected covariations between the PLPV movement proteins, p7 and p9.7, and the CP. More intriguing is the covariation found between the RdRp and the p7. Evidence for interaction between the viral polymerase and the movement protein has been obtained for Cucumber mosaic virus which has led to the suggestion that both proteins cooperate in regulating the intercellular movement of progeny viral RNA by an unknown mechanism (Hwang et al., 2005), a possibility that could also apply to PLPV. Future work will be aimed at corroborating the inferred interactions and at assessing whether the covarying amino acids identified in this work are actually involved in protein-protein contact interfaces as suggested from the present results.

Acknowledgements
We are indebted to Dr. Jan van der Meij (Ball Flora Plant, Chicago) for providing PLPV isolate from USA, and to Dr. Marisé Borja (Fundación Promiva, Madrid) for the Spanish isolates and for valuable comments in the course of this work. We thank to Dr. Selma Gago (IBMCP, Valencia) for critical reading of the manuscript. We are also grateful to Dolores Arocas and Isabella Avellaneda for excellent technical assistance. This research was supported by grants BFU2006-11230 and BFU2009-11699 from Ministerio de Ciencia e Innovación (MICINN, Spain), ACOM09/040 from Generalitat Valenciana (to C.H.), and by grant BFU2009-06993 (MICINN) (to S.F.E.). A.C. was recipient of predoctoral fellowships from the Generalitat Valenciana and from CSIC-Fundación Bancaja and L.R. received a postdoctoral contract from the Juan de la Cierva program of MEC.
References


Li, W., Wong, S.M., 2006. Analyses of subgenomic promoters of Hibiscus chlorotic ringspot virus and demonstration of 5’-untranslated region and 3’-terminal sequences functioning as subgenomic promoters. J. Virol. 80, 3395-3405.


FIGURE LEGENDS

**Fig. 1.** Minimum evolution unrooted phylogenetic tree inferred from nucleotide sequences derived from 10 isolates of PLPV (see Table 1). Phylogenetic analysis was conducted with programs included in the MEGA4 package. The numbers at the nodes are bootstrap support values based on 10,000 pseudoreplicates; only values >50% are shown.

**Fig. 2.** Alignment of partial amino acid sequences of p9.7 (A) and CP (B) from different isolates of PLPV. The reference sequence derived from isolate PV-0193 (Castaño and Hernández, 2007) is shown at the top of each alignment. Numbers above the reference sequence correspond to positions in the complete protein. Those residues conserved in all isolates are indicated by dots. The two amino acid sequences inferred from two cDNA clones selected from each isolate are shown. Underlined residues are under negative selection. In (A), the arrow demarcates the residues whose coding sequence overlaps with that of p7 gene. The two predicted hydrophobic regions of the protein are depicted within brackets at the bottom. In (B), the domains of the CP at which the residues belong are indicated at the top.

**Fig. 3.** MFOLD-predicted RNA secondary structures of the putative subgenomic promoter (A) and the 3’ UTR (B) of PLPV. The distribution of polymorphic positions is indicated on the most stable folding of the reference sequence corresponding to isolate PV-0193 (Castaño and Hernández, 2007). Numbers denote positions in the PLPV gRNA. The minus strand is shown in (A).
Fig. 4. Covariations within and between p7, p9.7, CP and the partial RdRp. The residues covarying within a given protein are connected by solid lines and those covarying among proteins are connected by dashed lines.
<table>
<thead>
<tr>
<th>Accession number</th>
<th>Original Host</th>
<th>Year(s)</th>
<th>Country</th>
<th>Isolate</th>
</tr>
</thead>
<tbody>
<tr>
<td>AV613852, EU835946</td>
<td>P. peletarium</td>
<td>1990s</td>
<td>Germany</td>
<td>P-Y0-1993</td>
</tr>
<tr>
<td>EU852927, EU852928</td>
<td>P. zonale</td>
<td>2004</td>
<td>Italy</td>
<td>ITA4</td>
</tr>
<tr>
<td>EU852925, EU852926</td>
<td>P. zonale</td>
<td>2004</td>
<td>Italy</td>
<td>ITA2</td>
</tr>
<tr>
<td>EU852923, EU852924</td>
<td>P. peletarium</td>
<td>2006</td>
<td>USA</td>
<td>USA1</td>
</tr>
<tr>
<td>EU852921, EU852922</td>
<td>P. zonale</td>
<td>2004</td>
<td>Spain</td>
<td>SPA8</td>
</tr>
<tr>
<td>EU852919, EU852920</td>
<td>P. zonale</td>
<td>2000</td>
<td>Spain</td>
<td>SPA7</td>
</tr>
<tr>
<td>EU852916, EU852918</td>
<td>P. zonale</td>
<td>2000</td>
<td>Spain</td>
<td>SPA4</td>
</tr>
<tr>
<td>EU852916, EU852917</td>
<td>P. zonale</td>
<td>2000</td>
<td>Spain</td>
<td>SPA3</td>
</tr>
<tr>
<td>EU852914, EU852915</td>
<td>P. zonale</td>
<td>2000</td>
<td>Spain</td>
<td>SPA1</td>
</tr>
<tr>
<td>EU852912, EU852913</td>
<td>P. zonale</td>
<td>2000</td>
<td>Spain</td>
<td>SPA0</td>
</tr>
</tbody>
</table>

Table 1. P. PLPy isolates used in this work
<table>
<thead>
<tr>
<th></th>
<th>Complete</th>
<th>3' UTR</th>
<th>CP</th>
<th>3'UTR</th>
<th>p7</th>
<th>Rdp</th>
</tr>
</thead>
<tbody>
<tr>
<td>NST = ( q_{ST} / q_{T} )</td>
<td>0.055 ± 0.007</td>
<td>0.035 ± 0.008</td>
<td>0.062 ± 0.009</td>
<td>0.052 ± 0.008</td>
<td>0.041 ± 0.008</td>
<td>0.057 ± 0.008</td>
</tr>
<tr>
<td>BST = ( Y_{ST} - Y_{TS} )</td>
<td>0.013 ± 0.002</td>
<td>0.026 ± 0.002</td>
<td>0.005 ± 0.002</td>
<td>0.031 ± 0.002</td>
<td>0.010 ± 0.002</td>
<td>0.055 ± 0.002</td>
</tr>
<tr>
<td>Ns = ( Y_{S} )</td>
<td>0.042 ± 0.005</td>
<td>0.026 ± 0.006</td>
<td>0.011 ± 0.007</td>
<td>0.041 ± 0.007</td>
<td>0.009 ± 0.007</td>
<td>0.042 ± 0.007</td>
</tr>
</tbody>
</table>

Table 2. Nucleotide diversity ± SEM (based on 1000 bootstrap replicates). Isolates from Spain, Italy, USA and Germany have been considered as subpopulations.
Fig. 1
Fig. 2