Supplementary Information

Files

- F1 - original data set in `Fasta (fasta)`
- F2 - pre-filtered data set in `Fasta (fasta)`
- F3 - phylogenetic tree in `Newick (nwk)` and `Nexus (nex)`
- F4 - representative genomes in `Fasta (fasta)`
- F5 - `AnchoRNA` output in `General feature (gff)` (nucleotide and protein sequences)
- F6 - nucleotide alignment in `ClustalW (aln)`, `Fasta (fasta)`, and `Stockholm (stk)` (latter with additional annotations such as RNA secondary structures and genes)
- F7 - nucleotide alignment of the CDS in `ClustalW (aln)`, `Fasta (fasta)`, and `Stockholm (stk)` (latter with additional annotations such as RNA secondary structures and genes)
- F8 - protein alignment in `ClustalW (aln)`, `Fasta (fasta)`, and `Stockholm (stk)`
- F9 - nucleotide alignment of the 5' UTR (51 sequences) in `Stockholm (stk)`, `ClustalW (aln)`, and `Fasta (fasta)`
- F10 - nucleotide alignment of the 3' UTR (52 sequences) in `Stockholm (stk)`, `ClustalW (aln)`, and `Fasta (fasta)`
- F11 - nucleotide alignment of the 3' UTR of species A (10 sequences) in `Stockholm (stk)`
- F12 - nucleotide alignment of the 3' UTR of species C (10 sequences) in `Stockholm (stk)`
- F13 - nucleotide alignment of the 3' UTR of species K, S, and the Zikole strain (7 sequences) in `Stockholm (stk)`

Clustering

Table S1. Clustering results using `ViralClust`. # cluster – number of clusters (with a size greater than one); min. size – size of smallest cluster; max. size – size of largest cluster; avg. size – average cluster size; median size – median cluster size; d. centroid – average phylogenetic distance between two representative genomes; unclustered – number of unclustered sequences (= cluster with only one element or determined as unclustered by `HDBSCAN`); cps – average cluster per species; cpg – average cluster per genus. The last two pieces of information give a measurement of how much a single species/genus is divided into different groups.

algorithm	# cluster	min. size	max. size	avg. size	median size	unclustered	d. centroid	cps	cpg
`cd-hit-est`	68	2	98	10.10	3	69	1.19	2.46	41.0
`HDBSCAN`	53	6	40	12.89	10	73	0.98	2.39	25.5
`MMSeqs2`	65	2	106	10.55	3	70	1.28	2.24	37.0
`sumaclust`	63	2	101	9.86	4	56	1.34	2.33	43.0
`vclust`	64	2	107	10.98	3	53	1.27	2.03	42.0

Representative genomes

Table S2. Selection of representative genomes.

RefSeqs	`NC_024018, NC_023176, NC_018713, NC_076029, NC_001461, NC_012812, NC_003678, NC_035432, NC_003679, NC_077024, NC_025677, NC_038964, NC_038912, NC_076032, NC_039237, NC_077000, NC_077001, NC_077015, NC_077023, NC_077026, NC_030653, NC_002657, MZ664274`
Reduce overrepresentation	`KT875160, AB894423, HQ174297, MH231142, LC649064, LT158404, JX297520, HQ174294, OR004808, MH221026, KC695814, HM237795, MH885413, MT799517, LT593753, MK121886, MK093249, MN558876, MW528229, JQ612704, KJ463422, AJ133738, AF091605`
Outlier	`MW256672, MN025505, ON684360, OM030319, JQ799141, OU592965, OM030320`
Highly similar sequences (temporarily removed)	`NC_076029, KC853440, JX419398, KT951841, ON165517, KX577637, AF526381, NC_076032, NC_039237, MH231152, KT875135, HG426490, MK599227, NC_038912, NC_002657, GU233732, KP343640, KC533775, AY805221, AF099102, AY775178, NC_038964, MH221025, MN099165, MN584738`

Jingmen *Crocidura shantungensis* genome The Jingmen sequence (isolate SYS_SheQu, accID OM030319) has been reported to be a *Pestivirus* 1 isolate in the NCBI database (accessed March 17, 2025). Although some of the protein-coding regions are alignable with high conservation to *Pestivirus* species (protein C, NS3, NS4A, NS4B, NS5B) other regions are harder to align computationally or manually (protein N^{pro}, E^{rns}, E1, E2, NS5A). Protein NS2 seems to have very little similarity to other *Pestivirus* species. The non-essential region (directly downstream of the 5' UTR) is very short (similar only to species M). The supposed 5' UTR contains neither any single RNA secondary structure element of the IRES, nor D I, nor any conserved motif, which leads us to the conclusion that either (a) this region is a remnant of the non-essential region (however, it is not translatable into a protein) and the 5' UTR of the Jingmen sequence is absent, (b) a chimeric sequence is present, or (c) it might be a sequencing artifact. For this reason, we decided to truncate the Jingmen sequence upstream of the start codon.

5' UTR

Additional information. Giangaspero *et al.* [66] claimed that for species D a phylogenetic reconstruction was possible only based on the 5' UTR. We generated a phylogenetic representation using SplitsTree for 51/55 pestivirus genomes (see Fig. S1, including all genomes with a present 5' UTR) and could confirm a high similarity of the split graph based on only the 5' UTR of pestiviruses compared to the tree based on entire genomes.

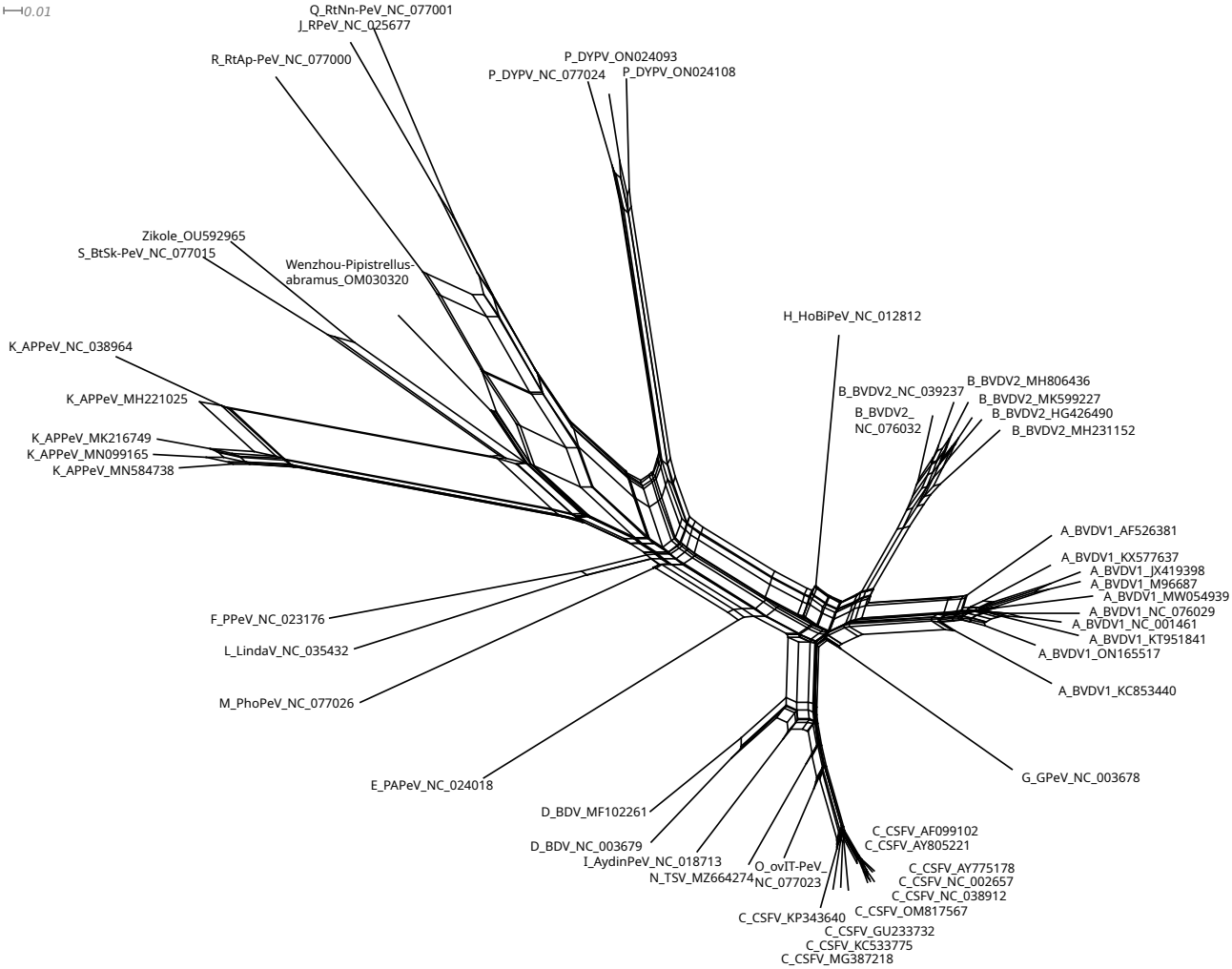

Figure S1. Phylogenetic reconstruction based on the 5' UTR alignment. The tree is based on the MAFFT alignment of the 5' UTR sequence of the representative genomes (if present) and served as input for SplitsTree to construct a split graph using the neighbor-net algorithm. The overall structure of the tree stays the same compared to the whole-genome phylogeny, see Fig. 1. Isolates of one species cluster together in one subtree, e.g., A, B, C. Species G and H are not clustered right next to each other, but still close. Species F, L, M, and E are still close to the root but located on the other side of the tree in this visualization. However, the placement of Wenzhou is different; instead of being between species F and L, it is now between species S and R.

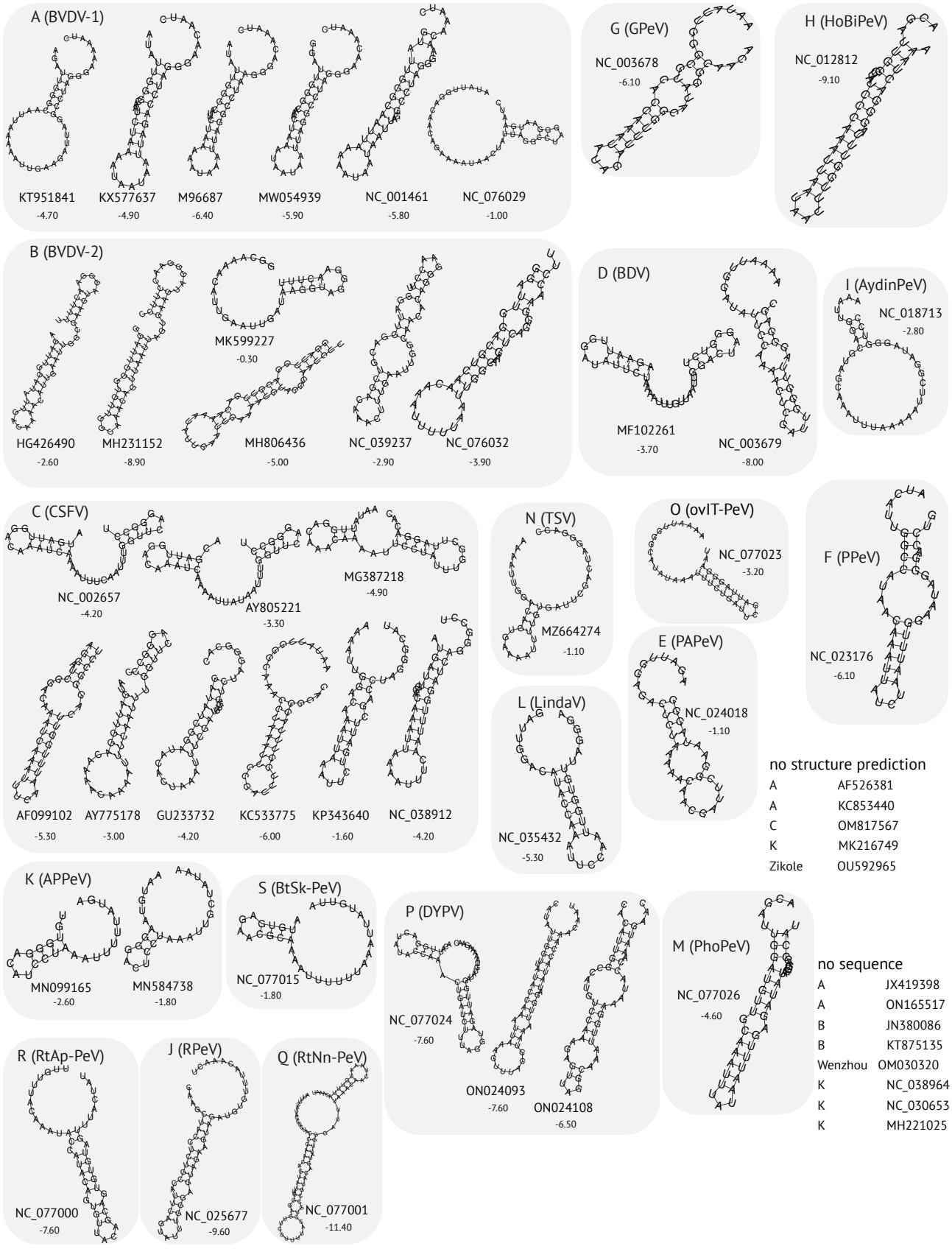

Figure S2. Single-sequence RNA secondary structure prediction of the region between SL I and SL II in the 5' UTR. MFE in kcal/mol of each structure is displayed below the Genbank ID.

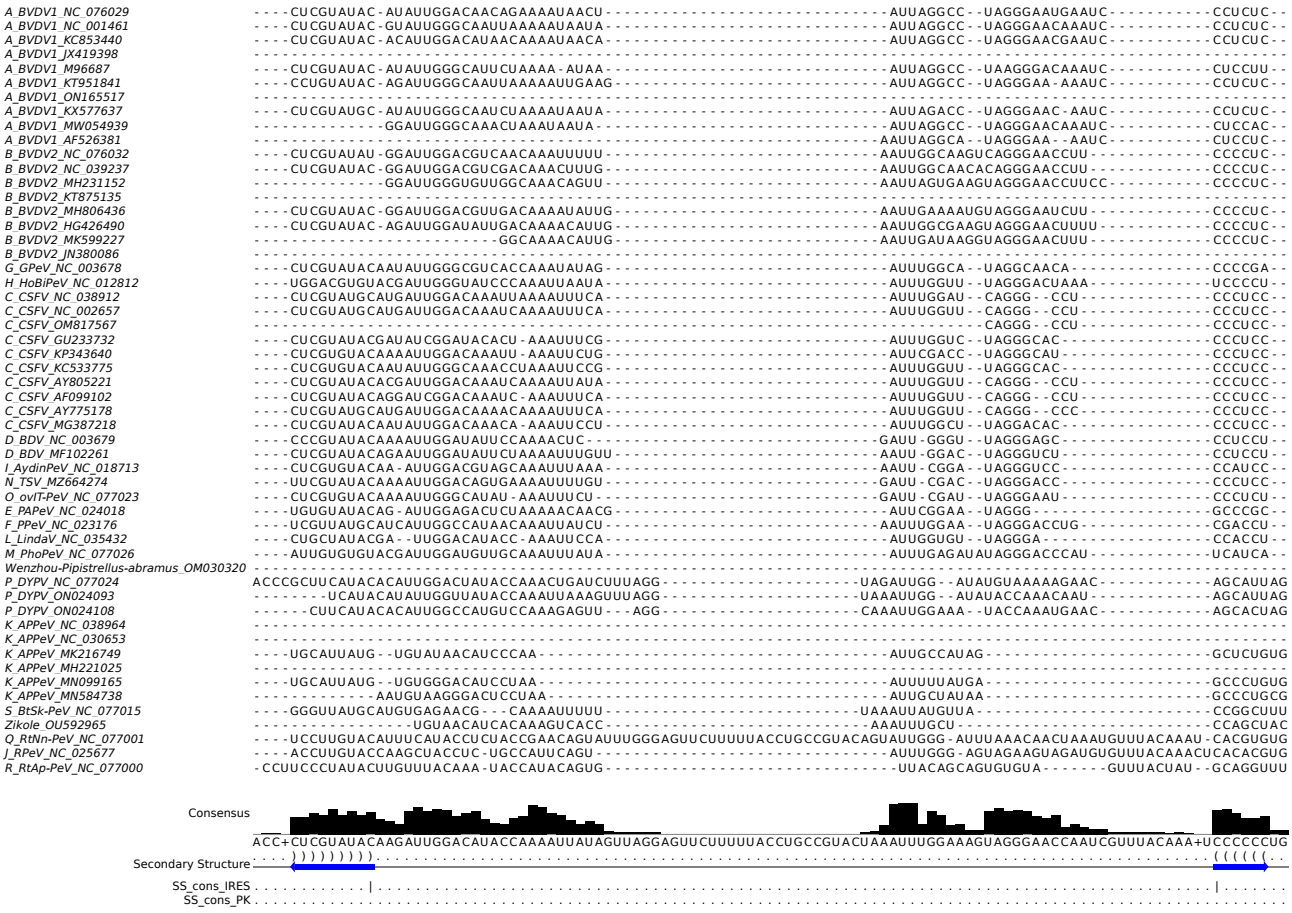

Figure S3. Multiple sequence alignment of the region between SL I and II in the 5' UTR (see Fig. 3 and S2).

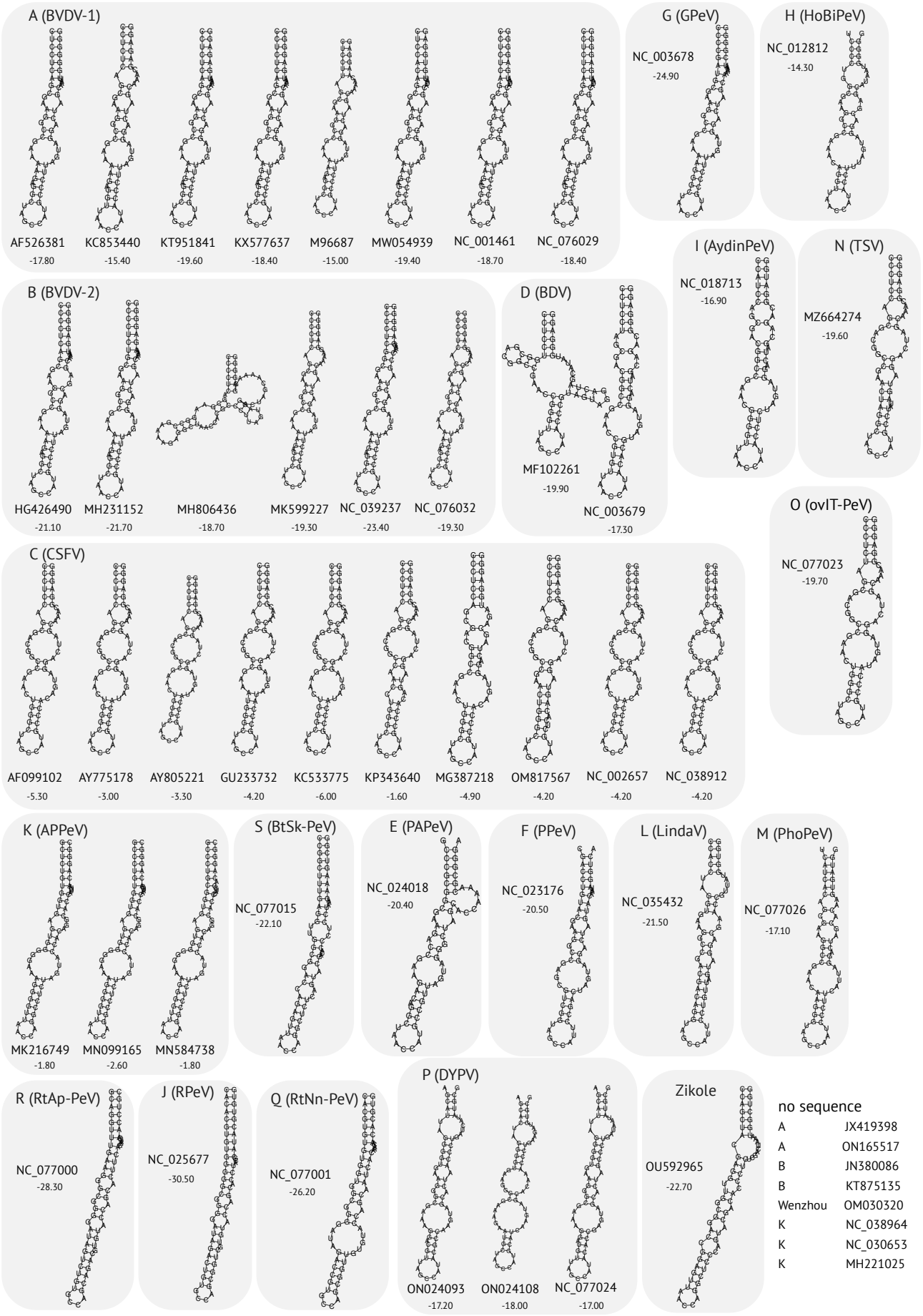

Figure S4. Single-sequence RNA secondary structure prediction of SL II in the 5'UTR using `RNAfold` with additional parameter for structural constraints forcing the four nucleotides in the hairpin loop to be unpaired. MFE in kcal/mol of each structure is displayed below the Genbank ID.

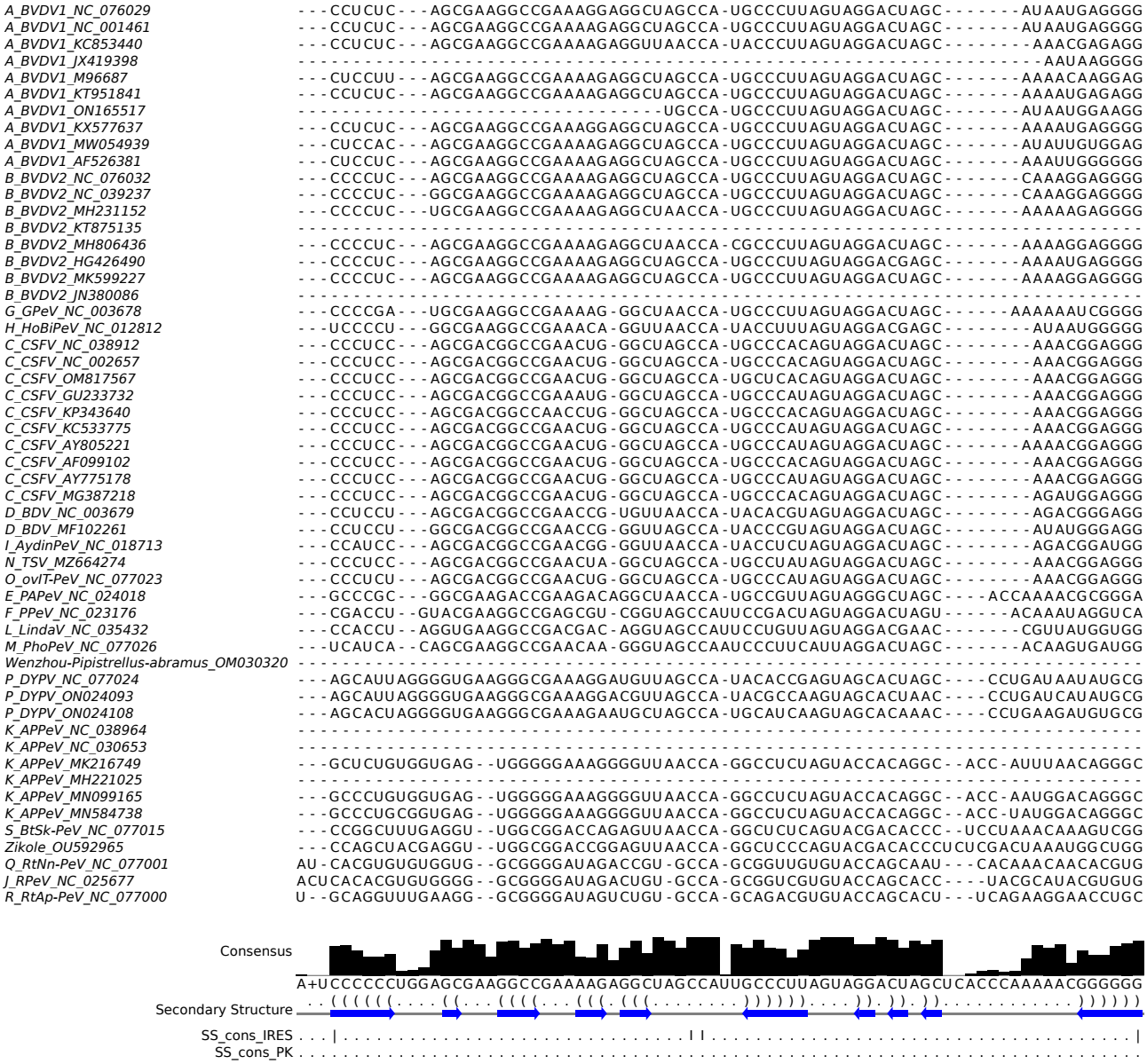

Figure S5. Multiple sequence alignment of SL II in the 5' UTR (see Fig. 3 and S4).

3' UTR

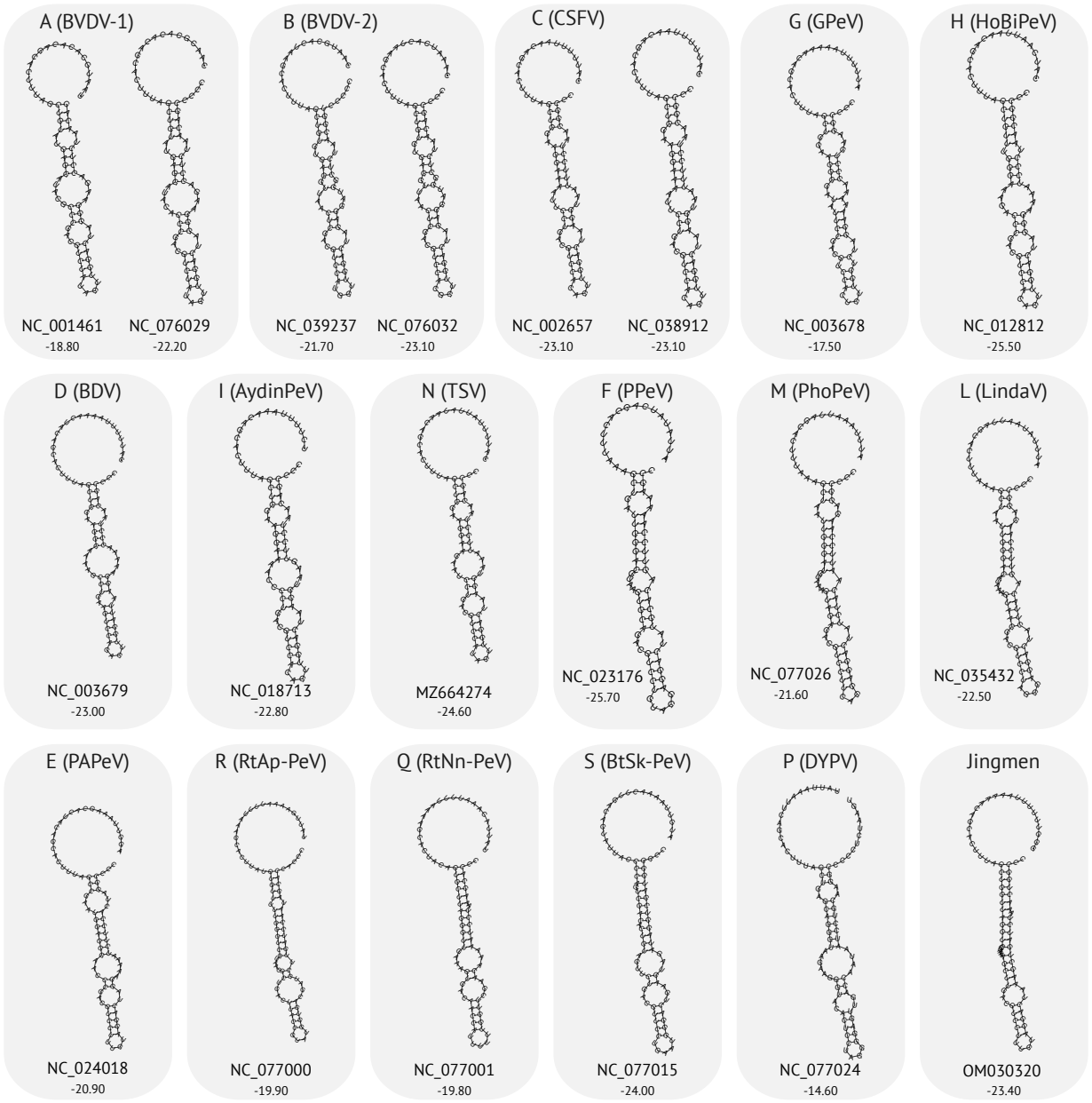

Figure S6. Single-sequence RNA secondary structure prediction of SL I in the 3' UTR using `RNAfold` with additional parameter for structural constraints forcing the miR-17 biniding site to be unpaired. MFE in kcal/mol of each structure is displayed below the Genbank ID.

New structure candidates

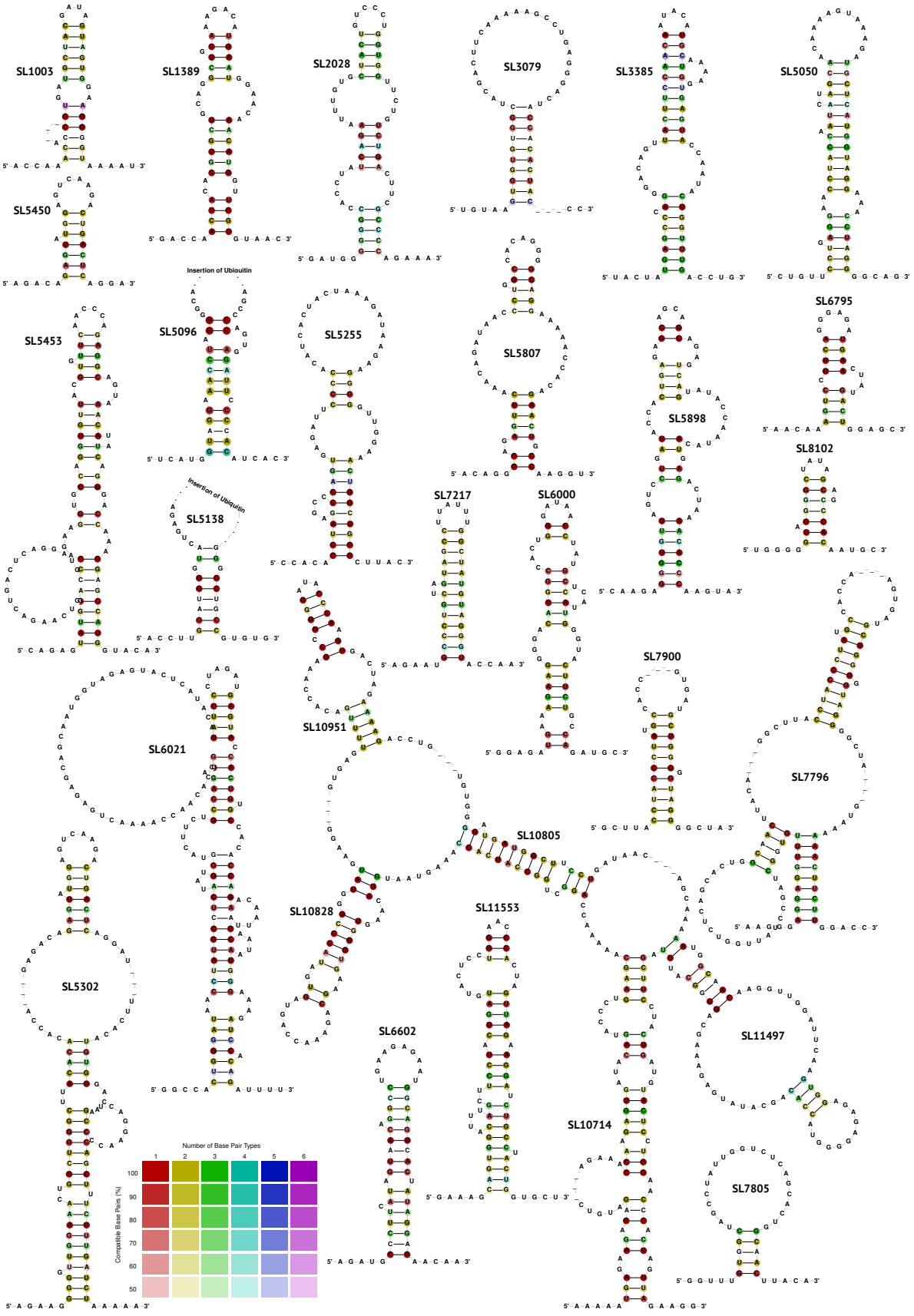

Figure S7. 29 novel conserved RNA secondary structure candidates from protein-coding region of the *Pestivirus* alignment.

Potential Primer Sites

Using our tool `AnchoRNA`, we identified a set of highly conserved regions at the nucleotide level across pestivirus sequences, which might serve as a potential primer in future experiments. Our 55 representative genomes served as input. To be considered a potential primer, we set the following parameters for `AnchoRNA`: minimum length of 18 nt with a maximum difference of 3 nt to the guiding sequence while at least 50% of the sequences have to contain the conserved region. These regions represent promising candidates for primer design and may serve as molecular markers for pestivirus detection via PCR (see Fig. S8). The corresponding list of conserved anchors for each sequence is summarized in Tab. S3. The table shows how many of the 55 representative pestivirus sequences contain each conserved region, indicating their broad applicability across pestivirus species. We emphasize that the proposed primers have not been experimentally validated. Rather, we present these conserved regions as potential starting points for the design and evaluation of molecular diagnostics. Experimental testing is required to assess their sensitivity, specificity, and overall performance.

Anchor No.	12	14	34	38	39	59	71	72	74	77	81
#Seq	53	46	43	44	38	39	46	52	51	45	38

Table S3. Detection of representative pestivirus sequences (out of 55) using primers derived from selected anchors for pestivirus identification. This table holds all potential primer sites for at least pestiviruses A, B, and C, see Fig. S8.

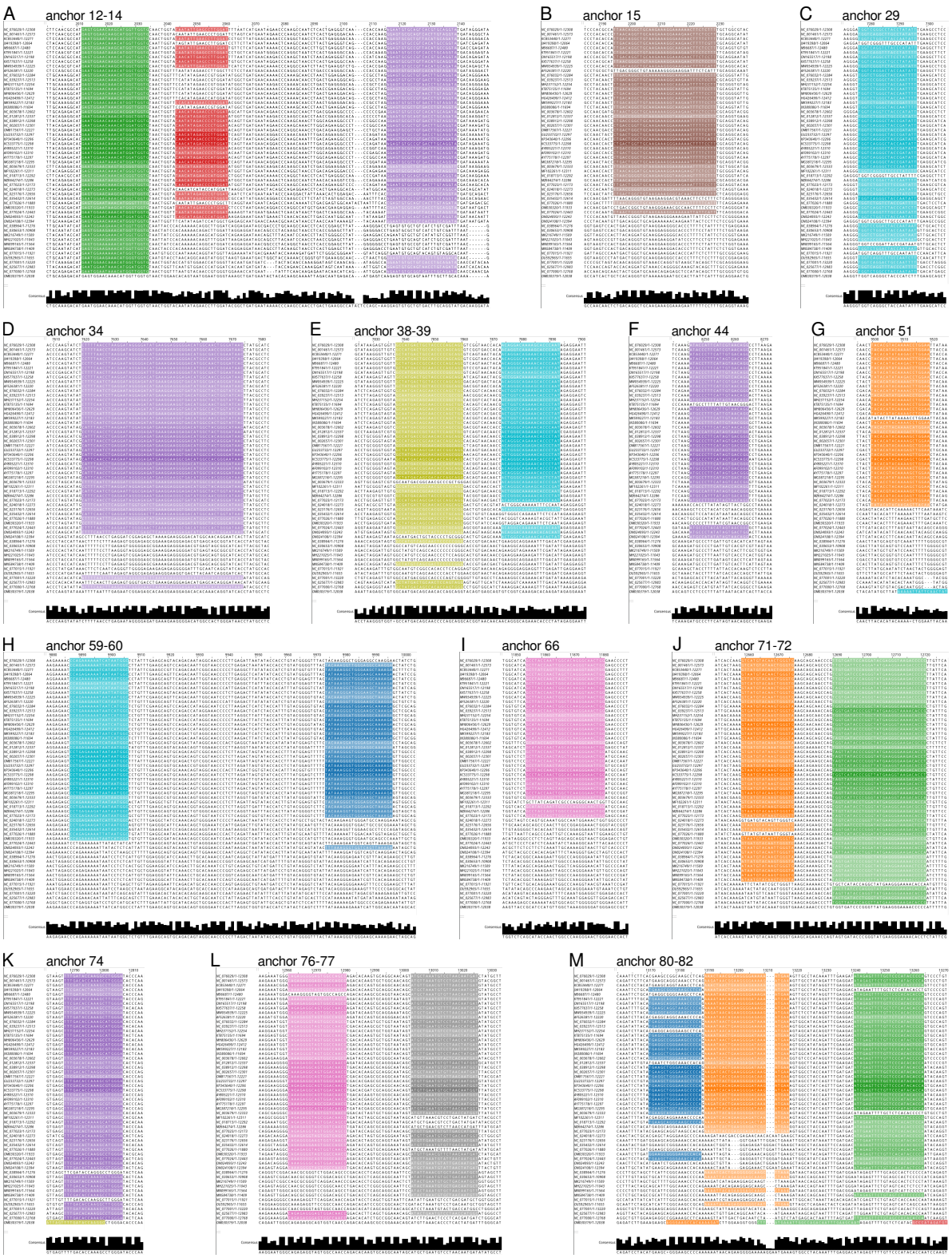

Figure S8. Potential primer sites for pestivirus detection identified by `AnchoRNA`.