Supplementary material for

"Viroid quasispecies revealed by deep sequencing"

Rajen J. J. Brass [1], Robert A. Owens [2], Jaroslav Matoušek [3], and Gerhard Steger [1]*

[1]Institut für Physikalische Biologie, Heinrich-Heine-Universität Düsseldorf, 40204 Düsseldorf, Germany

[2]United States Department of Agriculture, Agricultural Research Service, Molecular Plant Pathology Laboratory, Beltsville, MD 20705, USA

[3]Biology Centre, CAS, v. v. i., Institute of Plant Molecular Biology, Branišovská 31, 37005 České Budějovice, Czech Republic

*To whom correspondence should be addressed. Tel: +49 211 8114597; Fax: +49 211 8115167;

Email: steger@biophys.uni-duesseldorf.de

Table S1: Parameters used in different programs

(a) PRINSEQ

Variable	Value
-min_len	18
-max_len	36
-ns_max_n	0
-lc_method	entropy
-lc_threshold	60
-trim_tail_right	2
-trim_tail_left	6

(b) TRIMMOMATIC

Variable	Value
SE	True
-phred33	True
ILLUMINACLIP	5:5:5
LEADING	3
SLIDINGWINDOW	4:30
MINLEN	18
AVGQUAL	30

(c) SEGEMEHL

Variable	Value
H	1
m	100
A	80, 85, 90, 95, 100

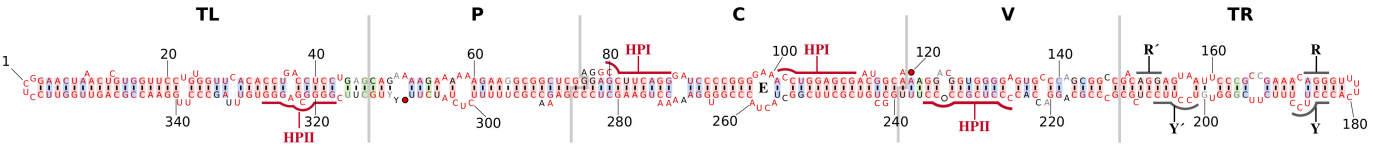
Table S2: Number of reads mapping to PSTVd variants at 90% and 100% accuracy.

Dataset	Cultivar	PSTVd	Count[a] (90%) / 10^5	Count[i] (100%) / 10^5	Ratio[b]	$\overline{\text{Ratio}}$[c]
D_s	Heinz1706	AS1	7.57	7.25	1.04	1.05
		C3	9.46	8.93	1.06	
		QFA→KF5M5	1.88	1.80	1.04	
		QFA	1.77	1.46	1.22	
D_p	Rutgers	I	5.74	5.31	1.08	1.10
		M→Var	1.29	1.16	1.11	
		M	1.29	1.16	1.11	
D_o	Heinz1706	RG1	7.01	6.19	1.13	1.14
	Moneymaker		4.27	3.76	1.14	
	Rutgers		5.10	4.34	1.18	
	UC82B		3.26	2.97	1.10	

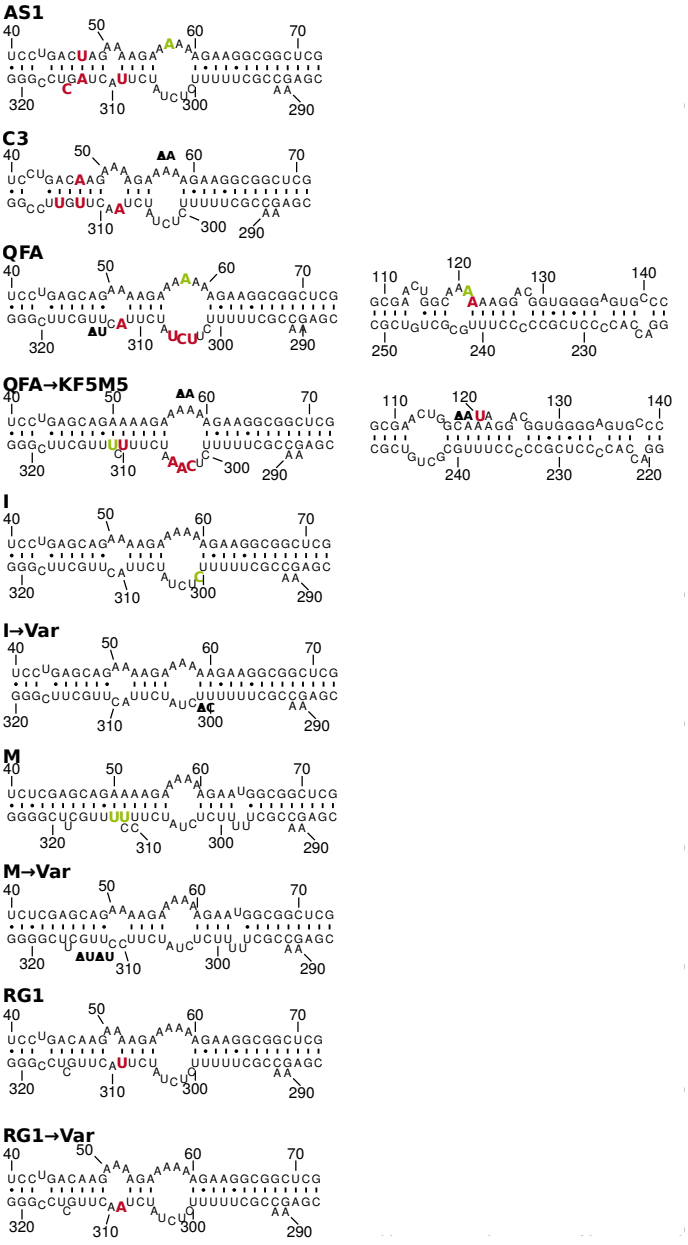
[a] number of reads mapping to the respective PSTVd variant at 90 and 100% accuracy, respectively.

[b] Ratio = count (90%)/count (100%)

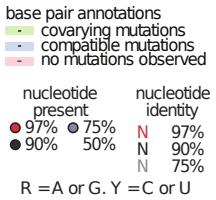
[c] mean ratio values; the mean value of the nine ratios is 1.10 ± 0.03 .



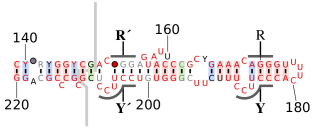
(A)



(D)



(B)



(C)

Figure S1: Structure of (+)-stranded PSTVd variants. (A) The consensus sequence and structure were predicted for an alignment of 266 variants (see Fig. S4) with CONSTRUCT[1] at 37 °C, excluding lonely base pairs, and drawn with R2R.[2] Borders of the five domains[3] are marked by gray lines (TL, terminal left; P, pathogenicity-related; C, central; V, variable; TR, terminal right). Base pairs on green and blue background are supported by covarying and compatible mutations, respectively; no mutations are observed for basepairs on red background; for further nucleotide annotations see (B). Nucleotides forming the extra-stable hairpins I (HPI) and II (HPII) in thermodynamically metastable structures are outlined. The two RY motifs in the TR critical for binding of the viroid RNA-binding protein VirP1 are outlined and marked by R, R' and Y, Y', respectively.[4] (C) This thermodynamically suboptimal structure of the TR domain emphasizes the similarity in sequence and structure of the two VirP1 binding sites. (D) Selected structural regions of PSTVd variants. Sequence differences between AS1 and C3, QFA and QFA→KF5M5, I and I→var, M and M→var, and RG1 and RG1→var, respectively, are marked in red (mutation), green (insertion), and by ΔN (deletion).

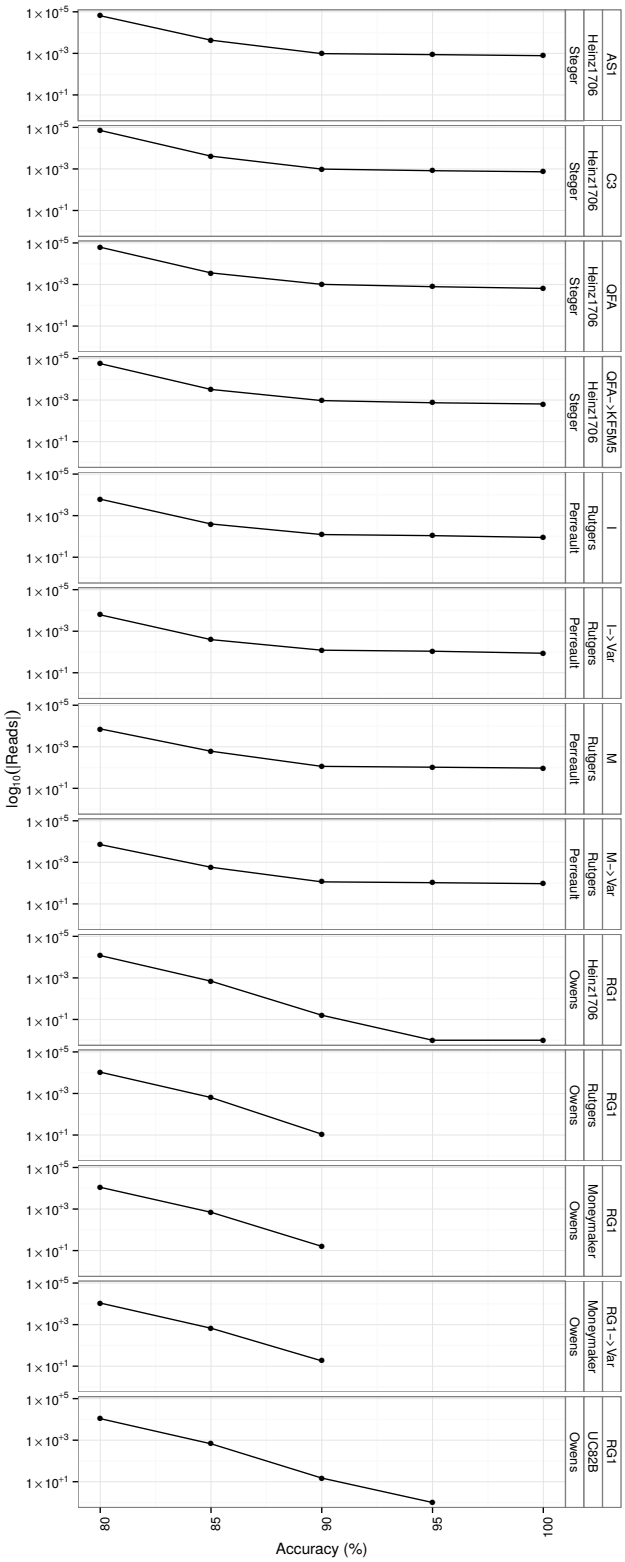


Figure S2: Mapping of small reads from mock-inoculated plants to the PSTVd variants at different accuracies.

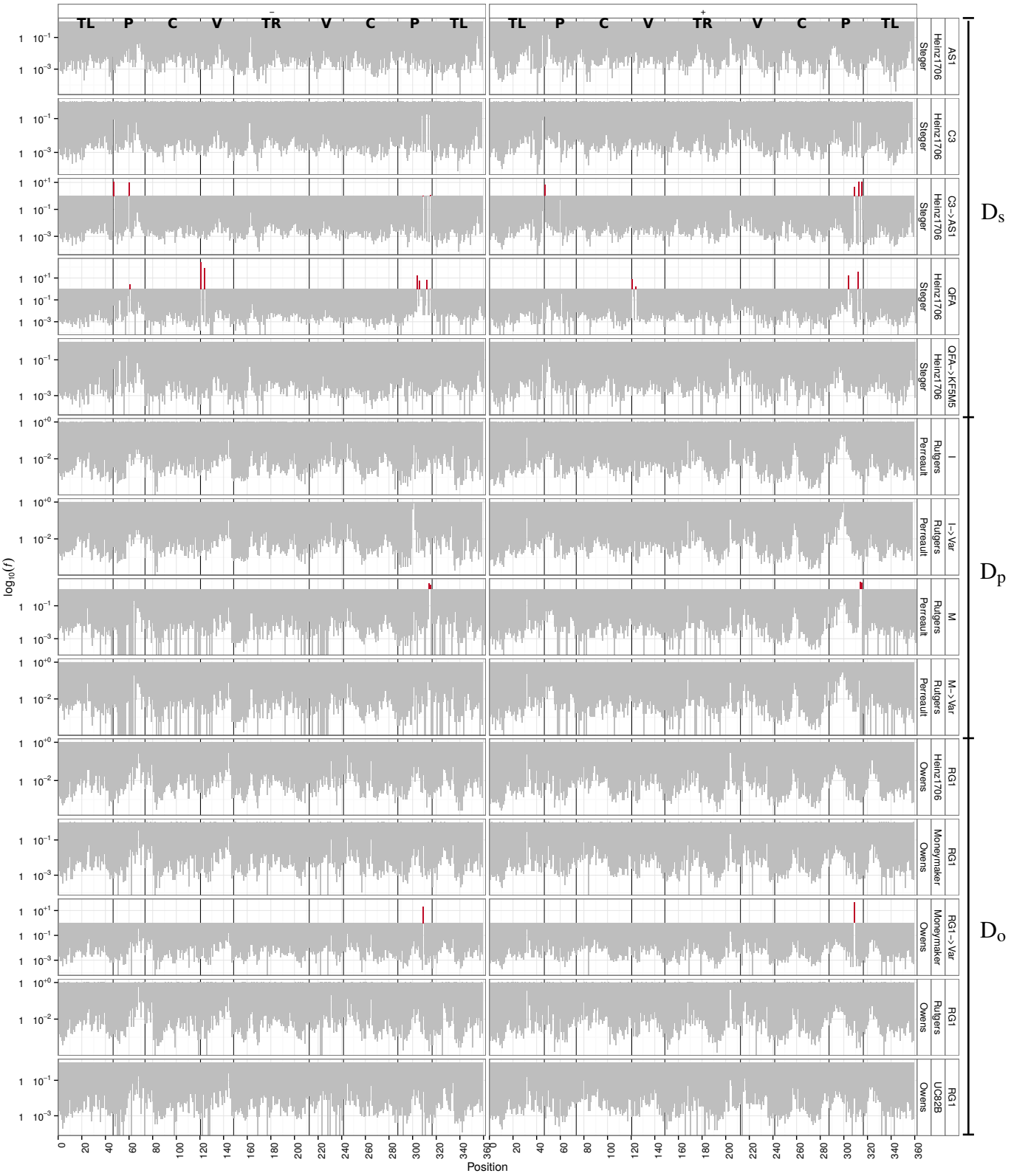


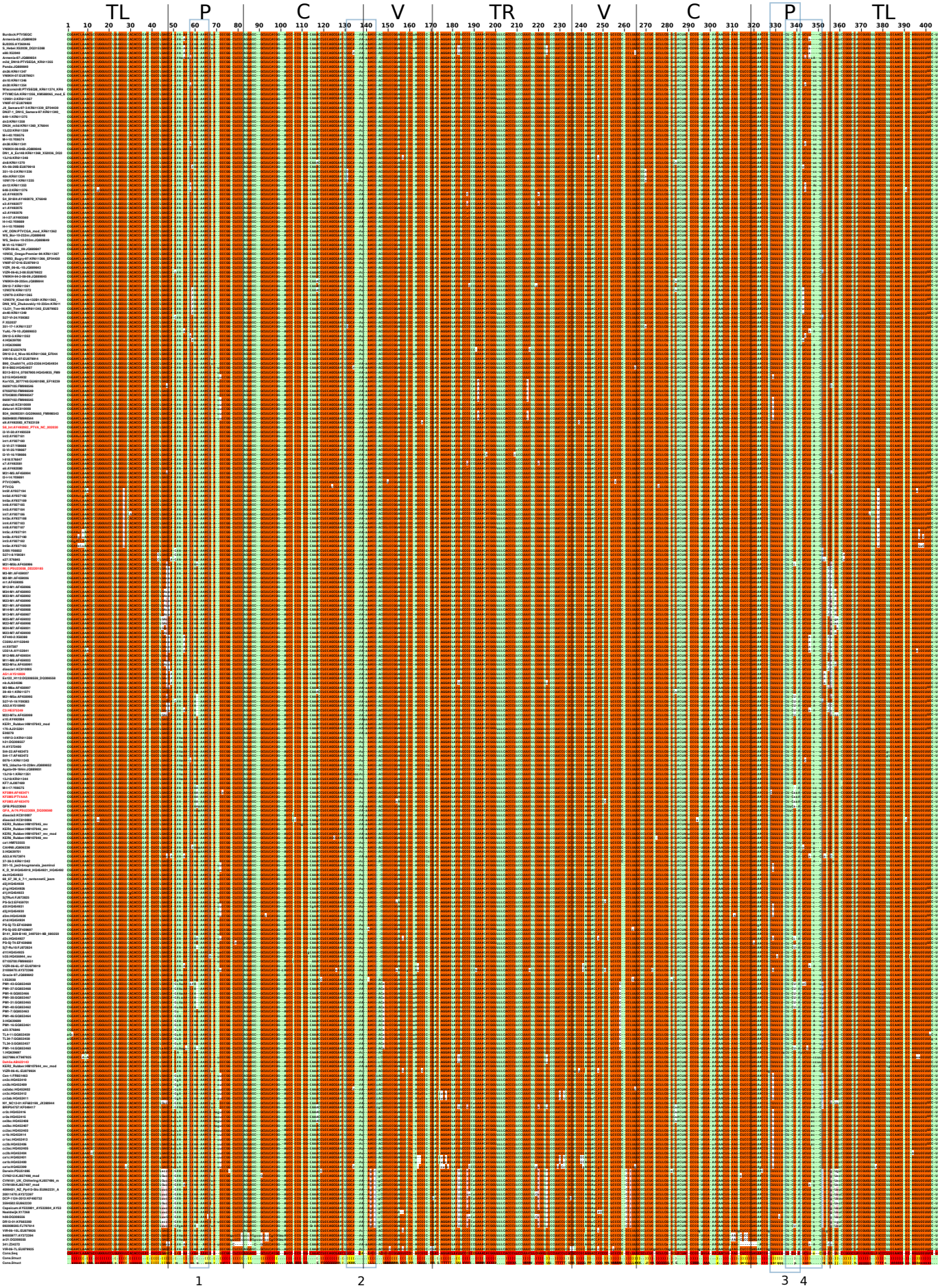
Figure S3: Ratio of read frequencies with a variant nucleotide relative to those with the reference nucleotide. Bars in red mark mutations that occur more often than the reference nucleotide, others are shown in gray. The y axis is in a logarithmic scale. Left and right columns show mappings to $(-)$- and $(+)$-stranded viroid sequences, respectively.

Table S3: Calculated error rate for two highly abundant *S. lycopersicum* miRNAs mapped with 90% accuracy. Due to the high rate of sequencing errors in the flanking positions, only positions 3–17 were considered for calculation of P_{error}.

Dataset	Condition	Variant	Cultivar	P_{error} / 10^{-4} sly-miR159	sly-miR162	
D_s	Healthy		Heinz1706	7.00	1.00	4
	Infected	AS1		14.8	4.84	7.94 ±4.23
		C3		9.39	4.02	
		QFA		10.0	4.59	
D_p	Healthy		Rutgers	6.00	5.00	5.5
	Infected	I		6.14	3.97	4.69 ±1.30
		M		5.36	3.27	
D_o	Healthy		Heinz1706	10.0	4.00	6.75 ±2.55
			Moneymaker	10.0	4.00	
			Rutgers	8.00	5.00	
			UC82B	8.00	5.00	
	Infected	RG1	Heinz1706	6.81	4.99	6.14 ±1.97
			Moneymaker	8.33	2.94	
			Rutgers	8.05	4.58	
			UC82B	8.04	5.41	

Figure S4: Structural alignment of 266 different PSTVd sequences. For figure see next page.
PSTVd sequences were retrieved from GenBank via BLAST and search in GenBank definitions in August, 2016. Several sequences had to be modified to fit the standard sequence enumeration and/or reverse complemented; for these sequences, the tags '_mod' and '_rev' were added. Identical sequences were removed but their GenBank loci were kept. Partial sequences were removed. The alignment of the remaining 266 unique sequences was performed with MAFFT/X-INS-I.[5] The alignment was drawn by CONSTRUCT.[1] A Vienna-formatted file with aligned sequences is available at http://www.biophys.uni-duesseldorf.de/~steger/final2supplement_motifs.vie .
Left: the sequence names consist of names as given in GenBank definition and comment plus GenBank ID. The sequence names of variants Intermediate (S8, Int, IDs AY492082, PTVA, NC_002030), RG1 (IDs PSU23058, DD220185), AS1 (AY518939), C3 (HE575349), KF5M5 (PTVAAA), QFA (Ar74, IDs PSU23059, DQ308560), and Dahlia (AB623143) are highlighted in red.
From top to bottom: borders of the five domains[3] are marked by gray lines (TL, terminal left; P, pathogenicity-related; C, central; V, variable; TR, terminal right). Sequences [with background colors green for loops, red for consensus base pairs, pink for consensus base pair changes (covarying pairs), and white for non-base pairs in paired regions], the consensus sequence, and the consensus structure in bracket-dot notation and character-encoded (both with background colors from white to red proportional to sequence conservation resp. pairing probability).
The "motifs" by which variants QFA and KF5M5 differ from each other are marked by cyan boxes; these motifs, number of PSTVd variants containing the motif, and the motif positions in the alignment are listed in the following table.

Motif		#	Position	Variants
1	GAAAAAA-G	239	58 – 66	KF5M3–5, Intermediate, RG1, Dahlia, AS1
	GAAAAAAaG	5	58 – 66	QFA
2	GGCAAu-AAGG	86	131 –144	KF5M3–5
	GGCAAaAAAGG	95	131 –144	QFA, Intermediate, RG1, C3
3	CUUUUUCUcaaA	3	333 –346	KF5M3–5
	CUUUUUCUCU-A	254	333 –346	Intermediate, RG1, Dahlia, AS1, C3
4	CU-CaaAUCUU-uCu	3	340 –356	KF5M3–5
	CU-CU-AUCUU-A-C	93	340 –356	Intermediate, RG1, AS1
	CUUCU-AUCUU-A-C	1	340 –356	QFA
	1+2	79		
	1+2+3+4	3		KF5M3–5


TL
P
C
V
TR
V
C
P
TL
1
2
3
4

References

1. Wilm A, Linnenbrink K, Steger G. ConStruct: Improved construction of RNA consensus structures. BMC Bioinformatics 2008; 9:219; PMID: 18442401; http://dx.doi.org/10.1186/1471-2105-9-219.

2. Weinberg Z, Breaker R. R2R—software to speed the depiction of aesthetic consensus RNA secondary structures. BMC Bioinformatics 2011; 12:3; PMID: 21205310; http://dx.doi.org/10.1186/1471-2105-12-3.

3. Keese P, Symons RH. Domains in viroids: evidence of intermolecular RNA rearrangements and their contribution to viroid evolution. Proc Natl Acad Sci USA 1985; 82:4582–86; PMID: 3860809.

4. Gozmanova M, Denti M, Minkov I, Tsagris M, Tabler M. Characterization of the RNA motif responsible for the specific interaction of potato spindle tuber viroid RNA (PSTVd) and the tomato protein Virp1. Nucleic Acids Res 2003; 31:5534–43; PMID: 14500815; http://dx.doi.org/10.1093/nar/gkg777.

5. Katoh K, Toh H. Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. BMC Bioinformatics 2008; 9:212; PMID: 18439255; http://dx.doi.org/http://dx.doi.org/10.1186/1471-2105-9-212.