Molecular Ecology (2005) **14**, 53–66 doi: 10.1111/j.1365-294X.2004.02392.x

# The molecular population genetics of the *Tomato spotted wilt virus* (TSWV) genome

M. TSOMPANA,* J. ABAD,* M. PURUGGANAN† and J. W. MOYER*
**Department of Plant Pathology, North Carolina State University, 2518 Gardner Hall, Raleigh, NC 27695–7616,*
*†Department of Genetics, North Carolina State University, 2618 Gardner Hall, Raleigh, NC 27695–7614*

## Abstract

**RNA viruses are characterized by high genetic variability resulting in rapid adaptation to new or resistant hosts. Research for plant RNA virus genetic structure and its variability has been relatively scarce compared to abundant research done for human and animal RNA viruses. Here, we utilized a molecular population genetic framework to characterize the evolution of a highly pathogenic plant RNA virus [*Tomato spotted wilt virus* (TSWV), *Tospovirus*, *Bunyaviridae*]. Data from genes encoding five viral proteins were used for phylogenetic analysis, and for estimation of population parameters, subpopulation differentiation, recombination, divergence between *Tospovirus* species, and selective constraints on the TSWV genome. Our analysis has defined the geographical structure of TSWV, attributed possibly to founder effects. Also, we identify positive selection favouring divergence between *Tospovirus* species. At the species level, purifying selection has acted to preserve protein function, although certain amino acids appear to be under positive selection. This analysis provides demonstration of population structuring and species-wide population expansions in a multisegmented plant RNA virus, using sequence-based molecular population genetic analyses. It also identifies specific amino acid sites subject to selection within *Bunyaviridae* and estimates the level of genetic heterogeneity of a highly pathogenic plant RNA virus. The study of the variability of TSWV populations lays the foundation in the development of strategies for the control of other viral diseases in floral crops.**

*Keywords*: *Bunyaviridae*, founder effect, mutation rate, population expansion, selection, TSWV

*Received 22 July 2004; revision received 29 September 2004; accepted 29 September 2004*

## Introduction

RNA viruses are characterized by high genetic variability, attributed to error-prone replication, high replication rates, short generation time and large population size (Domingo & Holland 1997). The high diversity of RNA viruses explains the emergence of new strains from old viruses and the ability to rapidly adapt to new or resistant hosts (Feuer *et al.* 1999). It is believed that the extent of viral variation is limited by the interaction of viral genes with plant proteins for the completion of an infection cycle (Schneider & Roossinck 2001). In plants, continuous and successful infection by RNA viruses results in great economic losses because of suppressed growth, yield and product quality (Goldbach & Peters 1994; Moyer 2000). The study of the variability in the genetic structure of plant RNA virus populations is crucial for the better understanding of plant virus evolution and plant–virus interactions. Additionally, it is relevant to the development of strategies for the long lasting control of virus-induced diseases in plants (Garcia-Arenal *et al.* 2001).

Research for plant RNA virus genetic structure and its variability has been relatively scarce compared to abundant research done for human and animal RNA viruses (Garcia-Arenal *et al.* 2001). A limited number of studies with plant RNA viruses have examined population genetic structure using monoclonal antibodies and molecular markers (Aranda *et al.* 1993; Kurath & Dodds 1995; Fraile *et al.* 1997; Alicai *et al.* 1999; Albiach-Marti *et al.* 2000; Rubio *et al.* 2001; Vives *et al.* 2002). Very few studies have used a sequence-based molecular population genetic framework to estimate population parameters (Arboleda & Azzam 2000; Moya *et al.* 1993; Fraile *et al.* 1996; Ayllon *et al.* 1999; Azzam *et al.* 2000a; 2000b; Fargette *et al.* 2004). Also, although

Correspondence: J. W. Moyer, Fax: 919 5157716; E-mail: james_moyer@ncsu.edu.

factors that shape genetic structure such as founder effects and selection have been suggested (Albiach-Marti *et al*. 2000; Kurath & Dodds 1995; Kofalvi *et al*. 1997; Ayllon *et al*. 1999; Azzam *et al*. 2000b; Choi *et al*. 2001), only a few cases have been rigorously tested (Fargette *et al*. 2004; Fraile *et al*. 1997; Guyader & Ducray 2002; Moury *et al*. 2002; Li & Roossinck 2004).

The *Bunyaviridae* is one of the largest RNA viral families, consisting of human, animal and plant viruses. It consists of five distinct genera with tripartite, ambisense, single-stranded (ss) genome: *Bunyavirus*, *Phlebovirus*, *Nairovirus*, *Hantavirus* and *Tospovirus*. Viruses in this family are distinguished by high mortality in their respective hosts. In addition, viruses in the *Tospovirus*, *Bunyavirus*, *Phlebovirus*, and *Nairovirus* genus are characterized by the ability to replicate in the mammal or plant as well as their insect hosts. The large number of viruses (more than 300) classified into *Bunyaviridae* is evidence of their evolutionary success and potential (Elliott 1996). Viruses in the *Tospovirus* genus of *Bunyaviridae* infect plants and their thrips vector (Moyer 2000). *Tospoviruses* have a tripartite RNA genome, composed of the large (L) segment (~9 kb), encoding an RNA-dependent RNA polymerase in negative sense, the medium (M) segment (~4.8 kb), encoding the NSm protein and the G1-G2 precursor glycoprotein in ambisense orientation, and the small (S) segment (~3 kb), encoding the NSs nonstructural protein and the N protein also in ambisense orientation (Fig. 1) (Moyer 1999). The NSm protein is involved in cell-to-cell movement (Kormelink *et al*. 1994) and the G1 and G2 glycoproteins play an important role in maturation/assembly of virions in both plants and thrips (Bandla *et al*. 1998). The NSs structural protein has RNA silencing suppressor activity (Takeda *et al*. 2002) and the N protein encapsidates the viral RNA within the viral envelope (Richmond *et al*. 1998). *Tomato spotted wilt virus* (TSWV) is the type member of the *Tospovirus* genus (Moyer 1999). TSWV is cosmopolitan and its host range exceeds 900 plant species, spanning monocots and dicots, and 10 thrips species (Moyer 2000; Ullman *et al*. 2002).

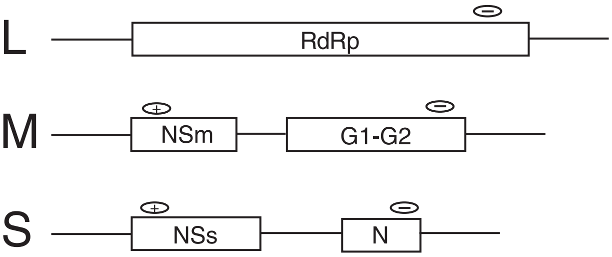

**Fig. 1** Genomic organization of TSWV. L, M, and S represent the three ss RNA genomic segments of the virus. L RNA (~9 kb) encodes an RNA-dependent RNA polymerase in negative sense (–). M RNA (~4.8 kb) encodes the NSm (+) protein and the G1-G2 (–) precursor glycoprotein. S RNA (~3 kb) encodes the NSs (+) nonstructural protein and the nucleocapsid protein, N (–). Both M and S RNAs are in ambisense orientation; ORFs of opposite polarity on the same genomic segment.

Heterogeneity and rapid adaptability are two prominent phenotypic characteristics that distinguish TSWV from other plant viruses. Genetic reassortment (Qiu *et al*. 1998) and multiplication in their hosts increase prospects for genetic heterogeneity in TSWV populations. These attributes make TSWV a good model for addressing specific evolutionary questions, such as defining the genetic structure of viral populations and the factors that shape it. However, until now these aspects of evolutionary biology of TSWV have not been studied and cannot be determined by the already available sequence data. Reported evolutionary TSWV, *Tospoviruses* and in general *Bunyaviridae* research has focused primarily on phylogenetic/taxonomic analysis or measures of genotypic variation (de Avila *et al*. 1993; Dewey *et al*. 1997; Pappu *et al*. 1998; Heinze *et al*. 2001; Sironen *et al*. 2002; Yashina *et al*. 2003). Limited research indicates that some viruses in *Bunyaviridae* evolve through mutations, antigenic shift (reassortment and recombination) (Plyusnin 2002) and host switching (Bohlman *et al*. 2002; Nemirov *et al*. 2002). There remains a need for statistically validated descriptions of the genetic structure of natural virus populations that can be used as the basis for investigation of the evolutionary forces acting on these viral species.

Here we utilize a molecular population genetic approach for a large number of host and geographically diverse isolates to study the genetic structure of TSWV and elucidate potential factors that shape variation in this viral species. We suggest a metapopulation model for TSWV, with a defined geographical structure attributed possibly to founder effects. The model is supported by phylogenetic analyses, significant genetic differentiation between geographical groups of isolates, an overall lack of nucleotide variation and high haplotype diversity within subpopulations, and deviations from the neutral equilibrium model for the five analysed genes. In addition, we demonstrate positive selection favouring divergence between *Tospovirus* species, and identify specific codon positions that are potential sites of positive selection pressure within the TSWV species. We believe that identification of the selection pressures acting on the TSWV genome together with functional/structural data will reveal accurate estimates of the biological function of certain genes or pieces of genes.

## Materials and methods

### Virus isolates

The 20 isolates sequenced in this paper were obtained from a range of geographical locations both in the United States [California (CA), Colorado (CO), North Carolina (NC)] and Europe (Spain). They arose from a variety of host plants (seed and vegetatively propagated), both from greenhouse and field environments, in the time period 2000–2002. Isolates from same locality have been collected at different times/years. Information about these isolates is provided in the Appendix. Stock cultures of all isolates were stored as infected *Nicotiana benthamiana* leaves at –80 °C.

### *RNA extraction, RT-PCR, and sequencing*

Total plant RNA was extracted from TSWV systemically infected *N. benthamiana* leaves (Qiu & Moyer 1999), after the first mechanical inoculation. RNA extracts were used as template for reverse transcription-polymerase chain reaction (RT-PCR). Primers were designed, according to TSWV sequences published in GenBank, to amplify overlapping regions of the S and M RNAs. Primers and their corresponding annealing temperature are given in the Appendix. The cDNA was synthesized in a 50-μL reaction volume with AMV (avian myeloblastosis virus) reverse transcriptase (Promega) (Law & Moyer 1990). PCR amplification was performed in a 50-μL reaction solution containing 1× *Taq* Buffer, 2.5 mM $MgCl_2$, 0.25 mM dNTP (each), 0.4 μM of the viral sense and complementary sense primer, with AmpliTaq DNA polymerase (Applied Biosystems) and 5 μL cDNA template. Thermal cycling reactions were carried out for 40 cycles of amplification, with each cycle consisting of 94 °C denaturation for 1 min, annealing at different temperatures (see Appendix) for 1 min, extension at 72 °C for 1 min, with 94 °C for 10 min at the beginning and 72 °C for 10 min at the final step. PCR products were electrophoresed in 0.8% agarose and stained with ethidium bromide. Qiagen PCR purifica-tion columns (Qiagen, Inc.) were used to purify the PCR products. Direct sequencing was carried out using BigDye terminator reactions (Applied Biosystems) run on an ABI 377 automated sequencer. Nucleotide consensus sequences were assembled and edited using VECTOR NTI (InforMax Inc.) and aligned using CLUSTALX version 1.81 (Thompson *et al*. 1997). Codon-based alignments were adjusted manually according to the amino acid sequence alignment using GENEDOC (Nicholas & Nicholas 1997). The sequences reported in this paper have been submitted to GenBank (accession numbers AY744468–AY744495).

### *Phylogenetic, population genetic and structural analysis*

Thirteen complete S and M RNA consensus sequences from 18 isolates, as well as G1-G2 consensus sequences of two additional isolates were obtained. Analyses were conducted using these sequences combined with 27 TSWV sequences from GenBank and four sequences previously sequenced from our laboratory (see Appendix). Analyses were conducted separately for each coding region of the S and M RNAs. Overall, for the NSs (1404 bp), N (777 bp), NSm (912 bp) and G1-G2 (3426 bp), we used 21, 41, 17 and 18 sequences, respectively. PAUP, version 4.0 beta 4 (Swofford 2000) was used to construct phylogenetic trees using a maximum-parsimony criterion (branch and bound search, stepwise addition of sequences). Bootstrap confidence limits were obtained by 1000 replicates. Neighbour-joining (NJ) phylogenies, based on the Kimura 2-parameter distance matrix were generated by MEGA version 2.01 (Kumar *et al*. 1993). Bootstrap confidence limits were also obtained by 1000 replicates.

DNASP version 3.51 (Rozas & Rozas 1999) was used to estimate Watterson's estimator of Θ (Θw) (Watterson 1975), the average pairwise nucleotide diversity π (Tajima 1983), and the haplotype diversity. This program was also used to examine selection on proteins between viral species with the McDonald and Kreitman test of neutrality ( 1991), and to run Tajima's D (Tajima 1989) and Fu and Li's D and F (1993) tests of neutrality.

The NNPREDICT program was used to predict the secondary structure for each amino acid residue of the NSs, N and G1-G2 coding regions (Kneller *et al*. 1990).

### *Tests of population differentiation*

Three independent statistical tests of population differentiation were run. Specifically, PERMTEST program (Hudson *et al*. 1992) was used to estimate $K_{ST}$ test statistic of genetic differentiation, using NJ trees as input data. $K_{ST}$ is equal to $1 - K_S/K_T$, where $K_S$ is a weighted average of the $K_1$ and $K_2$ (average number of differences between sequences from within subpopulations 1 and 2, respectively) and $K_T$ represents the average number of differences between two sequences regardless of their subpopulation. Under the null hypothesis (no genetic differentiation) we expect $< K_{ST} >$ (observed value of $K_{ST}$) to be near zero and so we reject the null if $< K_{ST} >$ is supported by a small *P*-value ($< 0.05$). Also, small observed values of $K_S$ lead to rejection of the null hypothesis ($P < 0.05$). Additionally, DNASP version 4.0 (Rozas *et al*. 2003) was used to estimate Z and Snn test statistics of genetic differentiation. Z is a weighed sum of Z1 and Z2, where Zi is the average of the ranks of all the $d_{ij,lk}$ values for pairs of sequences from within locality i (Hudson *et al*. 1992). The null hypothesis of no genetic differentiation is rejected if the Z statistic is too small, supported by a *P*-value less than 0.05. Finally, Snn measures the frequency that the nearest neighbours of sequences are found in the same locality (Hudson 2000). Snn test statistic values may range between one, when populations from different localities are genetically distinct, to one-half in the case of panmixia. Statistical significance for all above tests was established using 1000 permutations. None of the tests of genetic differentiation makes assumptions about collection time of samples, and can thus be reliably used to test genetic differentiation of isolates collected at different time intervals. Also, DNASP version 4.0 (Rozas *et al*. 2003) was used to estimate the coefficient of gene differentiation $F_{ST}$ for all subpopulations and loci (Wright 1951). $F_{ST}$ measures the amount of interpopulation diversity and takes values between zero and one.

### *Test of recombination*

Variations in the estimation of phylogenies and detection of positive selection arise from perturbations resulting from recombination events. Detection of recombination events was carried out by the LDHAT algorithm. LDHAT implements a coalescent-based method to estimate recombination from gene sequences (McVean *et al*. 2002) based on Hudson's (2001) composite-likelihood method (Hudson 2001), extended to allow for finite-sites mutation models. LDHAT estimates the proportion of permutated data sets that have a composite likelihood equal to or greater than the original estimated ($P_{LPT}$). The null hypothesis is rejected if $P_{LPT}$ is less than a given threshold value (< 0.05), and the alternative hypothesis of recombination is accepted.

### *Identification of selection pressure*

The CODEML algorithm from the PAML package was used to test whether variable selective pressures occur on amino acids of the coding regions (Yang *et al*. 2000). This likelihood-based approach identifies selection acting on specific codons. Statistical significance for the model is established using the likelihood-ratio test (Yang *et al*. 2000). We used comparisons between two sets of maximum-likelihood (ML) models to test for positive selection. These models assume that $d_N/d_S$ (ratio of the rate of nonsynonymous over the rate of synonymous substitutions) is constant across all lineages but varies across codons. The first comparison involved models M2 and M3 and the second models M7 and M8. Model M2, assumes codons are invariant ($d_N/d_S = 0$), neutral ($d_N/d_S = 1$), or have a floating $d_N/d_S$ ratio estimated from the data. Model M3, assumes that codons have one of the three different $d_N/d_S$ ratios estimated from the data, any of which can be greater than one. Model M7 utilizes a discrete beta distribution with 10 categories of $0 \le d_N/d_S \le 1$. Finally, M8 uses a discrete beta distribution with one more category of $d_N/d_S > 1$. If either M3 or M8 are statistically favoured over M2 and M7, respectively, and depict codons with $d_N/d_S > 1$, we conclude that positive selection has occurred.

## Results

### *Phylogenetic analyses support geographical structuring for isolates of each ORF*

To examine patterns of genetic differentiation among isolates, phylogenetic trees were constructed using maximum parsimony from the nucleotide polymorphisms observed in NSs (21 sequences), N (41 sequences), NSm (17 sequences) and G1-G2 (18 sequences) coding regions. None of the nucleotide sequences of haplotypes appears to represent an intragenic recombinant. Specifically for all genes, we estimated the $P_{LPT}$ value, which corresponds to the proportion of permutated data sets that have a composite likelihood equal to or greater than the original estimate. The $P_{LPT}$ value for the NSs, N, NSm and G1-G2 genes were 0.36, 0.92, 0.93 and 0.60, respectively, and are not statistically significant ($P > 0.05$).

The NSs phylogenetic tree (Fig. 2), based on 21 sequences, delineated four geographical subpopulations; one from North Carolina, a second from Spain, a third from California and a fourth containing isolates from Bulgaria, Netherlands and Germany. Interestingly, the last subpopulation also included two isolates that have been collected in North Carolina (NC1 and NC3). The N gene phylogenetic tree (41 sequences) (Fig. 3) uncovered five geographical subpopulations with isolates from: (i) Georgia and North Carolina (ii) California (iii) Spain (iv) Japan and (v) Bulgaria, Netherlands, and Germany. The last subpopulation included the two isolates collected in North Carolina and one isolate collected in South Africa.

Analyses of M RNA were consistent with the analyses of S RNA coding regions. The NSm gene phylogenetic tree (17 sequences) (Fig. 4) uncovered four geographical subpopulations including isolates from: (i) California (ii) North Carolina (iii) Spain and (iv) Netherlands and Bulgaria. Finally, the G1-G2 gene phylogenetic tree (18 sequences)

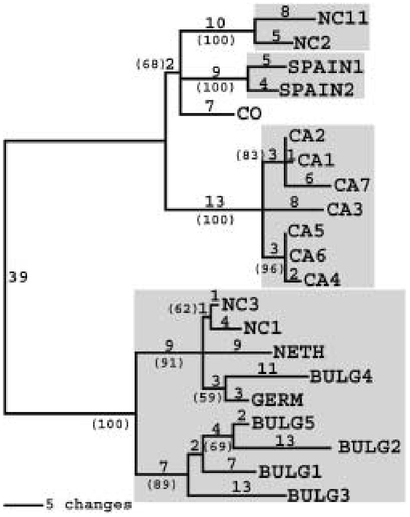

**Fig. 2** Parsimony phylogenetic tree (branch and bound search, stepwise addition of sequences) constructed from nucleotide polymorphisms observed at the NSs gene (21 sequences). Values on branches indicate number of polymorphisms separated by inferred intermediate haplotypes. Bootstrap values more than 50% are indicated in parentheses. Grey-shaded boxes highlight the independent geographical subpopulations.

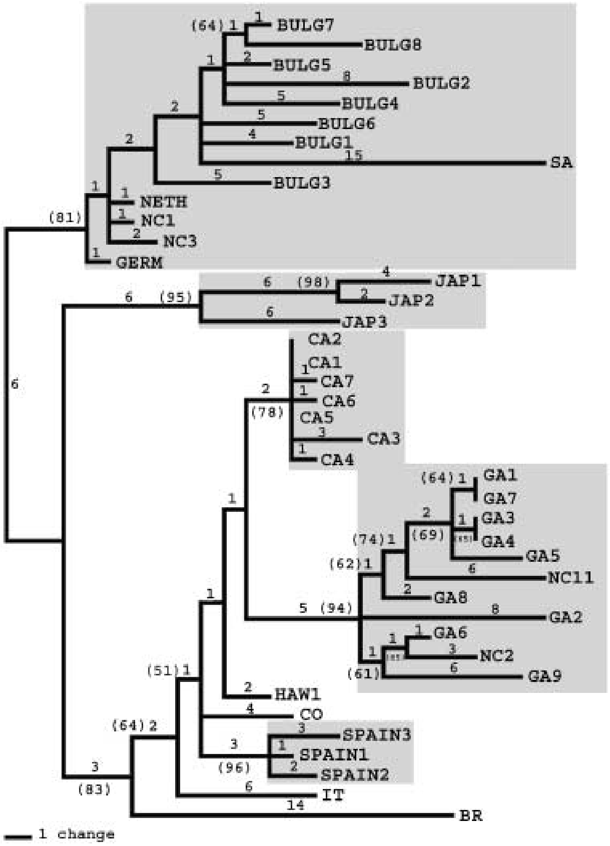

Fig. 3 Parsimony phylogenetic tree (branch and bound search, stepwise addition of sequences) constructed from nucleotide polymorphisms observed at the N gene (41 sequences). Values on branches indicate number of polymorphisms separated by inferred intermediate haplotypes. Bootstrap values more than 50% are indicated in parentheses. Grey-shaded boxes highlight the independent geographical subpopulations.

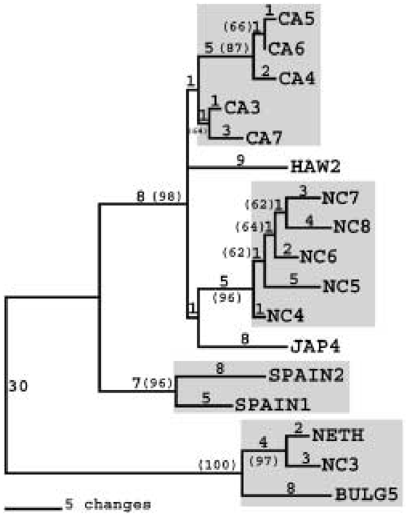

Fig. 4 Parsimony phylogenetic tree (branch and bound search, stepwise addition of sequences) constructed from nucleotide polymorphisms observed at the NSm gene (17 sequences). Values on branches indicate number of polymorphisms separated by inferred intermediate haplotypes. Bootstrap values more than 50% are indicated in parentheses. Grey-shaded boxes highlight the independent geographical subpopulations.

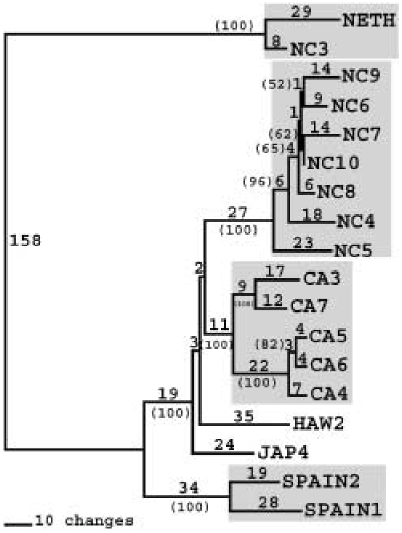

Fig. 5 Parsimony phylogenetic tree (branch and bound search, stepwise addition of sequences) constructed from nucleotide polymorphisms observed at the G1-G2 gene (18 sequences). Values on branches indicate number of polymorphisms separated by inferred intermediate haplotypes. Bootstrap values more than 50% are indicated in parentheses. Grey-shaded boxes highlight the independent geographical subpopulations.

(Fig. 5) yielded four subpopulations consisting of isolates from: (i) North Carolina (ii) California (iii) Spain and (iv) Netherlands. For both the NSm and G1-G2 phylogenetic trees, again one isolate collected in North Carolina (NC3) clusters within the Netherlands/Bulgaria subpopulation. Each geographical subpopulation consists of isolates collected at different times indicating that time is not a factor affecting the geographical clustering of isolates. With the exception of the two isolates collected in North Carolina (NC1 and NC3), and one collected in South Africa, all other isolates tend to cluster according to their geographical origin, and the affinities are consistent for all coding regions of S and M RNAs. The possible cause of the NC1, NC3 and South Africa clustering with the European isolates will be discussed later.

### *Significant genetic differentiation between geographical subpopulations for all coding regions of TSWV S and M RNAs*

To test if geographical isolates are genetically differentiated populations, three independent statistical tests of population

**Table 1** Summary of test statistics and parameter estimates examined for population differentiation between geographical isolates

| Gene | Comparison between geographical groups | $K_{ST}$† | $K_S$† | *P*-value | Z† | *P*-value | Snn† | *P*-value | $F_{ST}$‡ | $F_{ST}$‡ |
|---|---|---|---|---|---|---|---|---|---|---|
| NSs | CA vs. Bulgaria | 0.689 | 0.010 | 0.000* | 15.789 | 0.003* | 1.000 | 0.000* | 0.785 | |
| | NC vs. Bulgaria | 0.184 | 0.026 | 0.005* | 13.047 | 0.004* | 0.889 | 0.003* | 0.277 | 0.758§ |
| | Spain vs. Bulgaria | 0.587 | 0.014 | 0.037* | 5.500 | 0.046* | 1.000 | 0.000* | 0.770 | |
| | CA vs. NC | 0.326 | 0.017 | 0.007* | 18.020 | 0.005* | 1.000 | 0.000* | 0.430 | 0.680§ |
| | CA vs. Spain | 0.544 | 0.006 | 0.028* | 10.452 | 0.036* | 1.000 | 0.000* | 0.738 | |
| N | GA vs. CA | 0.457 | 0.007 | 0.000* | 32.585 | 0.000* | 1.000 | 0.000* | 0.630 | |
| | GA vs. Spain | 0.394 | 0.009 | 0.004* | 18.977 | 0.007* | 1.000 | 0.000* | 0.645 | |
| | GA vs. Japan | 0.474 | 0.011 | 0.004* | 19.443 | 0.007* | 1.000 | 0.000* | 0.662 | |
| | GA vs. Bulgaria | 0.508 | 0.012 | 0.000* | 31.711 | 0.000* | 1.000 | 0.000* | 0.651 | |
| | CA vs. Spain | 0.590 | 0.003 | 0.008* | 11.905 | 0.019* | 1.000 | 0.000* | 0.710 | |
| | CA vs. Japan | 0.612 | 0.006 | 0.014* | 12.000 | 0.011* | 1.000 | 0.000* | 0.700 | |
| | CA vs. Bulgaria | 0.565 | 0.008 | 0.000* | 23.367 | 0.000* | 1.000 | 0.000* | 0.712 | |
| | Spain vs. Bulgaria | 0.427 | 0.011 | 0.005* | 14.311 | 0.003* | 1.000 | 0.000* | 0.663 | |
| | Japan vs. Bulgaria | 0.367 | 0.014 | 0.005* | 15.162 | 0.005* | 1.000 | 0.000* | 0.559 | |
| NSm | CA vs. NC | 0.190 | 0.018 | 0.008 * | 17.762 | 0.005* | 0.909 | 0.002* | 0.317 | 0.564§ |
| | CA vs. Spain | 0.487 | 0.009 | 0.045* | 4.500 | 0.045* | 1.000 | 0.000* | 0.620 | |
| | NC vs. Spain | 0.244 | 0.023 | 0.040* | 11.333 | 0.270 | 0.875 | 0.000* | 0.464 | 0.667§ |
| G1-G2 | NC vs. CA | 0.263 | 0.019 | 0.001* | 25.162 | 0.001* | 0.923 | 0.000* | 0.434 | 0.620§ |
| | NC vs. Spain | 0.263 | 0.022 | 0.020* | 17.714 | 0.126 | 0.900 | 0.000* | 0.536 | 0.692§ |
| | CA vs. Spain | 0.474 | 0.015 | 0.045* | 4.950 | 0.045* | 1.000 | 0.000* | 0.635 | |

*$P < 0.05$, determined using 1000 permutations; †$K_{ST}$, $K_S$, Z and Snn are test statistics of genetic differentiation; ‡$F_{ST}$ examines the extent of genetic differentiation between geographical isolates; §$F_{ST}$ re-estimated for pairwise comparisons including only the NC isolates that did not cluster within the Bulgaria/Germany/Netherlands subpopulation.

differentiation were applied (Hudson *et al*. 1992; Hudson 2000). The analysis focused on the largest geographical groups of isolates of each coding region, as the power of the tests increases as the subpopulation size also increases. The null hypothesis was rejected for the majority of comparisons, with all the test statistics ($K_{ST}$, $K_S$, Z, and Snn) supported by *P*-values less than 0.05 (Table 1). Only in two cases (comparisons between North Carolina and Spain for NSm and G1-G2 genes) the *Z*-test statistic was not significant. All other test statistics were significant for the latter comparisons, supporting genetic differentiation between the North Carolina and Spain geographical groups in agreement with the phylogenetic analyses (Figs 4 and 5). The failure of the *Z*-test to differentiate the two geographical groups may be attributed to lack of statistical power. Overall, these results indicate significant genetic differentiation between geographical groups of isolates.

Additionally, the coefficient of gene differentiation $F_{ST}$ was used to estimate the extent of genetic differentiation between geographical isolates. The overall values of $F_{ST}$ for the NSs, N, NSm and G1-G2 genes were 0.58, 0.66, 0.47 and 0.53, respectively, indicating considerable genetic differences between geographical isolates. $F_{ST}$ values calculated for each pair of geographical group of isolates for each gene are presented in Table 1. For the majority of pairwise comparisons, $F_{ST}$ values are high supporting population differentiation for TSWV. For example, the $F_{ST}$ value for the NSs California/Bulgaria comparison is 0.79 indicating that most of the molecular diversity is distributed among these geographical groups. Interestingly, in most cases where North Carolina isolates were compared against another geographical group (for example NSs: North Carolina/Bulgaria comparison and NSm: California/ North Carolina comparison) the $F_{ST}$ values are lower than all other comparisons. When pairwise comparisons were repeated using only the North Carolina isolates that did not cluster within the Bulgaria/ Germany/Netherlands subpopulation, the $F_{ST}$ values were as high as in every other comparison (Table 1). This result, together with the phylogenetic analyses presented before, demonstrates the occurrence of gene flow from the Bulgaria/ Germany/ Netherlands to the North Carolina subpopulation. The possible causes of gene flow will be discussed later. Overall, estimation of the test statistics of genetic differentiation and of the coefficient of genetic differentiation supports a defined geographical structure for TSWV.

### *Levels and patterns of intraspecific polymorphism for TSWV*

The level of intraspecific polymorphism of TSWV was examined by estimating genetic diversity (Θw and π) at all sites and silent sites only, for the NSs, N, NSm and G1-G2

**Table 2** Genetic variation of NSs, N, NSm and G1-G2 coding regions measured by Θw and π

| Gene | $\pi_{Total}$* | $\pi_{Silent}$* | $\Theta w_{Total}$† | $\Theta w_{Silent}$† |
|---|---|---|---|---|
| NSs | 0.034 | 0.091 | 0.037 | 0.091 |
| N | 0.024 | 0.079 | 0.043 | 0.132 |
| NSm | 0.030 | 0.113 | 0.036 | 0.131 |
| G1-G2 | 0.031 | 0.104 | 0.042 | 0.133 |

*π is estimated by the average pairwise difference among sequences in a sample; †Θw is a function of both the number of polymorphic sites and the number of sequences in a sample; *,†Θw and π were calculated based on the total number of sites ($\Theta w_{Total}$ and $\pi_{Total}$) as well as on the silent sites only ($\Theta w_{Silent}$ and $\pi_{Silent}$).

coding regions using the same sequence data as above. The level of diversity for all coding regions can be seen in Table 2.

The patterns of molecular diversity were evaluated using Tajima's D and Fu and Li's D and *F*-test statistics at segregating sites, and haplotype diversity and nucleotide diversity at all sites (Table 3). Tajima's D and Fu and Li's D and *F* statistics test the distribution of nucleotide polymorphisms in the genome. Negative Tajima's D and Fu and Li's D and *F*-values indicate an excess of low frequency polymorphism caused either by background selection, genetic hitchhiking or population expansions. Because selection events such as genetic hitchhiking and background selection affect relatively small fractions of the genome, a multilocus trend of negative Tajima's D and Fu and Li's D and *F*-test statistic values indicates that demographic forces are acting on the population (Hey & Harris 1999; Tajima 1989). For the majority of geographical groups across the genome values of Tajima's D and Fu and Li's D and *F*-test, statistics are negative (Table 3), indicating the occurrence of recent TSWV population expansions.

Levels of haplotype diversity and nucleotide diversity were also compared to determine if TSWV population expansions are taking place. Nucleotide diversity estimates the average pairwise differences among sequences, and haplotype diversity refers to the frequency and number of haplotypes in a sample. Estimates of nucleotide diversity can range from zero when no variation exists to 0.100 under cases of extreme divergence between alleles, whereas haplotype diversity values may vary between zero and 1.000 (Grant & Bowen 1998). The haplotype diversity and nucleotide diversity values for all subpopulations across loci are presented in Table 3. In most cases, haplotype diversity values are high and nucleotide diversity values are low. Specifically, haplotype diversity values across loci range from 0.857 to 1.000, and nucleotide diversity estimates range from 0.002 to 0.025. The elevated nucleotide diversity for the North Carolina subpopulation observed at the NSm and G1-G2 genes may be the result of the introduction of distantly related alleles from the Bulgaria/Germany/Netherlands subpopulation, as also indicated by the phylogenetic analyses. Overall, the deviations from the neutral equilibrium model for the five analysed genes, together with the combination of high haplotype diversity and overall lack of nucleotide diversity within individual geographical groups are consistent with a model of recent population expansion events.

### *Interspecific divergence between TSWV and Impatiens necrotic spot virus INSV*

Neutral theory predicts that the ratio of silent to replacement substitutions should be the same for polymorphisms within

**Table 3** Summary of parameter estimates and test statistics examined for demographic trends. Numbers in parentheses are standard deviations of estimates

| Gene | Geographic Group | Tajima's D* | Fu & Li's D† | Fu & Li's F† | Haplotype diversity | $\pi_{Total}$§ |
|---|---|---|---|---|---|---|
| NSs | CA | −0.495 | −0.705 | −0.725 | 0.952 (0.096) | 0.006 (0.001) |
| | Bulgaria | −0.532 | −0.714 | −0.734 | 1.000 (0.126) | 0.017 (0.003) |
| N | GA | −1.000 | −1.116 | −1.220 | 0.944 (0.070) | 0.010 (0.002) |
| | CA | −1.524* | −1.609* | −1.732* | 0.857 (0.137) | 0.002 (0.001) |
| | Bulgaria | −1.645* | −1.779* | −1.948* | 1.000 (0.063) | 0.013 (0.002) |
| NSm | CA | 0.304 | 0.304 | 0.323 | 1.000 (0.126) | 0.008 (0.002) |
| | NC | −1.169 | −1.266 | −1.366 | 1.000 (0.096) | 0.025 (0.011) |
| G1-G2 | NC | −1.569* | −1.722* | −1.884* | 1.000 (0.063) | 0.023 (0.011) |
| | CA | 0.228 | 0.228 | 0.248 | 1.000 (0.126) | 0.010 (0.002) |

*$0.05 < P < 0.10$; †Tajima's D compares the nucleotide diversity π with the proportion of polymorphic sites, which are expected to be equal under selective neutrality; ‡Fu and Li's *D*-test statistic is based on the differences between the number of singletons (mutations appearing only once among the sequences) and the total number of mutations. The *F*-test statistic is based on the differences between the number of singletons and the average number of nucleotide differences between pairs of sequences; §$\pi_{Total}$ is estimated by the average pairwise difference among sequences in a sample, based on all sites.

**Table 4** Examination of selection on proteins between viral species (TSWV and INSV) with the McDonald and Kreitman test for NSs, N, NSm and G1-G2 coding regions

| Gene | Fs[a] | Fr[a] | PS[b] | PR[b] | Fisher's | G-test | G-Williams | G-Yates |
|---|---|---|---|---|---|---|---|---|
| NSs | 221 | 266 | 101 | 63 | 0.0* | 0.0* | 0.0* | 0.0* |
| N | 126 | 165 | 94 | 40 | 0.0* | 0.0* | 0.0* | 0.0* |
| NSm | 123 | 132 | 88 | 23 | 0.0* | 0.0* | 0.0* | 0.0* |
| G1-G2 | 495 | 503 | 347 | 118 | 0.0* | 0.0* | 0.0* | 0.0* |

**Table 5** Comparison of the different ML models used in the NSs, N, NSm and G1-G2 analyses for detection of positive selected amino acid sites

| Gene | M2 vs M3 | M7 vs M8 | Positively selected sites | Amino acid Changes | $d_N/d_S$† |
|---|---|---|---|---|---|
| NSs | 14.663* | 15.238* | 458 | K → P | 0.367 |
| | | | | K → S | |
| | | | 460 | A → T | |
| | | | | A → G | |
| | | | | A → D | |
| N | 8.092* | 11.586* | 10 | S → N | 0.221 |
| | | | 174 | Y → T | |
| | | | | Y → C | |
| | | | 255 | T → A | |
| NSm | 0.002 | 0.598 | none | none | 0.082 |
| G1-G2 | 2.570 | 6.398* | 528 | I → A | 0.137 |
| | | | | I → V | |
| | | | 530 | R → D | |
| | | | | R → H | |
| | | | | R → N | |
| | | | 1002 | P → L | |
| | | | | P → S | |

*The models that allow for adaptive evolution (M3 or M8) are statistically favoured at the 95% confidence level; †Reported values of $d_N/d_S$ are based on the discrete (M3) ML model, which is the most sensitive in identifying positive selected sites.

species and fixed differences between species (McDonald & Kreitman 1991). Significant deviations from this expectation may lead to rejection of the null hypothesis of neutrality. This prediction forms the basis of the McDonald-Kreitman tests in Table 4. We compared intraspecific data from the NSs (21 sequences), N (41 sequences), NSm (17 sequences) and G1-G2 (18 sequences) coding regions using the INSV sequence, with GenBank accession number X66972, as interspecific data (de Haan *et al*. 1992). Comparisons using one sequence as interspecific data are routinely done and provide statistically reliable results. The total number of mutational events observed across TSWV and INSV revealed a significant excess of replacement fixed differences. Specifically, 266, 165, 132 and 503 replacement fixed differences were observed between the two species for the NSs, N, NSm and G1-G2 coding regions, respectively (Table 4). Within species (TSWV), there was an excess of synonymous substitutions. These results are statistically significant (Fisher's Exact Test, *G*-test, G-Williams Test and G-Yates Test, $P < 0.05$).

### *Evidence of selection in TSWV genes*

CODEML algorithm from PAML (Yang *et al*. 2000) was used to highlight variable selective constraints exerted on the NSs, N, NSm and G1-G2 genes. The first comparison involved models M2 and M3 and the second models M7 and M8. Notably, differences were observed between the selective constraints exerted on the NSm gene and the NSs, N and G1-G2 genes. For the NSs and N coding regions, the models accounting for codons with a $d_N/d_S$ value greater than 1 were significantly better ($P < 0.05$) than those that do not account for a $d_N/d_S > 1$ (Table 5). Change in NSs codon positions 458 (K→P or K→S) and 460 (A→T or A→G or A→D) were predicted to be driven by positive selection. Amino acid positions 10 (S→N), 174 (Y→T or Y→C), and 255 (T→A) in the N gene were under positive selection. In the G1-G2 coding region, comparison between models M2 and M3 did not favour any evidence for positively selected sites, whereas comparison between M7 and M8 marginally did ($P < 0.05$). The latter comparison predicted positive selection at sites 528 (I→A or I→V), 530 (R→D or R→H or R→N) and 1002 (P→L or P→S). Neither of the models predicted positive selection for codon sites in the NSm ORF.

Interestingly, all of the positively selected amino acid sites of the NSs, N and G1-G2 coding regions are predicted to reside either in helix or turn domains of the proteins. Specifically, the NSs protein amino acid positions 458 and 460 are predicted to reside in a helix and turn protein domain, respectively. For the N gene, positively selected amino acids 10 and 255 are predicted to be part of helices, whereas codon 174 resides in a turn. Finally, all positively selected sites of the G1-G2 coding region are found in turns.

## Discussion

Using sequence data from genes encoding five viral proteins, we utilized a molecular population genetic framework to examine the evolution of a highly pathogenic plant RNA virus (TSWV). Our specific goals were to determine the genetic structure of TSWV, identify the evolutionary forces that shape it and estimate the level of variation within this viral species. The analysis has defined the geographical structure of TSWV, attributed possibly to founder effects. Also, we demonstrate positive selection favouring divergence between *Tospovirus* species. At the species level, purifying selection has acted to preserve protein function, although certain amino acids appear to be under positive selection.

The phylogenetic analyses, together with the three independent tests of genetic differentiation and the estimation of $F_{ST}$ reveal that TSWV has a defined geographical structure. Only two isolates sampled from North Carolina (NC1 and NC3) and an isolate sampled from South Africa deviate from this prediction. These three isolates are genetically similar to the isolate from the Netherlands, demonstrating the occurrence of gene flow from Europe to North Carolina. It is possible that these 'foreign' isolates were introduced to other countries or states through importation of infected plant material from Netherlands, which is a centre of plant propagation material. Because North Carolina is the most well represented group, more exceptions like the ones noticed above may be identified if more isolates from other geographical areas are analysed. In addition, all analysed geographical subpopulations consist of isolates collected at different times/years indicating that time is not a factor affecting the geographical clustering of isolates. We thus demonstrate using a molecular population genetic framework that populations of a plant RNA virus are geographically structured.

At the population level, deviations from the neutral equilibrium model for the five analysed genes together with a combination of high haplotype diversity and overall lack of nucleotide diversity within individual geographical groups indicate the existence of population expansions for TSWV. It is important to note that the data set examined for each locus is not homogenous. Although this could slightly bias the current analyses, recent epidemiological data shows that the virus has been causing economically significant epidemics, supporting our findings on TSWV population expansions. The observed temporal loss of polymorphism within the subpopulations is expected following genetic drift, which accompanies bottleneck transmission (Novella *et al.* 1995). Evolutionary bottlenecks/founder effects may arise as a result of plant or vector associated effects (Bergstrom *et al.* 1999). Transmission bottlenecks by the thrips vector, may occur when only a few individuals are transmitted horizontally from one host to another in the initiation of an infection (Bergstrom *et al.* 1999). The high specificity required between thrips and virus for replication and transmission may necessitate adaptation of TSWV to the specific thrips biotypes, which could explain the emergence of geographical variants. In addition, TSWV population bottlenecks could result because of epidemiological factors, such as seasonal expansion and crashes of the plant and insect host populations. Overall, these analyses demonstrate the occurrence of population expansions following a bottleneck event for a multisegmented plant RNA virus, and suggest that the inferred differentiation between viral populations is because of demographically related events.

The level of intraspecific polymorphism among coding regions of TSWV S and M RNAs has been estimated. Comparisons made with reported estimates of nucleotide diversity for other plant RNA viruses, such as *Pepper mild mottle virus* (0.018), *Tobacco mild green mosaic virus* (0.057), *Yam mosaic virus* (0.117), *Wheat streak mosaic virus* (0.031) and *Rice yellow mottle virus* (0.194) (Garcia-Arenal *et al.* 2001) demonstrate a relatively high TSWV genetic variability ($\pi_{Silent}$ ranged from 0.079 to 0.113). The latter can be related to the wide host range of TSWV, which includes both insects and plants, and its distinct ability to be transmitted both by thrips, in a persistent manner, and by vegetative propagation of plants. Comparisons between the TSWV genetic diversity and that of other animal RNA viruses, such as HIV and *Influenza A* (Sharp 2002), reveal a relatively lower level of heterogeneity for TSWV. This can be attributed to the additional selection pressure imposed to animal RNA viruses by the immune system. Finally, when comparisons are made among different species, TSWV exhibits a relatively higher level of nucleotide polymorphism. For example, *Drosophila melanogaster* diversity ranges from 0.00 to 0.009 (Aquadro 1992), which is eight to 13-fold more diverse at the DNA sequence level than humans (Zwick *et al.* 2000). Diversity of maize ranges from 0.0028 to 0.036 (Tenaillon *et al.* 2001). The elevated levels of TSWV genetic diversity compared to other species may be attributed to the high mutation frequencies of RNA viruses.

At the species level there is evidence of high selective constraints, restricting recurrent evolution of the NSs, N, NSm and G1-G2 coding regions, although specific amino acids appear to be under positive selection. Strong selective constraints can be attributed to the key roles of the analysed genes in viral functions. The NSs structural protein, has RNA silencing suppressor activity and it also affects symptom expression in TSWV-infected plants (Takeda *et al*. 2002). The N protein, that encapsidates the viral RNA within the viral envelope (Richmond *et al*. 1998), is the predominant protein that antisera recognize, functions in the replication complex, and in *Bunyaviruses* has been shown to control the switch from transcription to translation (Kolakofsky & Hacker 1991). The G1 and G2 glycoproteins play an important role in maturation/assembly of virions in both plants and thrips and also in attachment to cell surface receptors (in thrips only) (Bandla *et al*. 1998).

The NSm protein is encoded only by *Tospoviruses*, representing presumably an evolutionary adaptation of the *Bunyavirus* genome to plant hosts (Goldbach & Peters 1994). This protein is involved in cell-to-cell movement (Kormelink *et al*. 1994) and stimulation of tubular structures both in protoplasts and insect cells (Storms *et al*. 1995). It is also involved in host range determination and symptomatology (Silva *et al*. 2001). Also, NSm specifically interacts with the TSWV N protein and binds ss RNA in a sequence-nonspecific manner (Soellick *et al*. 2000). NSm is the only protein where no evidence for positive selection was identified. If plant host species have any role in TSWV evolution, we would expect to see positively selected sites on the NSm gene. The lack of positive selected sites on this gene may be an indication that host associated selection is not a major factor affecting TSWV evolution. An additional explanation for the failure to detect positive selection in the NSm protein might be the lack of statistical power of the method used. Overall, this analysis confirms that at the species level, purifying selection is acting to maintain functional integrity.

Additionally, we showed evidence for positive selection on specific amino acids of the NSs, N and G1-G2 genes. Prediction of the secondary structure type of each amino acid residue of the NSs, N and G1-G2 proteins, indicates that all positively selected sites are part of helix or turn domains. The absence of structural data and a reverse genetics system for TSWV is a crucial limitation for the prediction of the impact of these amino acid replacements. These amino acid changes may affect the binding of the virus to cells and subsequently influence its pathogenicity. Although purifying selection is predominantly acting at the species level to preserve the function of the encoded proteins, positive selection is favouring divergence between *Tospovirus* species. For the four analysed genes, NSs, N, NSm and G1-G2, we found that there is more amino acid replacement than synonymous substitutions between species. Until now, species characterization within *Bunyaviridae* has been based on molecular characteristics and estimations of identity between sequences. Our current analysis confirms that TSWV and INSV are different species within the same family, and demonstrates that selection may be a key evolutionary factor leading to speciation between these viral groups.

It is important to note that in this paper the analysed TSWV isolates were passaged to a common host (*Nicotiana benthamiana*) prior to RNA extraction. This step was necessary in order to increase the viral concentration, and it is difficult to overcome it experimentally. Although, the mechanical passage may have added an additional selection pressure on the virus, our current phylogenetic analyses suggest that it did not have a severe effect on the population genetics of TSWV as the geographical distinctions remained intact. In addition, for the analyses, we used multiple population genetic tests, a number of which have been developed initially for DNA genomes. These tests have been used currently for the analysis of a number of human and animal RNA viruses and represent a reliable way to infer the population dynamics of RNA viral species. The only limitation in our analysis is that the CODEML algorithm, used for the identification of variable selective constraints exerted on the analysed genes, does not detect selection pressures on the RNA sequence level which also occur in RNA viruses. Selection at the RNA level would manifest as negative selection pressure in the CODEML analysis, further emphasizing the importance of the identified positively selected sites in the NSs, N and G1-G2 genes.

In conclusion, TSWV displays geographical structuring of isolated populations that may have arisen by founder effects. We identified positive selection that favours divergence between *Tospovirus* species. At the species level, purifying selection has acted to preserve protein function, although certain amino acids appear to be under positive selection. This analysis provides the first demonstration of population structuring and species-wide population expansions in a multisegmented plant RNA virus, using molecular population genetic analyses. This paper also identifies specific amino acid sites subject to selection within *Bunyaviridae* and estimates the level of genetic heterogeneity of a highly pathogenic plant RNA virus. The study of the variability of TSWV populations lays the foundation in the development of strategies for the control of other viral diseases in floral crops. Taking into consideration the limited amount of population studies in plant RNA virology we believe that this research will set the stage for similar studies for other plant RNA viruses.

## Acknowledgements

We would like to acknowledge M. J. Buck, S. Guyader, A. Lawton-Rauh, and K. Olsen for their indispensable help with some of the software applications used for the analysis of the sequence data. Additionally, we would like to thank W. R. Atchley, and K. Olsen for the edition of the manuscript. This work was supported in part by the Cooperative Agreement grant from the USDA-ARS and by the American Floral Endowment grant.

## References

Albiach-Marti MR, Guerri J, de Mendoza H *et al.* (2000) Aphid transmission alters the genomic and defective RNA populations of *Citrus tristeza* virus isolates. *Virology*, **90**, 134–138.

Alicai T, Fenby NS, Gibson RW *et al.* (1999) Occurrence of two serotypes of sweet potato chlorotic stunt virus in East Africa and their associated differences in coat protein and HSP70 homologue gene sequences. *Plant Pathology*, **48**, 718–726.

Aquadro CF (1992) Why is the genome variable? Insights from *Drosophila*. *Trends in Genetics*, **8**, 355–362.

Aranda MA, Fraile A, Garcia-Arenal F (1993) Genetic variability and evolution of the satellite RNA of cucumber mosaic virus during natural epidemics. *Journal of Virology*, **67**, 5896–5901.

Arboleda M, Azzam O (2000) Inter and intrasite genetic diversity of natural field populations of rice tungro bacilliform virus in the Philippines. *Archives of Virology*, **145**, 275–289.

de Avila AC, de Haan P, Kormelink R *et al.* (1993) Classification of *Tospoviruses* based on phylogeny of nucleoprotein gene sequences. *Journal of General Virology*, **74**, 153–159.

Ayllon MA, Rubio L, Moya A, Guerri J, Moreno P (1999) The haplotype distribution of two genes of *Citrus tristeza* virus is altered after host change or aphid transmission. *Virology*, **255**, 32–39.

Azzam O, Arboleda M, Umadhay KM *et al.* (2000a) Genetic composition and complexity of virus populations at tungro-endemic and outbreak rice sites. *Archives of Virology*, **145**, 2643–2657.

Azzam O, Yambao ML, Muhsin M, McNally KL, Umadhay KM (2000b) Genetic diversity of rice tungro spherical virus in tungro-endemic provinces of the Philippines and Indonesia. *Archives of Virology*, **145**, 1183–1197.

Bandla MD, Campbell LR, Ullman DE, Sherwood JL (1998) Interaction of tomato spotted wilt *Tospovirus* (TSWV) glycoproteins with a thrips midgut protein, a potential cellular receptor for TSWV. *Phytopathology*, **88**, 98–104.

Bergstrom CT, McElhany P, Real LA (1999) Transmission bottlenecks as determinants of virulence in rapidly evolving pathogens. *Proceedings of the National Academy of Sciences USA*, **96**, 5095–5100.

Bohlman MC, Morzunov SP, Meissner J *et al.* (2002) Analysis of hantavirus genetic diversity in Argentina: s segment-derived phylogeny. *Journal of Virology*, **76**, 3765–3773.

Choi IR, Hall JS, Henry M *et al.* (2001) Contributions of genetic drift and negative selection on the evolution of three strains of wheat streak mosaic tritimovirus. *Archives of Virology*, **146**, 619–628.

Dewey RA, Semorile LC, Grau O, Crisci JV (1997) Cladistic analysis of *Tospovirus* using molecular characters. *Molecular Phylogenetics and Evolution*, **8**, 11–32.

Domingo E, Holland JJ (1997) RNA virus mutations and fitness for survival. *Annual Review of Microbiology*, **51**, 151–178.

Elliott RM (1996) *The Bunyaviridae*. Plenum Press, New York and London.

Fargette D, Pinel A, Abubakar Z *et al.* (2004) Inferring the evolutionary history of rice yellow mottle virus from genomic, phylogenetic, and phylogeographic studies. *Journal of Virology*, **78**, 3252–3261.

Feuer R, Boone JD, Netski D, Morzunov SP, St Jeor SC (1999) Temporal and spatial analysis of Sin Nombre virus quasispecies in naturally-infected rodents. *Journal of Virology*, **73**, 9544–9554.

Fraile A, Alonso-Prados JL, Aranda MA *et al.* (1997) Genetic exchange by recombination or reassortment is infrequent in natural populations of a tripartite RNA plant virus. *Journal of Virology*, **71**, 934–940.

Fraile A, Malpica JM, Aranda MA, Rodriguez-Cerezo E, Garcia-Arenal F (1996) Genetic diversity in tobacco mild green mosaic *Tobamovirus* infecting the wild plant *Nicotiana glauca*. *Virology*, **223**, 148–155.

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. *Genetics*, **133**, 693–709.

Garcia-Arenal F, Fraile A, Malpica JM (2001) Variability and genetic stucture of plant virus populations. *Annual Review of Phytopathology*, **39**, 157–186.

Goldbach R, Peters D (1994) Possible causes of the emergence of *Tospovirus* diseases. *Seminars in Virology*, **5**, 113–120.

Grant WS, Bowen BW (1998) Shallow population histories in deep evolutionary lineages of marine fishes: insights from sardines and anchovies and lessons for conservation. *Journal of Heredity*, **89**, 415–426.

Guyader S, Ducray DG (2002) Sequence analysis of potato leafroll virus isolates reveals genetic stability, major evolutionary events and differential selection pressure between overlapping reading frame products. *Journal of General Virology*, **83**, 1799–1807.

de Haan P, de Avila AC, Kormelink R *et al.* (1992) The nucleotide sequence of the S RNA of *Impatiens* necrotic spot virus, a novel *Tospovirus*. *Federation of European Biochemical Societies (FEBS) Letters*, **306**, 27–32.

Heinze C, Letschert B, Hristova D *et al.* (2001) Variability of the N-protein and the intergenic region of the S RNA of tomato spotted wilt *Tospovirus* (TSWV). *New Microbiology*, **24**, 175–187.

Hey J, Harris E (1999) Population bottlenecks and patterns of human polymorphism. *Molecular Biology and Evolution*, **16**, 1423–1426.

Hudson RR (2000) A new statistic for detecting genetic differentiation. *Genetics*, **155**, 2011–2014.

Hudson RR (2001) Two-locus sampling distributions and their application. *Genetics*, **159**, 1805–1817.

Hudson RR, Boos DD, Kaplan NL (1992) A statistical test for detecting geographic subdivision. *Molecular Biology and Evolution*, **9**, 138–151.

Kneller DG, Cohen FE, Langridge R (1990) Improvements in protein secondary structure prediction by an enhanced neural network. *Journal of Molecular Biology*, **214**, 171–182.

Kofalvi SA, Marcos JF, Canizares MC, Pallas V, Candresse T (1997) Hop stunt viroid (HSVd) sequence variants from *Prunus* species: evidence for recombination between HSVd isolates. *Journal of General Virology*, **78**, 3177–3186.

Kolakofsky D, Hacker D (1991) Bunyavirus RNA synthesis: genome transcription and replication. *Current Topics in Microbiology and Immunology*, **169**, 143–159.

Kormelink R, Storms M, Van Lent J, Peters D, Goldbach R (1994) Expression and subcellular location of the NSM protein of tomato spotted wilt virus (TSWV), a putative viral movement protein. *Virology*, **200**, 56–65.

Kumar S, Tamura K, Nei M (1993) MEGA: molecular evolutionary genetic analysis. *Pennsylvania*. State University Pres. University Park.

Kurath G, Dodds JA (1995) Mutation analyses of molecularly cloned satellite tobacco mosaic virus during serial passage in plants: evidence for hotspots of genetic change. *RNA*, **1**, 491–500.

Law MD, Moyer JW (1990) A tomato spotted wilt-like virus with a serologically distinct N protein. *Journal of General Virology*, **71**, 933–938.

Li H, Roossinck MJ (2004) Genetic bottlenecks reduce population variation in an experimental RNA virus population. *Journal of Virology*, **78**, 10582–10587.

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature*, **351**, 652–654.

McVean G, Awadalla P, Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, **160**, 1231–1241.

Moury B, Morel C, Johansen E, Jacquemond M (2002) Evidence for diversifying selection in potato virus Y and in the coat protein of other *Potyviruses*. *Journal of General Virology*, **83**, 2563–2573.

Moya A, Rodriguez-Cerezo E, Garcia-Arenal F (1993) Genetic structure of natural populations of the plant RNA virus Tobacco mild green mosaic virus. *Molecular Biology and Evolution*, **10**, 449–446.

Moyer JW (1999) *Tospoviruses* (Bunyaviridae). In: *Encyclopedia of Virology* (ed. Webster R, Granoff A), pp. 1803–1807. Academic Press, London.

Moyer JW (2000) Tospoviruses. In: *Encyclopedia of Microbiology* (ed. Hull R), pp. 592–597. Academic Press, London.

Nemirov K, Henttonen H, Vaheri A, Plyusnin A (2002) Phylogenetic evidence for host switching in the evolution of hantaviruses carried by *Apodemus* mice. *Virus Research*, **90**, 207–215.

Nicholas KB, Nicholas HBJ (1997) GENEDOC: a tool for editing and annotating multiple sequence alignments. Distributed by the author.

Novella IS, Elena SF, Moya A, Domingo E, Holland JJ (1995) Size of genetic bottlenecks leading to virus fitness loss is determined by mean initial population fitness. *Journal of Virology*, **69**, 2869–2872.

Pappu H, Pappu S, Jain R *et al.* (1998) Sequence characteristics of natural populations of tomato spotted wilt *Tospovirus* infecting flue-cured tobacco in Georgia. *Virus Genes*, **17**, 169–177.

Plyusnin A (2002) Genetics of hantaviruses: implications to taxonomy. *Archives of Virology*, **147**, 665–682.

Qiu WP, Geske SM, Hickey CM, Moyer JW (1998) Tomato spotted wilt *Tospovirus* genome reassortment and genome segment-specific adaptation. *Virology*, **244**, 186–194.

Qiu W, Moyer JW (1999) Tomato spotted wilt *Tospovirus* adapts to the TSWV N gene-derived resistance by genome reassortment. *Phytopathology*, **89**, 575–582.

Richmond KE, Chenault K, Sherwood JL, German TL (1998) Characterization of the nucleic acid binding properties of tomato spotted wilt virus nucleocapsid protein. *Virology*, **248**, 6–11.

Rozas J, Rozas R (1999) *DNASP: an Integrated Program for Molecular Population Genetics and Molecular Evolution Analysis*, Version 3. Bioinformatics, 15, 174–175.

Rozas J, Sanchez-DeI, Barrio JC, Messeguer X, Rozas R (2003) DNASP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, **19**, 2496–2497.

Rubio L, Abou-Jawdah Y, Lin HX, Falk BW (2001) Geographically distant isolates of the crinivirus Cucurbit yellow stunting disorder virus show very low genetic diversity in the coat protein gene. *Journal of General Virology*, **82**, 929–933.

Schneider WL, Roossinck MJ (2001) Genetic diversity in RNA virus quasispecies is controlled by host–virus interactions. *Journal of Virology*, **75**, 6566–6571.

Sharp PM (2002) Origins of human virus diversity. *Cell*, **108**, 305–312.

Silva MS, Martins CRF, Bezerra IC *et al.* (2001) Sequence diversity of NSm movement protein of *Tospoviruses*. *Archives of Virology*, **146**, 1267–1281.

Sironen T, Plyusnina A, Andersen HK *et al.* (2002) Distribution of *Puumala hantavirus* in Denmark: analysis of bank voles (*Clethrionomys glareolus*) from Fyn and Jutland. *Vector-Borne and Zoonotic Diseases*, **2**, 37–45.

Soellick T, Uhrig JF, Bucher GL, Kellmann JW, Schreier PH (2000) The movement protein NSm of tomato spotted wilt *Tospovirus* (TSWV): RNA binding, interaction with the TSWV N protein, and identification of interacting plant proteins. *Proceedings of the National Academy of Sciences of the USA*, **97**, 2373–2378.

Storms MM, Kormelink R, Peters D, Van Lent JW, Goldbach RW (1995) The nonstructural NSm protein of tomato spotted wilt virus induces tubular structures in plant and insect cells. *Virology*, **214**, 485–493.

Swofford DL (2000) PAUP*. Phylogenetic Analysis Using Parsimony (*and other Methods), Version 4.*0b8a*. Sinauer, Sunderland, MA.

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**, 437–460.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Takeda A, Sugiyama K, Nagano H *et al.* (2002) Identification of a novel RNA silencing suppressor, NSs protein of tomato spotted wilt virus. *Federation of European Biochemical Societies (FEBS) Letters*, **532**, 75–79.

Tenaillon MI, Sawkins MC, Long AD *et al.* (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proceedings of the National Academy of Sciences of the USA*, **98**, 9161–9166.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research*, **25**, 4876–4882.

Ullman DE, Meideros R, Campbell LR, Whitfield AE, Sherwood JL (2002) Thrips as vectors of *Tospoviruses*. In: *Advances in Botanical Research* (ed. Plumb R), pp. 113–140. Elsevier Science Ltd.

Vives MC, Rubio L, Galipienso L *et al.* (2002) Low genetic variation between isolates of *Citrus* leaf blotch virus from different host species and of different geographical origins. *Journal of General Virology*, **83**, 2587–2591.

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256–276.

Wright S (1951) The genetical structure of populations. *Annals of Eugenics*, **15**, 323–354.

Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, **155**, 431–449.

Yashina L, Petrova I, Seregin S *et al.* (2003) Genetic variability of Crimean-Congo hemorrhagic fever virus in Russia and Central Asia. *Journal of General Virology*, **84**, 1199–1206.

Zwick ME, Cutler DJ, Chakravarti A (2000) Patterns of genetic variation in Mendelian and complex traits. *Annual Review of Genomics and Human Genetics*, **1**, 387–407.

## Appendix I

Tomato spotted wilt isolates used in this paper

| Isolate | Region analysed | Host | Collected from | Acces. No. | Date |
|---|---|---|---|---|---|
| CA-1 | NSs, N | Aster | California[a] | AY744468 | 12/08/2000[d] |
| CA-2 | NSs, N | Buttercup | California[a] | AY744469 | 11/21/2000[d] |
| CA-3 | NSs, N, NSm, G1-G2 | Chrysanthemum | California[a] | AY744470, AY744481 | 11/21/2000[d] |
| CA-4 | NSs, N, NSm, G1-G2 | Chrysanthemum | California[a] | AY744471, AY744482 | 01/08/2002[d] |
| CA-5 | NSs, N, NSm, G1-G2 | Chrysanthemum | California[a] | AY744472, AY744483 | 01/08/2002[d] |
| CA-6 | NSs, N, NSm G1-G2 | Chrysanthemum | California[a] | AY744473, AY744484 | 01/08/2002[d] |
| CA-7 | NSs, N, NSm, G1-G2 | Dahlia | California[a] | AY744474, AY744485 | 03/22/2002[d] |
| CO | NSs, N | Falso lulo | Colorado[a] | AY744475 | 01/31/2001[d] |
| NC-1 | NSs, N | Dahlia | N. Carolina[a] | AY744476 | 06/20/2001[d] |
| NC-2 | NSs, N | Peanut | N. Carolina[a] | AY744477 | 07/18/2001[d] |
| NC-3 | NSs, N, NSm, G1-G2 | Dahlia | N. Carolina[a] | AY744478, AY744486 | 03/27/2002[d] |
| NC-4 | NSm, G1-G2 | Tobacco | N. Carolina[a] | AY744487 | 05/29/2002[d] |
| NC-5 | NSm, G1-G2 | Tobacco | N. Carolina[a] | AY744488 | 05/31/2002[d] |
| NC-6 | NSm, G1-G2 | Pepper | N. Carolina[a] | AY744489 | 06/05/2002[d] |
| NC-7 | NSm, G1-G2 | Tobacco | N. Carolina[a] | AY744490 | 05/27/2002[d] |
| NC-8 | NSm, G1-G2 | Tomato | N. Carolina[a] | AY744491 | 05/28/2002[d] |
| NC-9 | G1-G2 | Pepper | N. Carolina[a] | AY744494 | 06/07/2002[d] |
| NC-10 | G1-G2 | Pepper | N. Carolina[a] | AY744495 | 05/30/2002[d] |
| NC-11 | NSs, N | Peanut | N. Carolina[b] | AF020659 | 08/23/1997[e] |
| SPAIN-1 | NSs, N, NSm, G1-G2 | Tomato | Spain[a] | AY744479, AY744492 | 6/12/2001[d] |
| SPAIN-2 | NSs, N, NSm, G1-G2 | Tomato | Spain[a] | AY744480, AY744493 | 6/12/2001[d] |
| SPAIN-3 | N | N/A | Spain[c] | X94550 | 12/28/1995[e] |
| BULG-1 | NSs, N | Tobacco | Bulgaria[c] | AJ418777 (AJ297610) | 08/18/2000[e] |
| BULG-2 | NSs and N | Tobacco | Bulgaria[c] | AJ418778 | 11/05/2001[e] |
| BULG-3 | NSs, N | Tobacco | Bulgaria[c] | AJ418779 (AJ297608) | 08/18/2000[e] |
| BULG-4 | NSs, N | Tobacco | Bulgaria[c] | AJ418780 (AJ297609) | 08/18/2000[e] |
| BULG-5 | NSs, N, NSm | Tobacco | Bulgaria[c] | D13926 (X93603) | 01/18/1991[e] |
| BULG-6 | N | Tomato | Bulgaria[c] | AJ296598 | 07/29/2000[e] |
| BULG-7 | N | Hippeastrum | Bulgaria[c] | AJ296601 | 07/29/2000[e] |
| BULG-8 | N | Dahlia | Bulgaria[c] | AJ296602 | 07/29/2000[e] |
| GER | NSs, N lysimachia | Germany[c] |  | AJ418781 (AJ297611) | 08/18/2000[e] |
| NETH | NSs, N, NSm, G1-G2 | Dahlia | Netherlands[b] | AF020660, AF208497 | 08/23/1997[e] |
| BR | N | Tomato | Brazil[c] | D00645 | 1990[e] |
| SA | N | Potato | South Africa[c] | AJ296600 | 07/29/2000[e] |
| IT | N | Tomato | Italy[c] | Z36882 | 08/22/1994[e] |
| HAW-1 | N | N/A | Hawaii[c] | X61799 | 09/01/1991[e] |
| HAW-2 | NSm, G1-G2 | Tomato | Hawaii[b] | AF208498 | 11/26/1999[e] |
| JAP-1 | N | Chrysanthemum | Japan[c] | AB038342 | 02/13/2000[e] |
| JAP-2 | N | Chrysanthemum | Japan[c] | AB038341 | 02/13/2000[e] |
| JAP-3 | N | N/A | Japan[c] | AB010997 | 02/09/1998[e] |
| JAP-4 | NSm, G1-G2 | N/A | Japan[c] | AB010996 | 02/09/1998[e] |
| GA-1 | N | Tobacco | Georgia[c] | AF064474 | 05/07/1998[e] |
| GA-2 | N | Tobacco | Georgia[c] | AF064473 | 1996[f] |
| GA-3 | N | Tobacco | Georgia[c] | AF064472 | 05/07/1998[e] |
| GA-4 | N | Tobacco | Georgia[c] | AF064471 | 05/07/1998[e] |
| GA-5 | N | Tobacco | Georgia[c] | AF064470 | 05/07/1998[e] |
| GA-6 | N | Tobacco | Georgia[c] | AF064469 | 05/07/1998[e] |
| GA-7 | N | Pepper | Georgia[c] | AF048716 | 02/17/1998[e] |
| GA-8 | N | Peanut | Georgia[c] | AF048715 | 02/17/1998[e] |
| GA-9 | N | Tomato | Georgia[c] | AF048714 | 02/17/1998[e] |

[a] TSWV isolate sequenced for this study.

[b] TSWV isolate previously sequenced by our laboratory.

[c] TSWV sequence from GenBank.

[d] Date TSWV sample was collected.

[e] Date TSWV sequence was submitted to GenBank (collection date was not avilable).

[f] Date paper containing the sequence was published (collection date was not available).

## Appendix II

Synthetic oligonucleotides used for RT- PCR and corresponding annealing temperature

| Primer pairs | Sequence 5′→3′ | Temp. °C |
|---|---|---|
| S RNA | | |
| Pair: S1 | CCTCTAGAAGAGCAATTGTGTCA[a] | 50 |
| S154 | AGATGCAGTTGATCCCCAGACTGAA | |
| Pair: S1 | CCTCTAGAAGAGCAATTGTGTCA | 50 |
| S691 | TGGCTTGAAACTGTACAGCCATTCA | |
| Pair: S574 | GTCTTGTGTCAAAGAGCATACCTATAA | 50 |
| S1433 | TGATCCCGCTTAAATCAAGCT | |
| Pair: S1275[b] | CACTTGAATGTCTTCC | 45.5 |
| S2067[b] | GGAAGTATTGCTATGG | |
| Pair: S1983[b] | CCCTCGAGGCTTTCAAGCAAGTTCTGCG | 55 |
| S2767[b] | GCTCTAGAGCCATCATGTCTAAGGTTAAGCTCAC | |
| Pair: S2500 | GGCTCCAATCCTGTCTGAA | 44 |
| S2916 | GGGGTACCAGAGCAAT | |
| Pair: S2739 | CCTTAGTGAGCTTAACCTTAGAC | 44 |
| S2916 | GGGGTACCAGAGCAAT | |
| MRNA | | |
| Pair: M1M | CCTCTAGAAGAGCAATCAGTGC | 55 |
| M108 | GTCAACATTTTGAGTTCAACAGCC | |
| Pair: M1 | AGAGCAATCAGTGCATCAGAAATATACCTATTATACA | 55 |
| M743 | CACTACCAAAAGAAACCCC | |
| Pair: M490 | TTCCAGGATTGTGATATG | 40 |
| M821 | TTAGAGGTATAACCATAC | |
| Pair: M658 | GAAGATGAACAACACCCC | 40 |
| M1393[c] | TATGTTAATGAAAGATACAA | |
| Pair: M944[c] | TCAGTTGAAGAGGAAGA | 41 |
| M1393[c] | TATGTTAATGAAAGATACAA | |
| Pair: M1315 | CTGTGACAAGCATCTTC | 45.5 |
| M2109 | GGTTTAGAGCAAATATCAG | |
| Pair: M1922 | CCGCATAGAAGACAGCC | 55 |
| M2665 | TACAGGAAACTGCGACAC | |
| Pair: M2565 | ACCAAGCTTCTTCACATCC | 46 |
| M3320 | TTTATGTTCCAGGCTGTCC | |
| Pair: M3206 | GTGCCAAAGATACTCTCTATG | 46 |
| M3920 | CTGAGGAAATGTTGGATGG | |
| Pair: M3832 | GCAATCTCTGACTCTTT | 46 |
| M4595 | ATGATGATTCTGCTGAG | |
| Pair: M4479 | GTATCTGACGGGTTCCAGG | 55 |
| M4821 | AGAGCAATCAGTGCAAACAAAAACCTTAATCC | |
| Pair: M4671 | CAGAACTCAGGGCAATTGTG | 48 |
| M4821M | GGGGTACCAGAGCAAT | |

[a] Underlined regions of primers correspond to added restriction sites to the primer sequence.
[b] Primers designed by Qiu *et al*. 1998.
[c] Primers designed by Bhat *et al*. 1999.