

Metagenomic analysis of nepoviruses: diversity, evolution and identification of a genome region in members of subgroup A that appears to be important for host range

Hily JM^{1,2}, Poulicard N³, Kubina J², Reynard JS⁴, Spilmont AS¹, Fuchs M⁵, Lemaire O² and Vigne E^{2*}.

1 IFV, Le Grau-Du-Roi, France

2 Université de Strasbourg, INRAE, SVQV UMR-A 1131, F-68000 Colmar, France

3 PHIM, Université Montpellier, IRD, INRAE, Cirad, SupAgro, Montpellier, France

4 Institute for Plant Production Science, Agroscope, 1260 Nyon, Switzerland

5 Cornell University, Geneva, NY, USA

* Corresponding author: emmanuelle.vigne@inrae.fr

Archives of Virology

Figure S1. Unrooted maximum likelihood trees of the complete set of sequences of the genus *Nepovirus* with 110 ORF1 and 167 ORF2 nucleotide sequences. Sequences from the genera *Fabavirus* (ORF1:BBW1_NC_005289 and BBW2_KJ789136; ORF2: BBW1_NC_005290 and BBW2_KJ789137) and *Comovirus* (ORF1: CPMV_NC_003549, ORF2: CPMV_NC_003550) were used as outgroups. Colors correspond to nepovirus species affiliation with red for subgroup A sequences, blue for subgroup B sequences and green for subgroup C sequences. The scale bars show genetic distance.

Figure S2. Schematic representation of intra- (triangles) and inter-species (circles) recombination events along ORF1 (left panel) and ORF2 (right panel) nucleotide sequences of representative sequences of each of the three nepovirus subgroups. Maps of the two ORFs are not drawn to scale. Numbers within triangles and circles correspond to recombination events predicted at hot spot locations. Details and precise locations of recombination events are provided in Tables S5 and S6.

Figure S3. Maximum likelihood trees using all available (as of January 2020) nucleotide sequences corresponding to the coding regions 2A, MP and CP of arabis mosaic virus (ArMV) isolates. The list of ArMV sequences is provided in Table S8. Numbers at each node indicate bootstrap values based on 100 replicates. Scale bar show genetic distance. ArMV isolates infecting monocotyledonous plants were shown in red and those infecting dicotyledonous plants are in green.

Figure S4. Genetic diversity analyses of grapevine fanleaf virus (GFLV) isolates from different countries using a corpus of 40 ORF1 (left panel) and 80 ORF2 (right panel) nucleotide sequences. Colors correspond to different regions of the world with New World isolates in yellow, Old World isolates in green, French isolates in blue and isolates from the rest of the world in red. Graphics represent π (substitution/sites) and Tajima's D (D_T) for evolution along the ORF1 and ORF2 sequences. Colored bars with # and * correspond to genetic regions being statistically valid (P-values at 0.05 and 0.001), respectively.

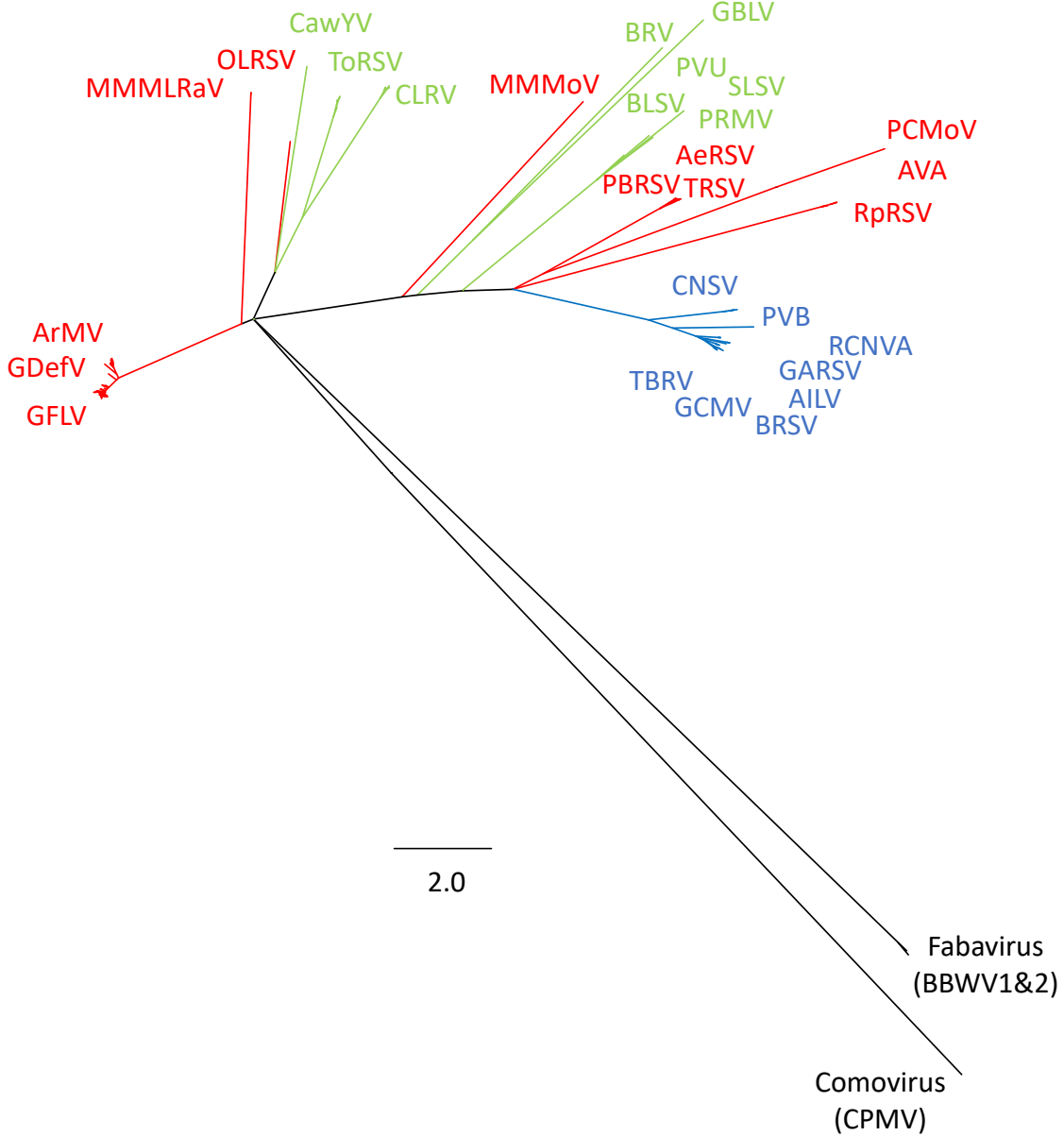
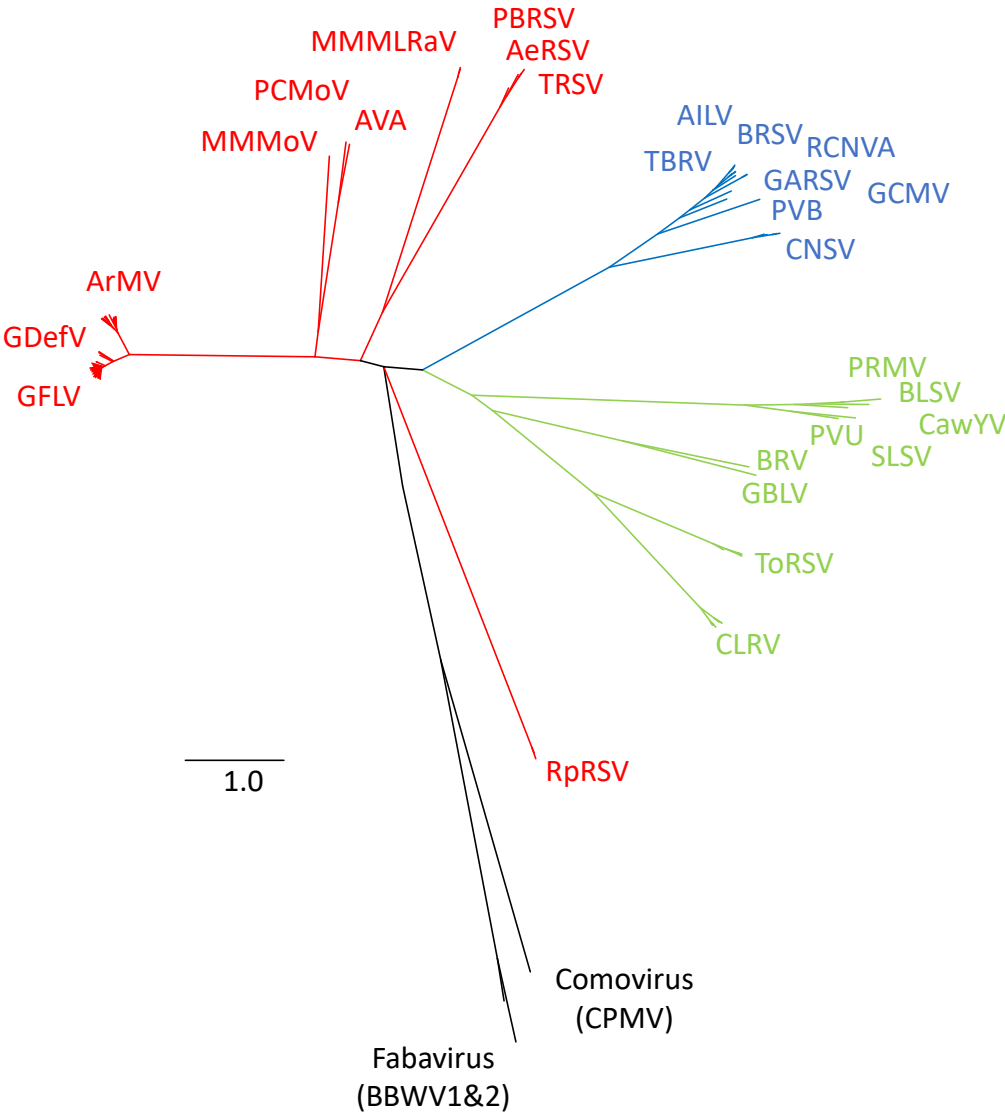
Figure S5. Phylogenetic and diversity analyses of the CP coding region using a corpus of 102 nucleotide sequences of grapevine fanleaf virus (GFLV), arabis mosaic virus (ArMV) and grapevine deformation virus (GDefV) isolates with π (substitution/sites) shown along the sequence (a), phylogenetic relationships using maximum likelihood tree (b), and consensus amino acids obtained by aligning regions previously predicted and/or validate in nematode-mediated transmission (R1-R5) (c). GFLV isolates are in blue, all ArMV isolates in yellow grapevine-infecting ArMV isolates (*Vitis*-ArMV) in yellow, and non-grapevine-infecting ArMV isolates (Non-*Vitis*-ArMV) in red. For panel a, the tree CP domains (A, B, C) are delineated. For the phylogenetic tree, numbers at each node indicate bootstrap values based on 100 replicates and the scale bar shows genetic distance.

Figure S6. Comparative analysis of the six possible reading frames of ORF1 from 40 grapevine fanleaf virus (GFLV) and 17 arabis mosaic virus (ArMV) sequences. Graphic representation of π (substitution/sites) along the sequence (top) and of the +1, +2, +3, -1, -2, -3 reading frames with stop codons in blue and gaps in gray (bottom). Within each reading frame that consists of 57 lines or sequences, the first 40 lines correspond to GFLV sequences, followed by nine *Vitis*-infecting ArMV sequences and finally by eight non-*Vitis*-infecting ArMV sequences. GFLV sequences are in blue, *Vitis*-infecting ArMV sequences in green and non-*Vitis*-infecting ArMV in red.

Figure S7. Phylogenetic tree reconstructed from the alignment of the distal 148 and 149 amino acids of the RNA-dependent RNA polymerase (Pol) from grapevine fanleaf virus (GFLV) and arabis mosaic virus (ArMV) isolates. The scale bar below the tree shows genetic distance. GFLV sequences are delineated in a blue area, *Vitis*-infecting ArMV sequences in a green area and non-*Vitis*-infecting ArMV in a red area.

Figure S8. Alignment of consensus amino acid sequences of the C-terminal part of the polymerase (Pol) from grapevine fanleaf virus (GFLV) isolates, *Vitis*-infected isolates of arabis mosaic virus (ArMV) and non-*Vitis*-infecting ArMV isolates. Arrows indicate identical amino acids between grapevine viruses but different from ArMV isolates infecting plants other than grapevines. The blue rectangle shows the distal 148/149 aa in the Pol domain.

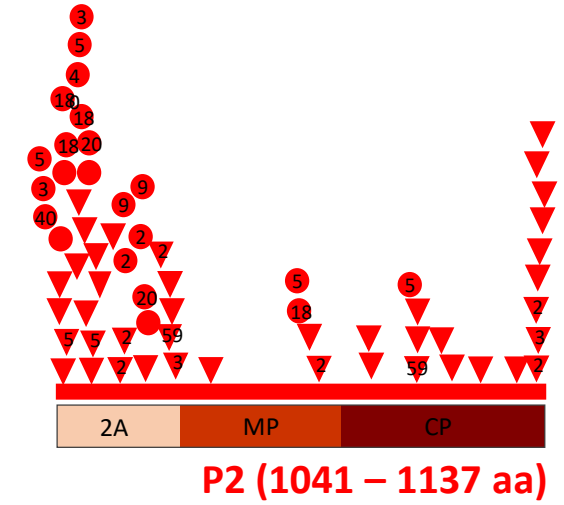
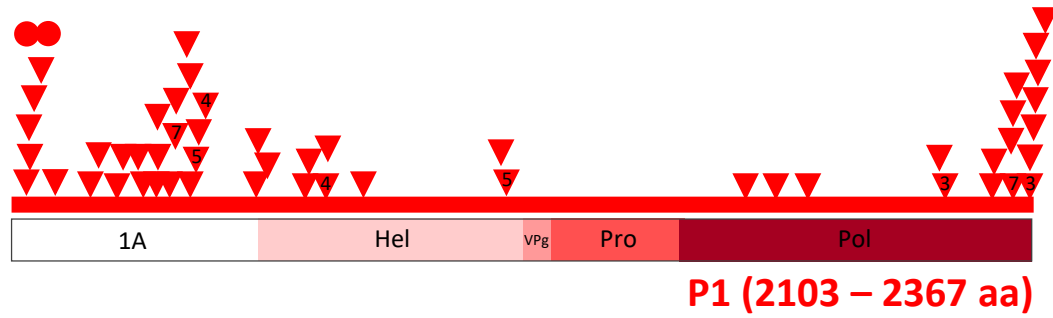
Figure S1



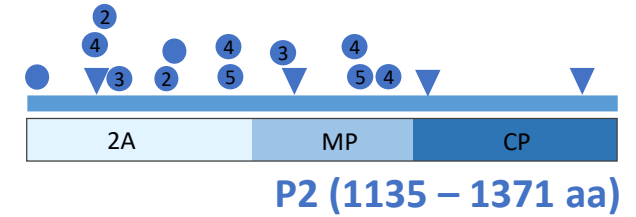
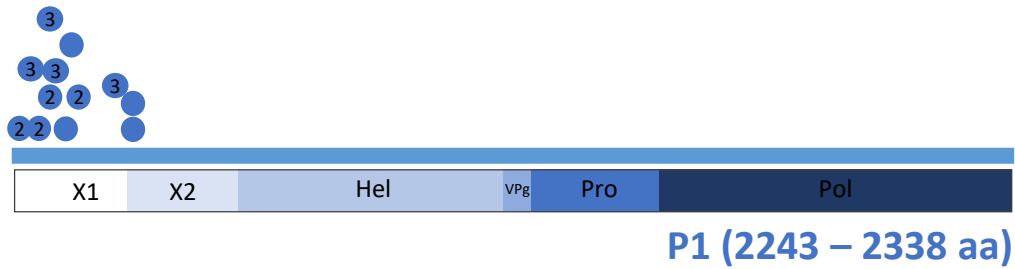
SubGroup A
SubGroup B
SubGroup C

Figure S2

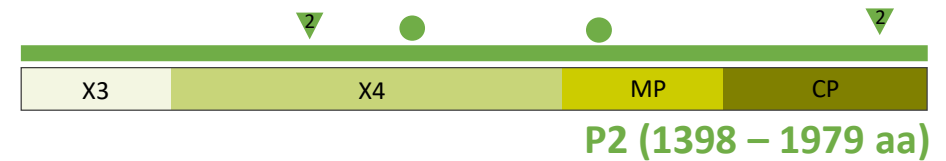
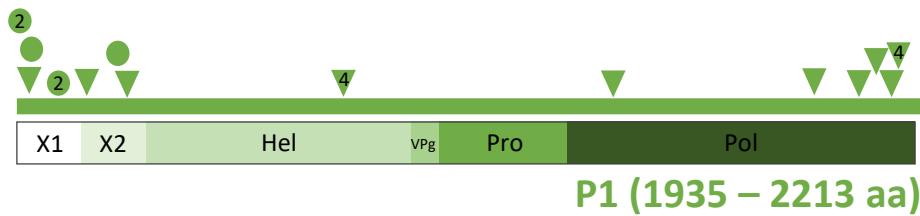
Subgroup A



Subgroup B



Subgroup C



▼ Intra-species ● Inter-species

Not perfectly scaled

Figure S3

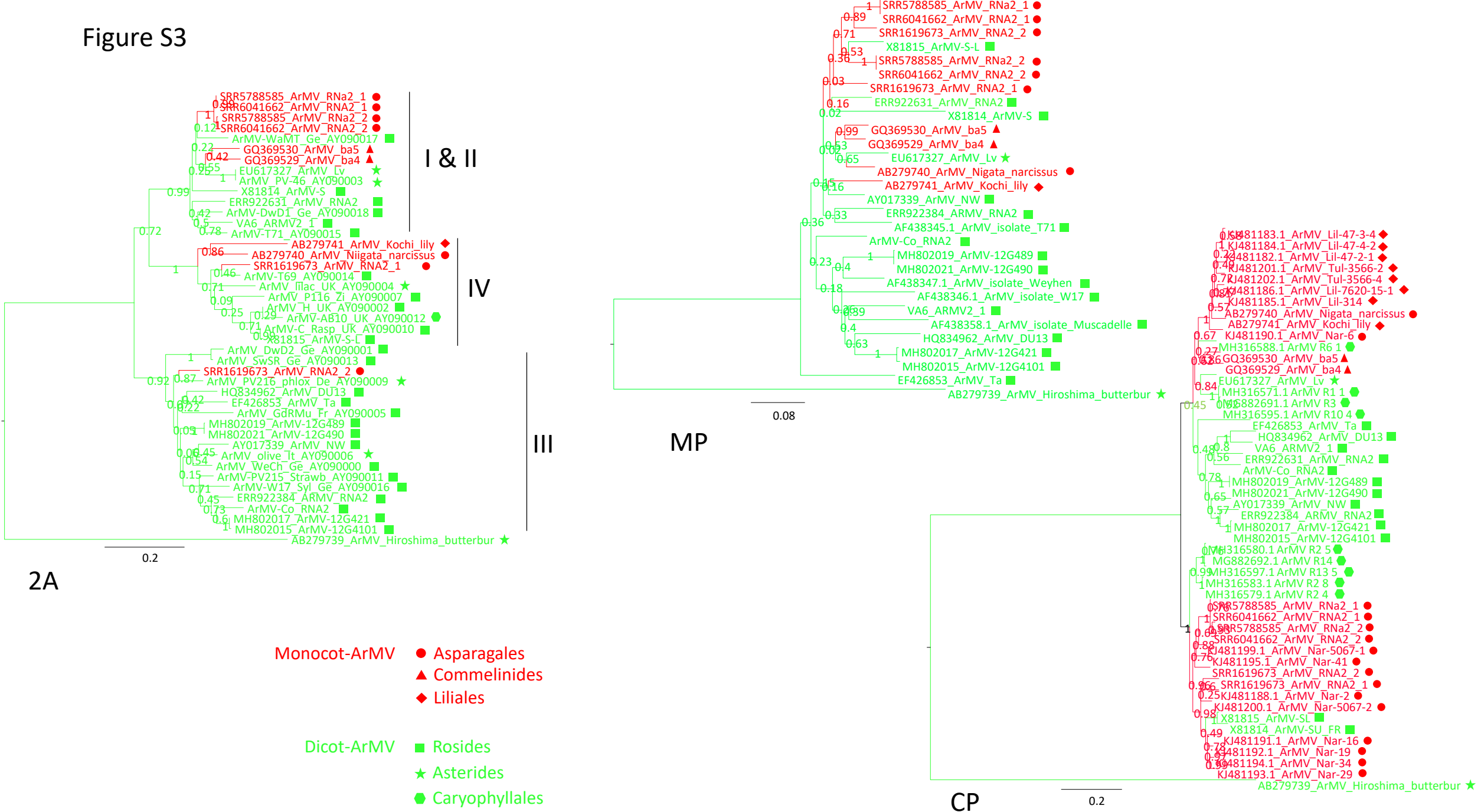
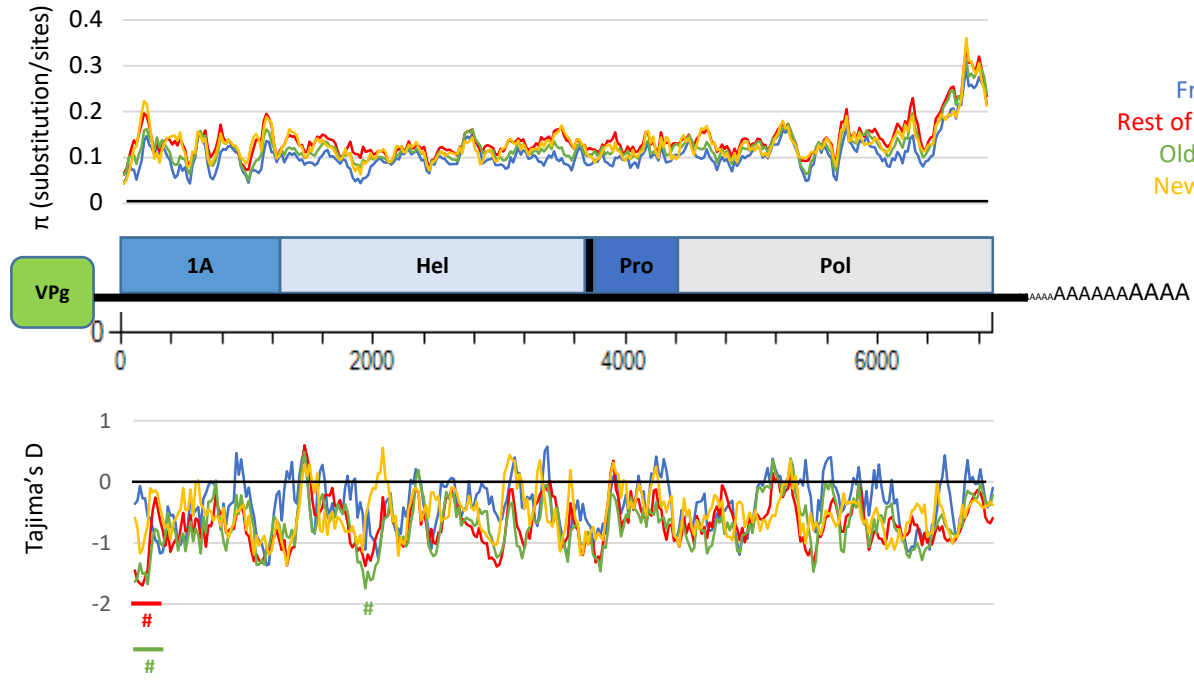


Figure S4

ORF1



ORF2

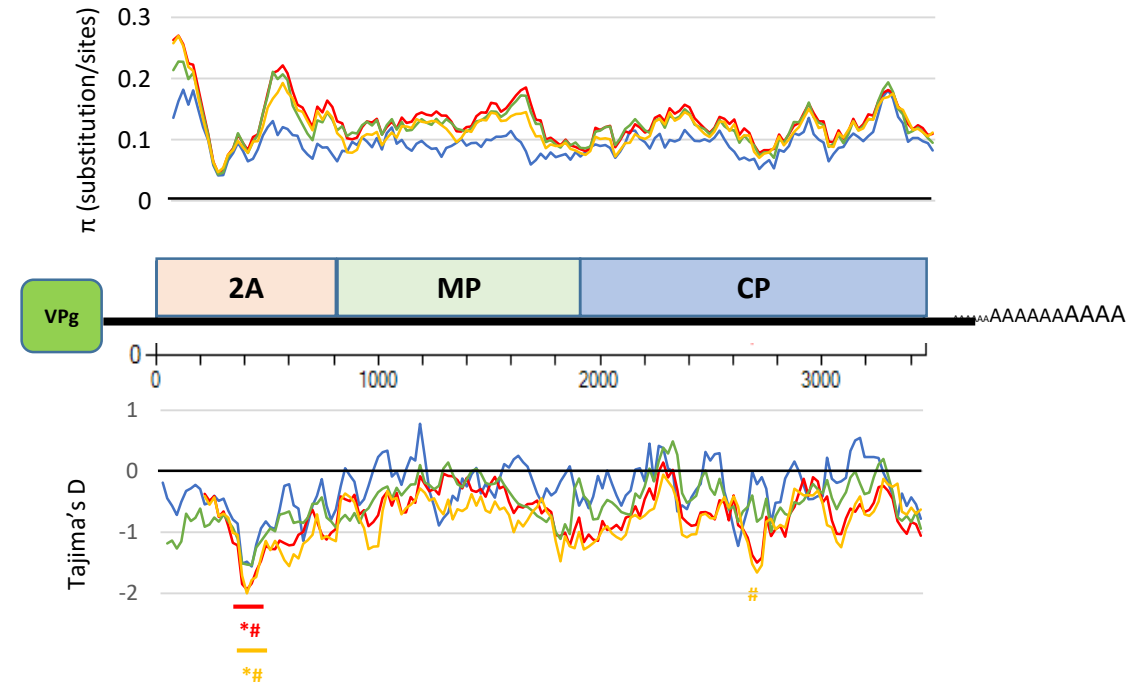


Figure S6

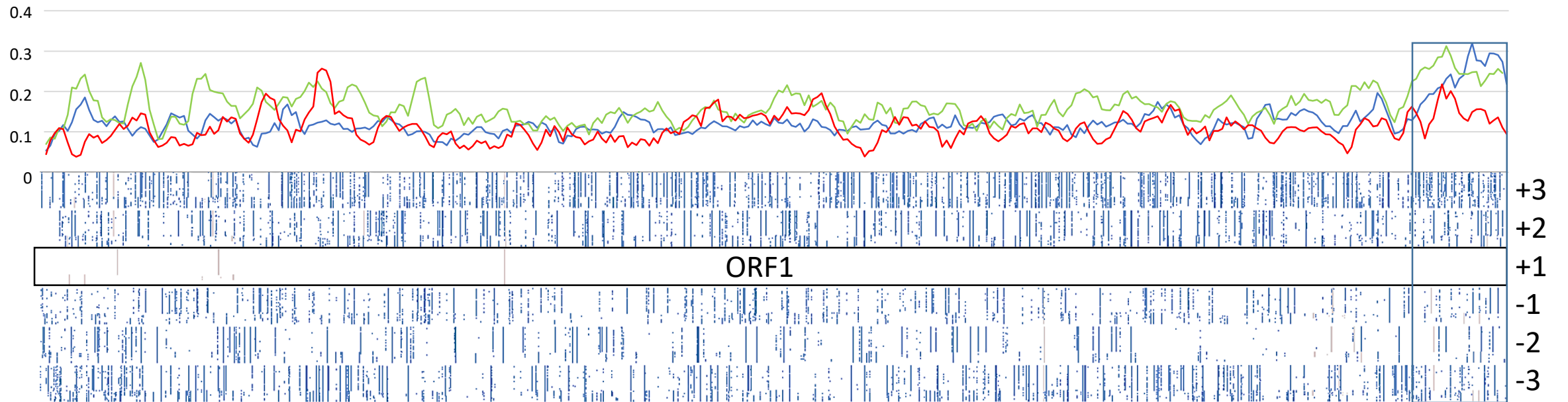
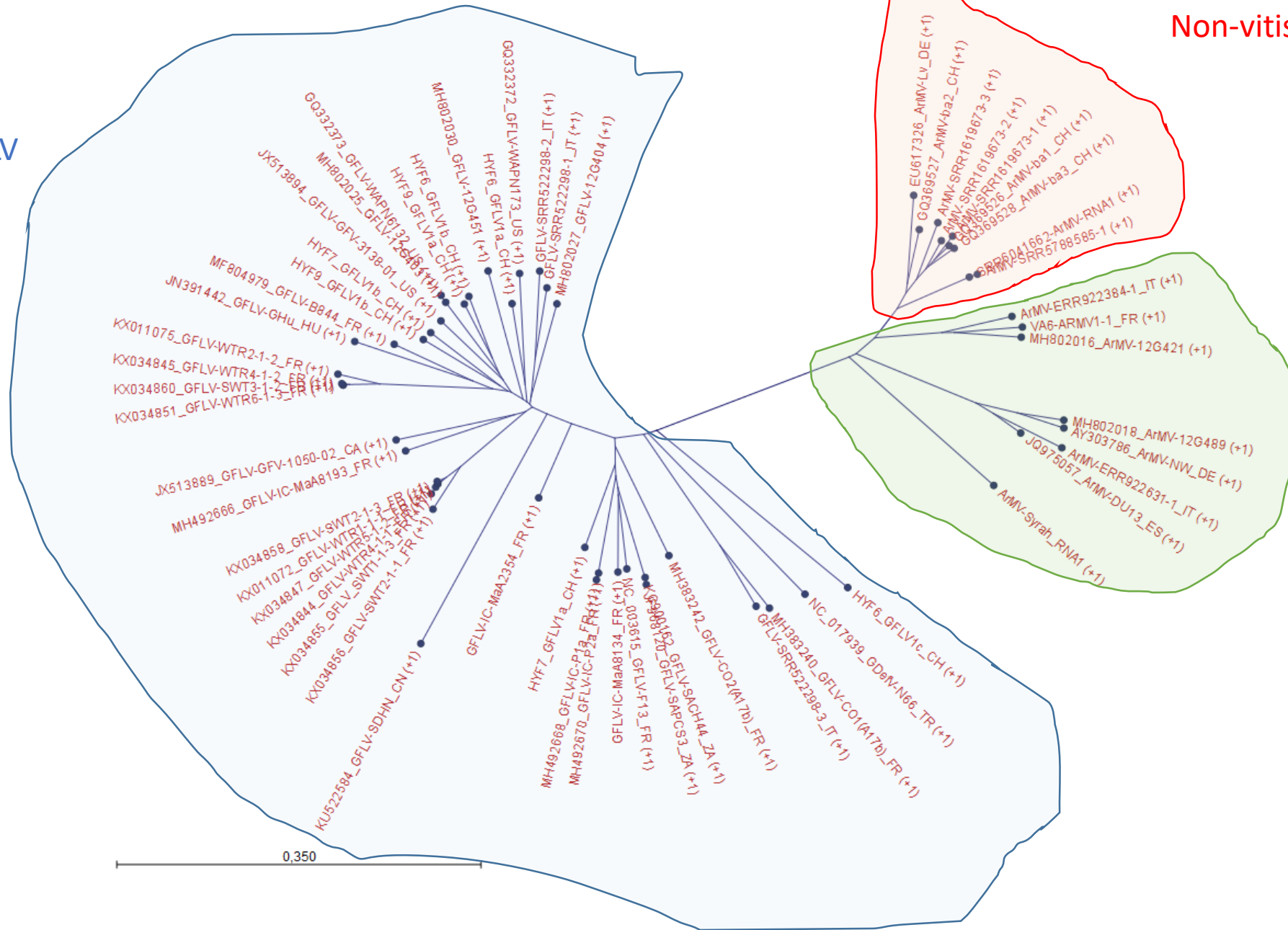


Figure S7

GFLV

Non-vitis-ArMV

Vitis-ArMV



Last 148/149 aa of the Pol coding region

Figure S8

