Selective constraint and genetic differentiation in geographically distant barley yellow dwarf virus populations

Gerod Hall
Cornell University, Department of Ecology and Evolutionary Biology, Corson Hall, Ithaca, NY 14853, USA

Numerous studies have documented molecular variability in plant virus populations, but few have assessed the relative contribution of natural selection and genetic drift in generating the observed pattern of diversity. To this end, gene function, environment and phylogenetic history were examined to observe the effect on genetic diversity and population structure of the PAV and PAS species of Barley yellow dwarf virus (family Luteoviridae). Three functional classes of gene were analysed: transcription-related (RdRp), structural (CP) and movement-related (MP). The results indicate that there were no inherent differences, in terms of total diversity or diversity at synonymous or non-synonymous nucleotide sites, between functional classes of genes or populations. Rather, selective constraints on a gene may be more or less relaxed depending on its function and the phylogenetic history of the population sampled. The CP of the PAS species, but not the PAV species, was differentiated genetically between regions. This is probably due to genetic drift, as there was no evidence that any gene deviated from a neutral model of evolution or is under positive selection. In general, the MP was under considerably less functional constraint than structural or replication-related proteins and four positively selected codon sites were identified. Mutations at these sites differentiate species and geographical subpopulations, so presumably they have aided the virus in adaptation to its host environment and contributed to intra- and interspecies diversification.

INTRODUCTION

Plant viruses may affect the fitness of their hosts by reducing host survivorship, fecundity or competitive ability relative to uninfected individuals in the population. Virus strains may differ markedly in the severity of symptoms that they induce (Anderson et al., 1991; Bencharki et al., 1999) and their ability to infect a given host (Moury et al., 2001; Sacristán et al., 2005). Variation among strains in virulence, infectivity and transmission may affect patterns of disease spread and, thereby, host population dynamics in natural systems (Raybould et al., 1999) or crop yield in agricultural systems. The behaviour and constraints of a virus upon host infection surely have a genetic underpinning. Thus, it is necessary to acquire knowledge of genetic diversity in pathogen populations to understand better the role that they play in ecological processes and as impediments to agricultural production. This study investigates qualitative (genetic differentiation) and quantitative (genetic variation) differences in genomic content among geographically distant populations of two species of Barley yellow dwarf virus (BYDV). This is a first step in understanding what evolutionary forces are acting on the viral genome.

Barley yellow dwarf (BYD) disease is caused by members of the family Luteoviridae in the genera Luteovirus and Polerovirus. Each virus species has a distinct aphid-transmission phenotype and the acronym for the species is derived from this specificity. Species in the family Luteoviridae commonly isolated from grain crops include GAV (Luteovirus), which is transmitted most efficiently by Schizaphis graminum and Sitobion avenae (Wang et al., 2001), MAV (Luteovirus), transmitted most efficiently by S. avenae (formerly Macrosiphum avenae), PAV (Luteovirus), transmitted most efficiently by Rhopalosiphum padi and S. avenae, SGV (unassigned to a genus within the family), transmitted most efficiently by S. graminum, and RPV (Polerovirus) transmitted most efficiently by R. padi (Rochow, 1969; Rochow & Muller, 1971). BYD disease has significant impacts in agricultural and natural plant communities. It is the most economically damaging viral disease of grain crops worldwide (Lister & Ranieri, 1995) and in grasslands it may contribute to shifts in community composition due to asymmetrical fitness effects on exotic...
and native grass species (Malmstrom et al., 2005a, b). Among the species listed above, PAV is the most widely distributed and economically important. To reflect significant variation in coat protein (CP) sequence among isolates, PAV has been divided into two species, PAV and PAS (Mayo, 2002). Sympatric populations of both species have been identified in Morocco (Bencharki et al., 1999), New York state (Chay et al., 1996a) and France (Mastari et al., 1998). The BYDV genome is composed of a single-stranded, positive-sense RNA with six open reading frames (ORFs) in total (Miller et al., 2002). Three viral genes were analysed in this study: ORFs 2, 3 and 4. ORF2 encodes the viral RNA-dependent RNA polymerase (RdRp) and is responsible for the replication of all viral RNAs (Koev et al., 2002). ORF3 encodes the major component of the CP, which is required for virion assembly (Mohan et al., 1995) and is thereby a prerequisite for aphid transmission (Gildow, 1987, 1993) and systemic plant infection (Filichkin et al., 1994). ORF4 encodes the movement protein (MP), which is required for the virus to spread systemically in the host (Chay et al., 1996b). ORF4 is embedded completely within ORF3, but is translated +1 bp out of the CP reading frame (Dinesh-Kumar & Miller, 1993).

There is currently a limited understanding of what factors influence genetic diversification and genome evolution in BYDV. This is due in part to a lack of studies that investigate variation in non-structural genes and explore how the genetic variation found in natural virus populations affects gene function. Through an analysis of diversity and selective constraint in multiple types of gene and across virus subpopulations, the present study was able to separate out the effects of gene function, environment and phylogenetic history in creating the pattern of diversity observed in the virus population. Thus, this study evaluates the role of selection in promoting or limiting genetic variability and provides a bridge between descriptive studies of population diversity and studies of protein structure and function (Moury, 2004).

METHODS

Virus isolates. The New York state virus isolates analysed in this study were obtained from field-caught R. padi or from field-collected corn (Zea mays), wheat (Triticum aestivum) and reed canary grass (Phalaris arundinacea) during the time period 1998–2003. Nucleotide sequences of members of the family Luteoviridae from other geographical regions were obtained from GenBank.

Immunocapture, RT-PCR and DNA sequencing. Virions were isolated from plant extracts by first grinding tissue in a mixture of dry ice and PBS, followed by immunocapture in anti-BYDV-PAV antibody-coated microfuge tubes. In the same tubes, RT-PCR was used to amplify the CP nucleotide sequence (603 bp) or a portion of the RdRp gene (654 bp). RT-PCR of the CP gene was performed by using oligonucleotide primers designed by Bencharki et al. (1999). The Primer Select program (DNASTAR software package) along with full-length PAV sequences of isolates PAS-129 and PAV-Aus, available in GenBank, were used to design primers to amplify the RdRp gene. Reverse transcription was performed by using oligonucleotide 5’-TTAGGGTCAACTCCGAAATGATTC-3’ and PCR using the former and oligonucleotide 5’-GGGCTAAGTGGGAAACCG-3’. Single-step RT-PCR was carried out in a 50 μl reaction volume containing 25 mM Tris (pH 8.8), 10 mM KCl, 10 mM dithiothreitol, 10 mM (NH₄)₂SO₄, 2.5 mM MgSO₄, 1 μM each primer, 0.4 mM BSA, 200 μM dNTPs, 30 units RNase inhibitor, 50 units SuperScript II and 5 units Taq polymerase (all enzymes were from Invitrogen). Thermocycling conditions were: one cycle of 45 min at 42°C for reverse transcription, one cycle of 2 min at 95°C for inactivation of reverse transcriptase, 36 cycles of 30 s at 94°C, 30 s at 53°C for the coat protein or 56°C for the RdRp, 1 min at 72°C, and a final extension of 10 min at 72°C. PCR products were purified by using a Qiagen PCR clean-up kit, then submitted to the Cornell University Biosource Center for direct sequencing. Sequencing was carried out in both the forward and reverse directions with the primers listed above.

Phylogenetic and nucleotide diversity analyses. Sequences were aligned with the CLUSTAL W algorithm of the MEGALIGN program (DNASTAR software package). Alignments of the RdRp, CP and MP genes were unambiguous, but manual adjustments were made to the alignment of the complete-genome sequences. In PAUP version 4.0b10 (Swofford, 2000), maximum-likelihood (ML) trees for the RdRp, CP and MP genes (PAV and PAS isolates only) were constructed by using the 2 ST model of nucleotide substitution with base frequencies and the shape of the gamma distribution estimated from the data. The resultant tree topologies were used for ML analysis of codon substitution. To examine phylogenetic relationships among BYDV species explicitly, a second set of trees was constructed by using the procedure outlined above, except that MAV, GAV, SGV, SbDV (Soybean dwarf virus, family Luteoviridae) and RPV (used as outgroup) sequences were included in the analysis. The addition of these sequences did not alter the topology of the PAV/PAS region of the tree. Thus, excluding the additional sequences, the trees depicted in Fig. 1(a–c) are identical to those used for the PAUP analysis. For each gene, robustness of the nodes of the phylogenetic tree was assessed by bootstrap percentages computed after 100 resamplings.

DnaSP version 4.0 (Rozas et al., 2003) was used to estimate total nucleotide diversity per nucleotide site (θ) (Nei, 1987). To compare diversity values, Friedman’s non-parametric test of population-central values and a series of pre-planned comparisons were implemented in SAS version 9.1. The first set of tests addressed whether there is a difference in diversity between the RdRp and CP or MP genes within species or homologous genes in different species (New York state population only). A second set of tests addressed whether diversity in the CP or MP differs within species in different geographical regions or between species in the same geographical region.

Tests of selective neutrality and population differentiation. Estimation of population parameters (s, k, θ) and Tajima’s D (Tajima, 1989) and Fu & Li’s F* and D* (Fu & Li, 1993) tests of selective neutrality were performed with DnaSP (Rozas et al., 2003). Tajima’s D test compares nucleotide diversity with the number of segregating sites, which are expected to be equal if mutations are selectively neutral. Fu & Li’s D* statistic is based on differences between the number of singletons (mutations appearing only once among the sequences) and the total number of mutations. Fu & Li’s F* statistic is based on the differences between the number of singletons and the mean number of nucleotide differences between pairs of sequences. SNAP Workbench (Price & Carbone, 2003) was used to implement the programs SEQTOMATRIX (Hudson et al., 1992), which facilitated the conversion of aligned CP sequences to a distance matrix, and PERMTEST (Hudson et al., 1992), which tested for geographical subdivision in the virus populations. PERMTEST calculates Hudson’s KST statistic of genetic differentiation. KST is equal to 1−Kd/Kt, where Ks is a weighted mean of K1 and K2 (mean
Selective constraint on luteovirus genes

Phylogenetic relationships among isolates

ML methods were used to determine the phylogenetic relationship among PAV and PAS isolates collected from a number of geographical regions. CP sequence identity between New York state isolates was similar to that number of differences between sequences in subpopulations 1 and 2, respectively) and $K_T$ represents the mean number of differences between two sequences regardless of their subpopulation. The null hypothesis of no genetic differentiation will be rejected ($P < 0.05$) when $K_S$ is small and $K_{ST}$ is close to 1.

Tests of positive selection and selective constraint. The ratio of non-synonymous (dN) to synonymous (dS) nucleotide substitutions ($\omega = dN/dS$) can be used to measure the degree of functional constraint for the maintenance of the encoded protein (Li, 1993). An $\omega$ ratio $> 1$ indicates that non-synonymous substitutions have a higher probability of fixation than synonymous substitutions and, presumably, offer a fitness advantage to the protein (positive selection). An $\omega$ ratio close to 0 indicates that non-synonymous substitutions are less likely than synonymous substitutions and the gene will be conserved at the amino acid level (negative selection). An $\omega$ ratio equal to 1 indicates neutral evolution. Two approaches were taken to determine the mode and strength of selection acting on the viral genome.

Fig. 1. ML trees of isolates of species in the family Luteoviridae, calculated by using (a) partial RdRp nucleotide sequences (ML score, 3126), (b) complete CP nucleotide sequences (ML score, 3101), (c) complete MP nucleotide sequences (ML score, 2099) and (d) complete-genome sequences (except SGV, ORFs 2, 3 and 4 only) (ML score, 44053). PAV isolates are shown in italics and PAS isolates in bold. All nodes with $< 70\%$ bootstrap support out of 100 replicates were collapsed to polytomies.
described for populations sampled in other geographical regions (Bencharki et al., 1999; Bisnieks et al., 2004) (Table 1). For the RdRp gene, there is low within-species diversity, but between species, diversity is greater than that found in the CP gene (Table 1). The RdRp tree (Fig. 1a) places PAV and MAV in a monophyletic clade that excludes PAS, but the CP (Fig. 1b) and MP (Fig. 1c) phylogenies place PAV and PAS in a monophyletic clade, with MAV sister to both species. The placement of CN-PAV also differs between gene trees. It clusters on a branch sister to all PAV and PAS isolates in the MP tree, but its RdRp is clearly derived from the same lineage as PAS. Analysis of the complete genomes of a number of different species in the family Luteoviridae supports the topology of the RdRp tree, i.e. MAV and PAV form a monophyletic clade and CN-PAV and PAS form a monophyletic clade (Fig. 1d). It also shows that SGV, which is unassigned to a genus in the family Luteoviridae, forms a monophyletic clade with luteovirus isolates and is only distantly related to the polerovirus RPV. Disagreement between RdRp and CP phylogenies could be the result of selection and/or genetic drift that has led to divergence of the CN-PAV and MAV CP sequences. It is also possible that ancestral luteoviruses have experienced recombination in the CP. The position of isolate ASL-1 in the CP and RdRp trees also warrants further investigation. ASL-1 shares greater nucleotide sequence identity with PAS isolates in its CP, but greater identity with PAV isolates in its RdRp sequence. ASL-1 may be a product of recombination between PAV and PAS or, alternatively, CP similarity could be the result of convergent evolution.

### Qualitative and quantitative differences in genomic content among species and geographical subpopulations

To better understand patterns of genetic diversity in the virus population, Hudson’s $K_{ST}$ test of population subdivision (Hudson et al., 1992) was performed. Hudson’s test showed evidence of genetic differentiation in the CP of the New York and Moroccan PAS populations ($P<0.01$, $K_{S}=4.1$, $K_{T}=16.0$, $K_{ST}=0.74$). It did not, however, detect differentiation in the PAV population ($P>0.16$, $K_{S}=15.5$, $K_{T}=16.3$, $K_{ST}=0.051$), despite clear segregation of Moroccan and New York isolates within cluster groups (Fig. 2). This may be due to the clustering of the PAV CP into two groups [described by Bisnieks et al. (2004)], each containing New York and Moroccan isolates, which could act to inflate $K_{S}$ and in turn lower the $K_{ST}$ value. I then compared total nucleotide diversity across genes and geographical regions between and within virus species (Table 2). It was found that, for the RdRp and CP comparison, there is a significant interaction between gene and species ($df=1$, $F=8.92$, $P<0.01$), thereby indicating that the probability that two randomly chosen isolates differ at a given nucleotide site is dependent upon the identity and phylogenetic history of the gene. This is reflected by the pre-planned comparisons found to yield significant $P$ values. The CP of PAS is less diverse than its RdRp ($P<0.01$) and less diverse than the CP of PAV ($P<0.01$). Similar results were obtained in comparison of the RdRp with the MP, in that there was no singular effect of gene or species and the MP of PAS was less diverse than the PAV RdRp ($P<0.01$) and less diverse than the MP of PAV ($P<0.01$). I next asked whether diversity in the CP or MP differs between species in the same geographical region and within species between geographical regions. For the CP gene, the effect of geographical region on nucleotide diversity was dependent upon the phylogenetic history of the isolate ($df=1$, $F=8.28$, $P<0.01$). It was also found that the PAS population in Morocco is significantly less diverse than the PAV population in New York ($P<0.01$) and the PAV population in Morocco ($P<0.01$). Diversity did not differ between the PAV population in New York state and Morocco. An identical pattern was found in analysis of the MP, where Moroccan PAS isolates had significantly lower nucleotide diversity than all other populations ($P<0.01$), but diversity did not differ between species in New York or between geographical regions for PAV.

### Tests of positive selection and selective neutrality and comparisons of selective constraint

To determine whether natural selection played a role in generating the genetic diversity observed in the virus population, I used ML models of codon substitution to test whether positive-selection models that allow for positively selected codon sites, M2 and M8, did not fit the data significantly better than models that assume that all sites are either conserved or neutral, M1 or M7 (Table 3). The BEB analysis did, however, identify four codon sites in the MP that are potentially undergoing positive selection (posterior probability of site belonging to class $K_{S}$ >95% with M8 and >90% with M2). No positively selected sites were detected in the RdRp or the CP. Mutations at codon site 9 (PAV-Aus used as reference) were region-specific, but not species-specific. All virus isolates collected in Morocco have lysine at this site, some New York PAV isolates, some New York PAS isolates and CN-PAV have glutamic acid and some New York PAS isolates have aspartic acid. The alanine mutation at codon 68 was unique to New York PAS isolates. All other PAS and PAV isolates have valine at this site. At two sites, MP substitutions were differentiated between species rather than between geographical regions. At codon

---

**Table 1. Percentage nucleotide identity within the CP (600 nt) (above diagonal) and RdRp (654 nt) (below diagonal) coding regions for PAV and PAS isolates collected in New York state**

<table>
<thead>
<tr>
<th>No. isolates</th>
<th>Virus</th>
<th>PAV</th>
<th>PAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>PAV</td>
<td>100–97</td>
<td>91</td>
</tr>
<tr>
<td>5</td>
<td>PAS</td>
<td>100–99</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>PAV</td>
<td>98</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>PAS</td>
<td>78–80</td>
<td>97–98</td>
</tr>
</tbody>
</table>
Numbers in parentheses are standard deviations of estimates. DnaSP was used to estimate \( \pi \) as the mean of pairwise comparisons among sequences in a population (Nei, 1987). Freidman’s non-parametric test of population-central values was used to compare \( \pi \) values. Values with the same letter were included in the same analysis and capitalization denotes values found to be significantly different \((P<0.05)\).

<table>
<thead>
<tr>
<th>Region</th>
<th>Species</th>
<th>RdRp</th>
<th>CP</th>
<th>MP</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York</td>
<td>PAV</td>
<td>0.017</td>
<td>0.025 (0.007)(^{bc})</td>
<td>0.024 (0.006)(^{bd})</td>
</tr>
<tr>
<td></td>
<td>PAS</td>
<td>0.002</td>
<td>0.01 (0.001)(^{bc})</td>
<td>0.009 (0.002)(^{bd})</td>
</tr>
<tr>
<td>Morocco</td>
<td>PAV</td>
<td>*</td>
<td>0.028 (0.006)(^{c})</td>
<td>0.026 (0.006)(^{d})</td>
</tr>
<tr>
<td></td>
<td>PAS</td>
<td>*</td>
<td>0.006 (0.002)(^{c})</td>
<td>0.003 (0.001)(^{d})</td>
</tr>
</tbody>
</table>

*No sequences available for analysis.

site 133, all PAS isolates have asparagine, PAV isolates in one cluster group have glycine and PAV isolates in the other cluster group have arginine. At codon site 145, all PAV isolates have threonine, all PAS isolates have methionine and CN-PAV has glutamic acid. These selection events may have occurred before populations were separated on different continents. None of the test statistics generated from Tajima’s D and Fu & Li’s \( F^* \) and \( D^* \) tests were significant \((P>0.10; \text{Table 4})\). This indicates that the hypothesis of selective neutrality for the CP, RdRp and MP in the New York and Moroccan PAV and PAS populations cannot be rejected.

To examine selective constraints experienced by the RdRp, CP and MP, I compared \( dS/(dS+\text{constant}) \) and \( dS/(dN+dS) \) for each gene across geographical regions and lineages. Both methods underestimate \( \omega \), but \( dN/(dN+dS) \) has the further shortcoming that it does not allow values >1. Similar distributions and \( P \) values were obtained by using either method, thus only values calculated by \( dS/(dS+\text{constant}) \) are presented in Table 5. To evaluate whether selective constraint is correlated with gene function, I compared the RdRp with the CP or MP. Results show that selective constraint is dependent upon the phylogenetic history of a gene and its function \((df=1, F=12.7, P<0.01)\).

For the MP, the effect of geographical location on selective constraint is not significant \((df=1, F=2.43, P<0.01)\). Thus, neither the RdRp nor the CP is necessarily more functionally constrained than the other. For the RdRp/MP comparison, there was a significant effect of function \((df=1, F=23.18, P<0.01)\), but not geographical location \((df=1, F=0.09, P=0.78)\), has a significant impact on selective constraint. For the MP, the effect of geographical location on selective constraint is dependent upon the phylogenetic history of the isolate \((df=1, F=11.08, P<0.01)\). This is reflected by the pre-planned comparisons, where the PAS population in

### Table 2. Nucleotide diversity per site \( (\pi) \) in the RdRp, CP and MP genes of BYDV-PAV and PAS isolates collected in New York and Morocco

<table>
<thead>
<tr>
<th>Region</th>
<th>Species</th>
<th>RdRp  ( (\pi) )</th>
<th>CP  ( (\pi) )</th>
<th>MP  ( (\pi) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York</td>
<td>PAV</td>
<td>0.017</td>
<td>0.025</td>
<td>0.024</td>
</tr>
<tr>
<td></td>
<td>PAS</td>
<td>0.002</td>
<td>0.01</td>
<td>0.009</td>
</tr>
<tr>
<td>Morocco</td>
<td>PAV</td>
<td>*</td>
<td>0.028</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>PAS</td>
<td>*</td>
<td>0.006</td>
<td>0.003</td>
</tr>
</tbody>
</table>

*No sequences available for analysis.

### Table 3. ML analysis of codon substitution (Yang, 1997) for detection of positive selection in the RdRp, CP and MP of PAV and PAS isolates collected in New York and Morocco

<table>
<thead>
<tr>
<th>Gene</th>
<th>Mean ( \omega )</th>
<th>2( \Delta )L ( ^{\text{f}} ) M2 vs M1</th>
<th>( P ) value ( ^{\text{i}} )</th>
<th>2( \Delta )L M8 vs M7</th>
<th>( P ) value</th>
<th>Positively selected site ( ^{\text{j}} )</th>
<th>Amino acid change</th>
</tr>
</thead>
<tbody>
<tr>
<td>RdRp</td>
<td>0.05</td>
<td>17.4</td>
<td>0.74</td>
<td>0</td>
<td>1</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>CP</td>
<td>0.37</td>
<td>1.2</td>
<td>0.55</td>
<td>0</td>
<td>1</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>MP</td>
<td>0.67</td>
<td>4.2</td>
<td>0.12</td>
<td>4</td>
<td>0.13</td>
<td>2928</td>
<td>E to K</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3105</td>
<td>E to D</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3300</td>
<td>A to V</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3361</td>
<td>M to Y</td>
</tr>
</tbody>
</table>

*Mean \( \omega \) (dN/dS) for best-fit model among M0, M1 and M2; model M0 assumes one ratio for all codon sites, M1 and M7 assume \( \omega = 0 \) or \( \omega = 1 \) for all sites, M2 and M8 allow \( \omega > 1 \) (Yang et al., 2000).

†Twice the difference in \(-\log \text{likelihood}\) between the general and reduced models.

‡Probability that \( 2\Delta L \) is smaller than a \( \chi^2 \) with two degrees of freedom.

§Codon belongs to site class \( \omega > 1 \) with 90% confidence for M2 and 95% confidence for M8; number indicates first nucleotide position of codon site, with PAV-Aus used as reference sequence.

http://vir.sgmjournals.org
Morocco was found to be more constrained than the PAS population in New York ($P < 0.01$) and PAV population in Morocco ($P < 0.01$).

**DISCUSSION**

**Phylogenetic relationships and genetic structure**

Previous studies of genetic variation in BYDV populations have focused exclusively on genes in the 3' half of the viral genome (Bencharki et al., 1999; Chay et al., 1996a; Mastari et al., 1998). The findings of the present study demonstrate that, in order to understand the phylogenetic history of luteoviruses, it is necessary to analyse genes in the 5' and 3' halves of the genome. For instance, analysis of the RdRp, which is near the 5' terminus, has unmasked potential recombination between the CP of PAV isolate ASL-1 and PAS. Findings that the species *Sugarcane yellow leaf virus* has undergone recombination with PAV in this same genomic region (Smith et al., 2000) lend some support to this hypothesis, but it is also possible that there has been convergence of the ASL-1 CP to a PAS-like sequence. This study has also found inconsistent evolutionary relationships among MAV, PAV and PAS isolates. In the RdRp phylogeny, MAV and PAV form a monophyletic clade independent of CN-PAV and PAS, which also form a monophyletic clade. In the MP phylogeny, however, PAV and PAS are sister taxa, CN-PAV is sister to both species and MAV is sister to all three virus types. Analysis of the complete-genome sequences from a number of species in the family *Luteoviridae* supported the topology of the RdRp tree. It could be that the position of MAV in the CP is the result of the loss or change in genetic motifs that govern transmission by *R. padi*. CP sequences that regulate aphid species-specific virus transmission have not yet been uncovered, but studies have demonstrated that vector selection does exert evolutionary pressure on luteovirus populations (Power & Gray, 1995; Lucio-Zavaleta et al., 2001).

Geographical subdivision has been reported for the global populations of *Tomato spotted wilt virus* (TSWV) (Tosmpana et al., 2005), *Cucurbit yellow stunting disorder virus* (Rubio et al., 2001) and *Turnip mosaic virus* (Tomimura et al., 2003), as well as for regional populations of *Rice yellow

![](https://www.microbiologyresearch.org)

**Table 4.** Population statistics and neutrality tests based on variation in the RdRp, CP and MP of PAV and PAS isolates collected in New York and Morocco

<table>
<thead>
<tr>
<th>Geographical region</th>
<th>Gene</th>
<th>Species</th>
<th>Population statistics</th>
<th>Tests of neutrality†</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>l</td>
<td>n</td>
</tr>
<tr>
<td>New York</td>
<td>RdRp</td>
<td>PAV</td>
<td>654</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>RdRp</td>
<td>PAS</td>
<td>654</td>
<td>7</td>
</tr>
<tr>
<td>New York</td>
<td>CP</td>
<td>PAV</td>
<td>603</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>PAS</td>
<td>603</td>
<td>7</td>
</tr>
<tr>
<td>Morocco</td>
<td>CP</td>
<td>PAV</td>
<td>603</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>PAS</td>
<td>603</td>
<td>7</td>
</tr>
<tr>
<td>New York</td>
<td>MP</td>
<td>PAV</td>
<td>462</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>MP</td>
<td>PAS</td>
<td>462</td>
<td>7</td>
</tr>
<tr>
<td>Morocco</td>
<td>MP</td>
<td>PAV</td>
<td>462</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>MP</td>
<td>PAS</td>
<td>462</td>
<td>7</td>
</tr>
</tbody>
</table>

*l*, Sequence length; *n*, sample size; *s*, no. segregating sites; *k*, mean no. nucleotide differences in pairwise sequence comparisons (Tajima, 1983); $\theta$, population mean mutation rate per site (Watterson’s $\theta$ estimator) is a function of the number of segregating sites and the sample size.

†Tajima’s D and Fu & Li’s D* and F* tests measure the departure from neutrality for all mutations in a genomic region (Tajima, 1989; Fu & Li, 1993). Values for neutrality tests were all not significant ($P > 0.1$).

<table>
<thead>
<tr>
<th>Region</th>
<th>Species</th>
<th>RdRp</th>
<th>CP</th>
<th>MP</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York</td>
<td>PAV</td>
<td>0.052</td>
<td>0.055 (0.025)</td>
<td>1.0 (0.32)</td>
</tr>
<tr>
<td></td>
<td>PAS</td>
<td>0.035</td>
<td>0.44 (0.07)</td>
<td>0.71 (0.14)</td>
</tr>
<tr>
<td>Morocco</td>
<td>PAV</td>
<td>*</td>
<td>0.11 (0.03)</td>
<td>1.05 (0.29)</td>
</tr>
<tr>
<td></td>
<td>PAS</td>
<td>*</td>
<td>0.30 (0.05)</td>
<td>0.025 (0.008)</td>
</tr>
</tbody>
</table>

*No sequences available for analysis.

**Table 5.** Relative selective constraints [dN/(dS + constant)] calculated for the RdRp, CP and MP of PAV and PAS isolates collected in New York and Morocco

*PAML* (Yang, 1997) was used to estimate dN and dS values from pairwise comparisons of sequences in a population by using the method of Nei & Gojobori (1986). Freidman’s non-parametric test of population-central values was used to compare relative selective constraint. Values with the same letter were included in the same analysis and capitalization denotes values found to be significantly different ($P < 0.05$). Numbers in parentheses are the standard errors of estimates.
mottle virus (RYMV) (Pinel et al., 2000), Sweet potato chlorotic stunt virus (Alicai et al., 1999) and Kenneyda yellow mosaic virus (Skotnicki et al., 1996). Several of these studies invoke virus adaptation in response to new hosts or transmission modalities to explain the observed pattern of variation, but few present evidence that geographical variants differ in their biology or that there is a correlation between positively selected codon sites and the geographical region from where the population was sampled. In the absence of such evidence, one could also attribute geographical subpopulation structure to transmission bottlenecks followed by genetic drift. As selection and genetic drift can both lead to a decrease in diversity within populations and an increase in diversity between populations, it is often difficult to distinguish the effects of these processes. However, selection and drift need not be mutually exclusive forces. A study by Choi et al. (2001) found considerable divergence between the American and Mexican populations of Wheat streak mosaic virus (WSMV), mostly at synonymous nucleotide sites. They concluded from these data that negative selection and drift acting simultaneously have contributed to the evolution of WSMV strains. The present study found significant genetic differentiation between the CP of the Moroccan and New York PAS populations, but no support for the hypothesis that either population deviates from a neutral model of evolution. Thus, it appears that for PAS, like WSMV, genetic drift rather than adaptation is responsible for divergence between the two populations.

Genetic diversity and selective constraint

Total nucleotide diversity was estimated to be 0.028 and 0.029 in the CP and 0.017 and 0.020 in the RdRp for the New York PAV and PAS populations, respectively. These values are similar to or lower than total nucleotide diversity values reported for other insect-transmitted plant RNA viruses, such as 0.035 for TSWV (Tsompana et al., 2005), 0.068 for Citrus tristeza virus, 0.07 for Groundnut rosette virus and 0.194 for RYMV (Garcia-Arenal et al., 2001). Diversity did not differ significantly across genes, species or geographical regions. The only exception was the Moroccan PAS population, which has much lower nucleotide diversity in its CP (0.006) when compared with the Moroccan PAV population (0.028) or the New York PAS population (0.025). This study did not find a significant difference in selective constraint on the New York or Moroccan PAS CP. Thus, it is likely that the Moroccan population has been through a genetic bottleneck more recently than the New York population, such that there has been less time for population expansion and re-establishment of diversity. In New York, selective constraint did not differ between the PAV CP (0.055), the PAV RdRp (0.052) and the PAS RdRp (0.035), but was significantly greater for the PAS CP (0.44). This indicates that genetic variation in either gene class is not necessarily more constrained due to the function of the encoded protein. I also found that selective constraint does not differ between the PAS CP in New York and Morocco and it is less constrained than the PAV CP in both environmental contexts. Because PAV and PAS are sympatric in these locations and share the same vector specificity, it might be the case that relaxed selective constraints on the PAS CP are due to differences in plasticity of the protein rather than differences in the source or magnitude of the selection pressure acting on the virus population.

Utilizing statistical models that allow heterogeneous $\omega$ ratios among codon sites, researchers have found evidence of diversifying selection in genes of CMV (Moury, 2004), Potato virus Y (Moury et al., 2002) and TSWV (Tsompana et al., 2005). Positively selected amino acid sites were detected in structural proteins of each of these viruses and in transcription-related proteins of CMV and TSWV. For BYDV-PAV and PAS, the MP was under the least selective constraint when compared with genes in other functional classes, except in the Moroccan PAS population, where it seemed to be more constrained than any gene sampled from any other population. As discussed above, this could be the result of a recent population bottleneck. Four codon sites in the MP are putatively affected by positive selection. These codon sites are located in the C-proximal part of the protein, which is known to be the nucleic acid-binding domain of the potato leaf roll virus (PLRV) (family Luteoviridae, genus Polerovirus) MP (Tacke et al., 1991). The other site is located in the N-terminal part of the protein. Experimental manipulation of the PLRV nucleotide sequence suggests that protein–protein interactions occur in this domain (Tacke et al., 1993). In interpreting the results of the PAML analysis, it is necessary to keep in mind that synonymous sites in the MP correspond to non-synonymous sites in the CP. This may influence estimates of $\omega$ for the MP by reducing the rate of synonymous substitutions. Relatively high $\omega$ estimates have been shown for overlapping reading frames in CMV (Moury et al., 2002) and PLRV (Guyader & Ducray, 2002).

In summary, the present study found that PAV and PAS are related more distantly with respect to their RdRp than their CP genes. It is not well understood how genetic variation in the BYDV RdRp relates to virus accumulation in host tissues, host range and disease severity. Future research must address this issue, as the RdRp is a significant source of genetic variation at the intra- and interspecies levels. The results of this study indicate that genetic variation in the RdRp is not correlated with the aphid-transmission phenotype of the virus, because analysis of the RdRp gene alone cannot differentiate between BYDV species. This reflects the different selection pressures acting on the RdRp and CP. Despite the different roles that they play in the virus infection cycle, selective constraint experienced by the RdRp or CP did not differ based solely upon their function. Rather, functional constraint was determined by gene function, phylogenetic history and selection pressures in a particular environment. Given our findings that New York PAS populations are genetically distinct from the Moroccan populations, it should not be assumed that isolates of the same species in different geographical regions will have
similar biological characteristics. It may be necessary to independently evaluate the infection properties of isolates present in each area where BYDV hinders grain production. This information will aid growers in the development of grain varieties with the appropriate resistance and will link features in the viral genome to virus-transmission dynamics and effects on host fitness.

ACKNOWLEDGEMENTS

I thank Micheal Milgroom for reviewing previous drafts of this manuscript, Alison Power for providing materials and facilities necessary for maintaining viruses culture plants and aphid colonies, and Kai Blaisdell, Steve Bogdanowicz, Erik Dogman and Damon Little for their assistance in developing molecular protocols and performing cladistic analyses. All molecular work was performed in the Evolutionary Genetics Core Facility in the Department of Ecology and Evolutionary Biology at Cornell University. Research was supported by the National Institutes of Health Ruth L. Kirschstein individual research fellowship and the Andrew W. Mellon student research grant.

REFERENCES


