Molecular evolution and phylogeography of potato virus Y based on the CP gene

José M. Cuevas,1 Agnès Delaunay,2† Matevz Rupar,3 Emmanuel Jacquot2† and Santiago F. Elena1,4

1Instituto de Biología Molecular y Celular de Plantas, Consejo Superior de Investigaciones Científicas-Universidad Politécnica de Valencia, Valencia 46022, Spain
2INRA-Agrocampus Ouest-Université Rennes 1, UMR1099 BIO3P (Biology of Organisms and Populations Applied to Plant Protection), Le Rheu 35653, France
3NIB, Department of Biotechnology and Systems Biology, Večna pot 111, 1000 Ljubljana, Slovenia
4The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

Potato virus Y (PVY) is an important plant pathogen with a wide host range that includes, among others, potato, tobacco, tomato and pepper. The coat protein (CP) of PVY has been commonly used in phylogenetic studies for strain classification. In this study, we used a pool of 292 CP sequences from isolates collected worldwide. After detecting and removing recombinant sequences, we applied Bayesian techniques to study the influence of geography and host species in CP population structure and dynamics. Finally, we performed selection and covariation analyses to identify specific amino acids involved in adaptation. Our results show that PVY CP diversification is significantly accounted for by both geographical and host-driven adaptations. Amino acid positions detected as positively selected concentrate in the N-terminal region of the protein. Some of these selected positions may discriminate among strains, and to a much lesser extent, between potato and non-potato isolates.

Potato virus Y (PVY) is responsible for serious diseases in potato, tobacco, pepper and tomato crops. PVY was originally classified into strain groups (e.g. PVYN, PVYO and PVYC) according to their biological properties, serological characteristics and/or genome sequences (Moury et al., 2002; Singh et al., 2008). Recombination is highly pervasive in PVY and additional genomic organizations have been described (Lorenzen et al., 2008; Schubert et al., 2007).

Molecular evolution studies are useful tools to shed light on the molecular bases of virus geographical spread and adaptation to new hosts and for designing better epidemic control strategies (Elena et al., 2011; Jones, 2009). We recently studied the phylogeography and molecular evolution of PVY whole-genomes (Cuevas et al., 2012), showing that host and geographical origin influenced PVY diversification, and detecting positively selected sites. Here, we revisit these topics but focus on coat protein (CP). Novelties of this study are: (i) a much larger dataset is available for CP, which is expected to allow a more robust characterization of phylogenetic and selection patterns, (ii) CP plays an important role in host adaptation for many plant viruses, and (iii) CP is the most diverse and well-studied gene in PVY and other potyviruses (Moury & Simon, 2011; Ogawa et al., 2008; Rohožková & Navrátil, 2010; Visser & Bellstedt, 2009).

A detailed description of the methods employed in this study can be found elsewhere (Cuevas et al., 2012). For this study, we retrieved 198 PVY CP sequences from GenBank, plus 94 additional sequences from worldwide isolates (PVYwide Organization, http://www.inra.fr/pvy_organization) (Table S1, available in JGV Online). This dataset was aligned with MUSCLE (Edgar, 2004) as implemented in MEGA 5 (Tamura et al., 2011). We ran recombination analyses to remove its effect from subsequent analyses. Bayesian Markov chain Monte Carlo (MCMC) coalescent analyses were performed with non-recombinant isolates to study the effect of local adaptation and host species in the observed diversity. Finally, we performed selection analyses to identify regions from the CP cistron that may be more likely involved in PVY adaptation dynamics.

Seventy-five of the 292 isolates (Table S1) showed a breakpoint, indicating ancestral recombination between PVYN and PVYC strains at position 9170 (considering the full genome) in CP (Schubert et al., 2007) and worldwide distributed. Five other isolates showed uncommon breakpoints detected by at least three of the methods implemented in RDP3 (Martin et al., 2010). N Nysa isolate

1Present address: INRA-CIRAD-Supagro, UMR385 BGT (Biology and Genetics of Plant-Pathogen Interactions), Montpellier 34398, France

Supplementary material is available with the online version of this paper.
showed a newly described breakpoint at position 8896 (Cuevas et al., 2012). IAC and v951204-N isolates showed a breakpoint at position 8735 (Mont and SASA-110 being the major and minor parents, respectively), almost coincident with other previously described breakpoints (Moury et al., 2002). Finally, S-RB96 and NN-UK-N isolates showed a new recombination point at position 8947 (SASA-110 and Mont are the major and minor parents, respectively). All recombinants were excluded, reducing the dataset to 212 isolates.

Phylogenetic analyses were performed using the GTR+$I_4$+I substitution model in the Bayesian MCMC framework, as implemented in BEAST 1.6 (Drummond & Rambaut, 2007). Substitution rates were estimated using the relaxed uncorrelated exponential clock model. The three typical PVY strain groups (PVYC, PVYO and PVYN) could be observed (Fig. S1), although the differentiation between PVYC and PVYO strains was poorly supported.

Chile3 occupies a basal position in the tree, outside any of the strain groups, supporting its ancestry (Moury, 2010). Within the PVYC clade, 17 of 22 isolates were collected from five different non-potato hosts. However, host species did not account for clustering within this clade, since most of the isolates from a given host were dispersed along the clade or closely grouped with isolates from other hosts. Only isolates PVY-MN and NC57 (from tobacco) formed a differentiated cluster, as observed previously (Keohoe & Jones, 2011; Mascia et al., 2010). PVYC clade has been subdivided into PVYC1 and PVYC2 subgroups depending on their ability to infect pepper (Blanco-Urgoiti et al., 1998). In our phylogenetic tree, only isolates PVY-C-CM and Adgen-C were of pathotype PVYC2, forming a differentiated cluster. Isolate CAA82, collected from pepper, grouped outside the PVYC1 subgroup. More isolates from subgroup PVYC2 are thus necessary to check the relative distance of isolate CAA82 to those from the non-pepper subgroup PVYC2. Most isolates in our dataset belong to PVYO. The globally low branch supports suggests a very genetically homogeneous group, compatible with a recent origin with minimal selection (Pagan et al., 2006; Roossinck et al., 1999). In fact, well-supported clusters within the PVYO clade included isolates with common geographical origins. Finally, a similar trend was observed in the PVYN clade, although internal branches close to the basis of the tree were usually well supported, thus differentiating several monophyletic clusters. Our study supports the classification proposed by Ogawa et al. (2008) into two PVYN main groups (i.e. N-Europe and N-North America). Some well-supported clusters were observed into each PVYN group, although this differentiation was not strictly associated with geographical origin.

A visual inspection of the maximum clade credibility (MCC) phylogeny did not show a clear structure in terms of geographical origin at the continent level (Fig. S1 and Table S1). For commercial and geographical reasons, North African and Middle East isolates were included in the European group. For the same reason, the only isolate from New Zealand was not included in any continental group. We used BATS 1.0b2 (Parker et al., 2008) to calculate three statistics (AI, association index; PS, parsimony score and MC, maximum monophyletic clade size) describing the correlation between the geographical and phylogenetic relationships. Significant signatures for geographical structure in the diversity of the CP cistron were observed when grouped by geographical origins (Table 1), as shown by the significant AI and PS values. Asian, European,
South African, and North American groups showed differentiated subpopulations (significant MC values). The South American group did not show a significant association, which is accounted for by the small sample size, and no inference was possible for the single New Zealand isolate.

Host-driven adaptation could also be tested using host as a grouping variable, and a significant signature was also observed (Table 1). In this case, the differentiation was due to three subpopulations of isolates derived from potato, tobacco and pepper. For tomato and black nightshade no significant association was detected, whereas no inference was possible for single isolates from aji and tamarillo. Since most of the samples in our dataset are potato isolates, the significance of AI and PS values could be a consequence of the global distribution of the same state across most of the branches in the tree (Parker et al., 2008). However, host structure explained the phylogeny quite well, since clade PVYC predominantly included non-potato isolates (17 of 22), whereas the remaining main clades only included 14 non-potato isolates (of 189). Twelve of the 14 non-potato isolates falling outside the PVYC clade were collected from tobacco. In this sense, tobacco infection could accidentally take place from potato crops early in the year, thus leading to misidentification of some tobacco isolates (M. Chrzanowska, personal communication). Besides, it is not surprising either that the tomato isolate GR_PVY12 fell outside clade PVYC, since tomato can be infected with most PVY potato isolates (Singh et al., 2008), and thus a recent introduction from potatoes cannot be excluded. Finally, the inclusion of black nightshade isolate SYR-Sn into the PVYO clade is surprising, although the biological properties of this isolate are not yet available.

Selective pressures at a codon level were estimated using FEL, IFEL and MEME methods (www.datamonkey.org). Intramolecular covariation analyses were carried out using CAPS 1 (Fares & Travers, 2006), as described previously (Cuevas et al., 2012). Table 2 shows the distribution of codon positions under purifying, neutral and positive selection, and covarying positions. As previously shown, most of the codons evolve neutrally, whereas purifying selection is the main force driving the evolution of CP (Cuevas et al., 2012). Negatively selected positions are scattered along the ORFs, suggesting that no domain is particularly constrained. FEL and IFEL predicted codon 1 as positively selected, whereas MEME detected three additional codons (68, 193 and 216) to be under episodic diversifying selection (Table 2). Finally, a covariation group of nine codons was also detected, all located in the first half of CP. Selected codon 1 was involved in this covariation group.

Previous phylogenetic studies showed that non-potato isolates mainly fell into the PVY\textsuperscript{C} clade (Ogawa et al., 2008; Schubert et al., 2007), highlighting the importance of host-driven adaptation. Our study, which included a significantly larger number of non-potato isolates, clearly showed that, in spite of the global consideration of non-potato isolates as belonging to the PVYC clade, several other non-potato isolates were dispersed in the phylogeny. In fact, the analysis of amino acid composition for positively selected and covarying positions showed no clear differences between potato and non-potato isolates (Tables S2 and S3). Globally, both groups, except for positions 24, 138 and 193, shared the same predominant amino acid at a given position. Whereas similar amino acid composition between both groups was found for positions 24 and 193, the main difference was found at position 138, since the predominant amino acid for non-potato isolates was absent in potato isolates (Table S3). Besides, with the exception of position 138, specific residues of potato and non-potato isolates were always present at low frequencies. We also obtained the amino acid composition of positively selected and covarying codons, but grouping in this case for the PVYC\textsuperscript{C}, PVYO\textsuperscript{O} and PVYN\textsuperscript{N} strains, which allowed us to check if selective forces were strain-specific (Tables S4 and S5). Globally, the same predominant amino acid at a given position was usually shared by the three strains. For those cases showing differences in the predominant amino acid, these predominant residues for a given strain were also usually present at low frequencies in at least one of the alternative strains. We observed positions 24 and 193 wherein the predominant amino acid for the PVYO\textsuperscript{O} strain was different from that of the PVYC\textsuperscript{C} and PVYN\textsuperscript{N} strains. Besides, the predominant amino acid from the PVYN\textsuperscript{N} strain was different from that observed in the PVYC\textsuperscript{C} and PVYO\textsuperscript{O} strains for positions 1, 11, 17, 26, 29 and 31. Finally, positions 99 and 138 showed different predominant residues for the three strains. For those cases showing differences in the predominant amino acid, these predominant residues for a given strain were also usually present at low frequencies in at least one of the alternative strains. We observed positions 24 and 193 wherein the predominant amino acid for the PVYO\textsuperscript{O} strain was different from that of the PVYC\textsuperscript{C} and PVYN\textsuperscript{N} strains. Besides, the predominant amino acid from the PVYN\textsuperscript{N} strain was different from that observed in the PVYC\textsuperscript{C} and PVYO\textsuperscript{O} strains for positions 1, 11, 17, 26, 29 and 31. Finally, positions 99 and 138 showed different predominant residues for the three strains. Interestingly, the predominant residue for the PVYC\textsuperscript{C} strain at these two positions was absent in the other two strains, although the predominant amino acids from the PVYO\textsuperscript{O} and PVYN\textsuperscript{N} strains were also present at low frequencies. Consequently, the analysis of amino acid composition at selected and covarying positions showed more partially discriminating residues among strains than among potato and non-potato isolates, which indicates that selective forces are mainly acting independently of the

**Table 2.** Results of the codon selection and covariation analyses at the CP gene

For selection methods (FEL, IFEL and MEME), the number of codons detected to be under negative, neutral or positive selection are given. The last column indicates the location of positively selected sites besides those positions showing covariation (CAPS).

<table>
<thead>
<tr>
<th>Method</th>
<th>Negative</th>
<th>Neutral</th>
<th>Positive</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>FEL</td>
<td>113</td>
<td>153</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>IFEL</td>
<td>76</td>
<td>190</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>MEME</td>
<td>NA*</td>
<td>NA*</td>
<td>4</td>
<td>1, 68, 193, 216</td>
</tr>
<tr>
<td>CAPS</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>1, 11, 17, 24, 26, 29, 31, 99, 138</td>
</tr>
</tbody>
</table>

*NA*, Not applicable.
potato/non-potato distinction. In this sense, as mentioned before, PVY does not have a narrow host range, which would account for the lack of association between selected positions and host usage.

Selection analyses at a branch level were performed using SWAPSC (Fares, 2004) to check the potential association between selective events and the phylogeny. Thirty-four branches showed evidence of positive selection (18 internal and 16 terminal branches; Fig. S1), and this selective signature was detected in 13 regions, often overlapping (Table 3). Most of them fell into the N-terminal region, congruently with the above selection and covariation analyses (Tables 2 and 3). With respect to the distribution of the selected branches in the phylogeny, we could differentiate between internal and terminal branches (Fig. S1). The frequency of selected internal branches was different among clades (20, 3.7 and 15.8% for PVYC, PVYO and PVYN clades, respectively; Fisher’s exact test, \( P = 0.003 \)), but not for terminal branches (with frequencies of 9.1, 6.1 and 10.5% for PVYC, PVYO and PVYN clades, respectively; Fisher’s exact test, \( P = 0.568 \)). These results suggest that selective forces are stronger into the PVYC and PVYN clades and milder into PVYO. It is worth mentioning that one selected internal branch lead to the PVYC clade (named as b2 in Table 3 and Fig. S1), except for the tamarillo isolate falling outside the selected cluster. We obtained the amino acid composition of the region involved in this branch specific selection event (codons 187–194) for the PVYC, PVYO and PVYN clades (Table S6). This region included the selected site 193, which have been discussed above. Besides, the predominant amino acid for the PVYN clade was different from that observed in the PVYC and PVYO clades at position 187. Finally, position 194 clearly discriminated between the PVYO and PVYN clades, but the two fixed residues present in these strains were also observed in the PVYC strain. In conclusion, branch selection analyses showed evidence of the differential effect of selective events among strains, but did not provide particular positions accounting for these differences at a strain level.

The role of CP in the pathology of potyviruses has been previously confirmed (Andrejeva et al., 1999; Hu et al., 2011; Ullah & Grumet, 2002) and symptom determinants may be different even between strains of PVY in a particular host (Bukovinszki et al., 2007). The N-terminal part of CP is a clear example of multifunctionality. It is exposed on the virion surface (potential function in binding ligands), besides being involved in vector transmission (Peng et al., 1998) and systemic plant colonization (Andersen & Johansen, 1998; López-Moya & Pirone, 1998), becoming a potential target of selection at both vector and plant levels. In addition, CP from PVY interacts with different chloroplast proteins (Feki et al., 2005). Consequently, it is not easy to discern if a given amino acid position is involved in one or more functions.

Regarding biological functions of CP, several commonalities were found when comparing our results with those described by Moury & Simon (2011). All positions showing positive selection in this previous study are within the N-terminal region of the CP cistron. In particular, positions 11, 24, 26, 68 and 138, were also detected to be under positive selection or covariation in our study. Position 11 is close to the DAG conserved motif involved in aphid transmission (Atreya et al., 1991, 1995), and it has been

<table>
<thead>
<tr>
<th>Region</th>
<th>Branch</th>
<th>FEL-IFEL-MEME</th>
<th>Covariation</th>
</tr>
</thead>
<tbody>
<tr>
<td>7–11</td>
<td>SASA-110, b3</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>8–13</td>
<td>PN-82</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>23–28</td>
<td>b12</td>
<td>24, 26</td>
<td></td>
</tr>
<tr>
<td>23–29</td>
<td>b14</td>
<td>24, 26, 29</td>
<td></td>
</tr>
<tr>
<td>25–28</td>
<td>PB_707, US05_30, SYR-NB-16</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>25–29</td>
<td>b15, b16, b17</td>
<td>26, 29</td>
<td></td>
</tr>
<tr>
<td>26–29</td>
<td>b5, b7, b11</td>
<td>26, 29</td>
<td></td>
</tr>
<tr>
<td>29–33</td>
<td>CAA141, PB_707, PB_602, PB_752, SC143, SC61, US05_30, US05_7, NN71_111, SYR-NB-16, 605, b4, b8, b9, b13, b18</td>
<td>29, 31</td>
<td></td>
</tr>
<tr>
<td>62–65</td>
<td>German_45, US06_55, b6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>135–138</td>
<td>b1</td>
<td>138</td>
<td></td>
</tr>
<tr>
<td>187–193</td>
<td>Nicola</td>
<td>193</td>
<td></td>
</tr>
<tr>
<td>187–194</td>
<td>b2, b3, b10</td>
<td>193</td>
<td></td>
</tr>
<tr>
<td>214–217</td>
<td>German_45, b6</td>
<td>216</td>
<td></td>
</tr>
</tbody>
</table>
shown that mutations in a neighbouring residue can reduce transmissibility substantially (Atreya et al., 1995). Furthermore, position 25 was shown to affect virus accumulation in host plants (Moury & Simon, 2011), and covarying positions detected in the vicinity could have some influence in this respect. Regarding position 68, it is worth mentioning that a mutation in this codon promoted differences in viral accumulation and transmissibility by aphids (Moury & Simon, 2011). Finally, the region spanning amino acid positions 133–148 of CP from soybean mosaic virus (positions 136–151 of PVY CP) is involved in binding to HC-Pro (See et al., 2010), and then a potential influence for the included covarying position 138 could be postulated.

Acknowledgements

This research was supported by the Spanish Ministry of Science and Innovation grant BFU2009-06993 to S. F. E. and J. M. C was supported by the JAE-Doc program from CSIC. We thank Mario A. Fares for fruitful bioinformatics guidance. PVY isolates used in this work were available thanks to PVIwide Organization members (Carole Balmelli, Peter Dedic, Stewart Gray, Christophe Lacomme, Kestin Linder, Benoît Moury, Maja Ravnakivi, Laura Tomassoli, René Van der Vlugt, and Christina Varaveri). The authors do not have a commercial or any other association that might pose a conflict of interest.

References


