Virus Research 260 (2019) 38–48

Contents lists available at ScienceDirect

Virus Research

journal homepage: www.elsevier.com/locate/virusres

Discovery of highly divergent lineages of plant-associated astro-like viruses sheds light on the emergence of potyviruses

Chris Lauber[a,b,c,*], Michael Seifert[a], Ralf Bartenschlager[b,d], Stefan Seitz[b,d]

[a] Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, 01307 Dresden, Germany
[b] Division of Virus-associated Carcinogenesis, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
[c] European Virus Bioinformatics Center (EVBC), 07743 Jena, Germany
[d] Department of Infectious Diseases, Molecular Virology, Heidelberg University, 69120 Heidelberg, Germany

ARTICLE INFO

Keywords:
Virus discovery
Virus evolution
Astrovirus
Potyvirus
Genome architecture

ABSTRACT

RNA viruses are believed to have originated from a common ancestor, but how this ancestral genome evolved into the large variety of genomic architectures and viral proteomes we see today remains largely unknown. Tackling this question is hindered by the lack of universally conserved proteins other than the RNA-dependent RNA polymerase (RdRp) as well as a limited RNA virus sampling. The latter is still heavily biased towards relatively few viral lineages from a non-representative collection of hosts, which complicates studies aiming to reveal possible trajectories during the evolution of RNA virus genomes that are favored over others.

We report the discovery of 11 highly divergent lineages of viruses with genomic architectures that resemble those of the astroviruses. These genomes were initially identified through a sequence homology search in more than 6600 plant transcriptome projects from the Sequence Read Archive (SRA) using astrovirus representatives as query. Seed-based viral genome assembly of unprocessed SRA data for several dozens of the most promising hits resulted in two viral genome sequences with full-length coding regions, nine partial genomes and a much larger number of short sequence fragments. Genomic and phylogenetic characterization of the 11 discovered viruses, which we coined plastroviruses (plant-associated astro-like viruses), showed that they are related to both astro- and potyviruses and allowed us to identify divergent Serine protease, RdRp and viral capsid domains encoded in the plastrovirus genome. Interestingly, some of the plastroviruses shared different features with potyviruses including the replacement of the catalytic Ser by a Cys residue in the protease active site. These results suggest that plastroviruses may have reached different points on an evolutionary trajectory from astro-like to poty-like genomes. A model how potyviruses might have emerged from (pl)astro-like ancestors in a multi-step process is discussed.

1. Introduction

RNA viruses utilize a variety of genomic architectures for the expression of their genes and for regulation of that expression (Dolja and Koonin, 2018; Koonin, 1991; Maia et al., 1996; Whelan et al., 2004). A recurrent genomic organization seen in different single-stranded, positive-sense RNA (ssRNA+) viruses is characterized by two partially overlapping 5′-proximal open reading frames (ORFs) and one or several additional ORFs encoded in the 3′ part of the genome. This architecture is employed by astro- and nidoviruses and members of several other virus families (Ali et al., 2014; Gorbalenya et al., 2006; Mäkinen et al., 1995; Willcocks et al., 1994). Genes in the 3′-proximal region, which in the case of astroviruses contains a single ORF2 that encodes the coat protein forming icosahedral capsids, are typically expressed via production of subgenomic RNA species, while the two ORFs located upstream are translated using the genomic RNA as template. The latter are named ORF1a and ORF1b in the case of astroviruses and encode a polyprotein with transmembrane, Serine protease and genome-linked viral protein (VPg) domains and an RNA-dependent RNA polymerase (RdRp), respectively. Translation of the genomic RNA may stop at the ORF1a termination codon resulting in polyprotein 1a (pp1a) or continue after -1 ribosomal frameshifting (RFS) at a slippery sequence located in the ORF1a/b overlap to produce pp1ab (Jiang et al., 1993). The rate of RFS determines the relative amounts of protein expressed from ORF1a and ORF1b which, together with subgenomic RNA production, enables the virus to regulate the intracellular concentrations of

[*] Corresponding author at: Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, 01307 Dresden, Germany.
E-mail address: chris.lauber@tu-dresden.de (C. Lauber).

https://doi.org/10.1016/j.virusres.2018.11.009
Received 31 August 2018; Received in revised form 15 November 2018; Accepted 15 November 2018
Available online 16 November 2018

its gene products (Brierley et al., 1987; Brierley and Vidakovic, 2003; van Marle et al., 1995). Astroviruses infect various vertebrate species including humans (Madeley and Cosgrove, 1975; Shi et al., 2018). Astro-like viruses have recently been detected in invertebrates (Shi et al., 2016), but it is not known whether similar viruses can infect fungi, protists or plants.

Another, very different genome architecture is characterized by a single ORF encoding a large polyprotein. This relatively simple monocistronic organization is employed by many different ssRNA + viruses including most members of the *Potyviridae* family (Wylie et al., 2017). The potyvirus polyprotein is cleaved by a set of virus-encoded proteases including a Serine-like cysteine protease (PRO_{Ser}) that is responsible for processing most of the polyprotein cleavage sites (Carrington et al., 1989; Gorbalenya et al., 1989; Hellmann et al., 1988; Verchot et al., 1991). The proteolytic cleavage by PRO_{Ser}, whose rate is dependent on the sequence context at the cleavage site (Rodamilans et al., 2018; Tözsér et al., 2005), presents an alternative means of regulating the relative concentrations of viral proteins compared to that used by the astroviruses. Downstream of PRO_{Ser}, the potyvirus genome encodes an RdRp domain and a coat protein forming filamentous nucleocapsids of helical symmetry, and the relative positions of these three proteins in the viral genome resemble those of their counterparts in astroviruses. Several additional domains including the other proteases, a VPg and a helicase are located upstream of PRO_{Ser} in the potyvirus polyprotein. In addition, a trans-frame protein (P3N-PIPO), is expressed in low quantities by a transcriptional slippage mechanism (Olspert et al., 2015). Potyviruses currently constitute the largest family of plant RNA viruses and include several economically important pathogens (Rybicki, 2015; Wylie et al., 2017).

If and how transitions from one genome architecture to another have occurred during RNA virus evolution is poorly understood, and approaching this delicate question is complicated by our fragmentary knowledge of the existing genetic diversity of viruses (Rosario and Breitbart, 2011; Suttle, 2007). To tackle this, extensive next generation sequencing (NGS)-based virus discovery efforts targeting specific host taxa or metagenomes have been undertaken during the recent years, but in the case of eukaryotic viruses, the discovered genome sequences mostly fall within the genetic diversity of known virus groups, while novel virus families are reported very rarely (Drexler et al., 2012; Phan et al., 2011; Shi et al., 2016, 2018). We have recently shown that primary NGS data from studies unrelated to virus research, which is obtainable in very large quantities from the Sequence Read Archive (SRA) (Leinonen et al., 2011), can be exploited for discovering unknown vertebrate viruses including members of novel virus families (Lauber et al., 2017). We now extended this approach to plant SRA projects in the current study. We discovered almost a dozen highly divergent viral genomes that show striking similarities to both astro- and potyviruses, suggesting unappreciated evolutionary links between these two important viral families.

2. Materials and methods

2.1. Public sequence data

We utilized unprocessed sequencing data from the Sequence Read Archive (SRA) hosted by the National Center for Biotechnology Information (NCBI) (Leinonen et al., 2011). We limited our search to transcriptome projects studying plants from various taxa (see Section 3.1 for taxonomic details). The NCBI Taxonomy Browser was utilized for selection of these projects. The SRA data was downloaded to and temporarily stored at the computing cluster Taurus hosted by the Centre for Information Services and High Performance Computing (ZIH) of the TU Dresden, where it was analyzed as described below.

2.2. Discovery of viral sequence fragments in transcriptome data

The SRA data sets were screened for the presence of unknown viral sequences using either BLAST (Altschul et al., 1990) or HMMER v3.1 (Eddy, 2011), respectively, with RdRp protein sequences of astrovirus representatives or protein profiles of RdRp alignments as query. Sequencing reads of an individual SRA project that were hit with an E-value below 10 were compared to the non-viral subset of the NCBI reference proteins (nr) database using blastx, and an E-value cut-off of 10^{-4} was used to filter out non-viral sequences. The remaining sequences were compared to the NCBI viral genomics database using tblastx and hits with an E-value below 1 were retained for manual inspection.

2.3. Genome assembly of discovered viruses

Sequencing adapters and low-quality bases were trimmed using Cutadapt (Martin, 2011). The assembly of viral sequences was done using a seed-based approach as implemented in GenSeed-HMM (Alves et al., 2016). An astrovirus RdRp or plastrovirus pp1b protein profile was used as seed in the GenSeed-HMM analysis. If a new virus was detected in several sequencing experiments of the same host species a super-assembly of the contigs was performed using CAP3. For selected sequencing experiments an additional, independent *de novo* assembly of the full set of reads was performed using SPAdes (Bankevich et al., 2012) or Trinity (Grabherr et al., 2011). Sequencing reads included in an assembly were mapped back to the respective contigs using Bowtie2 (Langmead and Salzberg, 2012) and assembly quality was assessed by visual inspection using Tablet (Milne et al., 2010).

The viral genome sequences were compared to all available non-vertebrate eukaryotic genome assemblies from the Whole Genome Shotgun (WGS) database (Benson et al., 2013) to exclude that the discovered sequences represented endogenous viral elements. We employed blastn with the viral contigs as query and considered hits that included at least 50 nucleotides of the viral contig termini and showed at least 95% sequence identity and an E-value of 1×10^{-3} or better.

2.4. Proteome characterization

ORFs and encoded peptides were extracted using the EMBOSS package (Rice et al., 2000). Protein domains were annotated using HHpred (Zimmermann et al., 2017), blastp (Altschul et al., 1990) or through alignment with homologous astro- and potyvirus proteins by a human expert. We used the Protein Data Bank (PDB), the Structural Classification of Proteins (SCOP) and the NCBI Conserved Domains (CD) as databases during the HHpred searches (Bernstein et al., 1977; Marchler-Bauer et al., 2017; Murzin et al., 1995). Transmembrane domains were predicted using TMHMM v2.0 (Krogh et al., 2001). Jpred 4 was used for protein secondary structure prediction (Drozdetskiy et al., 2015).

We used PSI-BLAST (Altschul et al., 1997) with each plastrovirus p2 sequence as query to retrieve viral structural proteins from RefSeq (Pruitt et al., 2007) that shared significant similarity with the putative plastrovirus capsid proteins (CPs). The longest hit annotated as a structural protein was 1218 amino acids in length, and we excluded all hits above that size threshold as these likely constituted various polyproteins. After adding 96 potyvirus CPs obtained from RefSeq to the sequence collection, we used CLANS (Frickey and Lupas, 2004) to visualize the pairwise sequence similarities between the viral CPs as a network graph. In addition, SWISS-MODEL (Waterhouse et al., 2018), which implements a homology modelling approach based on experimentally determined structures, was used to predict the structural fold of the plastrovirus p2 sequences.

2.5. Phylogeny reconstruction

Astro- and potyvirus reference genomes and proteins were obtained from the NCBI Viral Genomes Resource (Brister et al., 2015). We added to this collection recently discovered astro- and astro-like vertebrate and invertebrate viruses (Shi et al., 2018, 2016). Multiple amino acid alignments were created and visualized with the help of SeaView (Gouy et al., 2010) which offers access to MUSCLE (Edgar, 2004) and Clustal (Chenna et al., 2003) for alignment computation. All alignments were inspected and curated by an expert.

The best fitting amino acid substitution model was selected using ProtTest v3.4 (Abascal et al., 2005) and used in subsequent phylogenetic reconstructions. Maximum likelihood trees were reconstructed using PhyML v3.0 (Guindon et al., 2010) with 100 bootstrap replicates. Bayesian trees were reconstructed using BEAST v1.8.0 (Drummond et al., 2012); two chains were run for 5 million steps using a relaxed molecular clock approach with log-normal distribution (Drummond et al., 2006) and a Yules speciation prior. Convergence of the BEAST runs was verified using Tracer and consensus trees were visualized using FigTree (tree.bio.ed.ac.uk/software/).

3. Results and discussion

3.1. Discovery of astro-like viral genomes in plant sequencing projects

To test our hypothesis about the existence of astro-like viruses that infect eukaryotic hosts other than animals, we screened public transcriptome data from various plant studies, obtained as unprocessed sequencing reads from the SRA. With this approach to virus discovery we take advantage of the fact that viral genomes will be sequenced as a by-product to sequencing the organism under study, if that organism was infected at the time of sampling. The sequence fragments produced during sequencing can then be assembled to full-length or partial viral genome sequences, depending on sequence coverage. The sheer number of data sets in the SRA – exceeding 3.1 million by the time of writing and originating from a large variety of organisms – offers an unprecedented resource that can be utilized for discovering unknown viruses, including those with highly divergent genomes. We have recently shown the value of this approach in a pilot study reporting the discovery of the non-enveloped nackednaviruses that constitute a so far unknown sister family to the enveloped hepadnaviruses (Lauber et al., 2017).

In an initial search for unknown astro-like viruses we employed tblastn to screen 5577 SRA data sets (referred to as SRA runs from here on) of land plants (*Embryophyta*, txid:3193) excluding flowering plants (*Magnoliophyta*, txid:3398), by using a representative set of RdRp protein sequences from 12 astrovirus species as query. We obtained hits in 1567 of these runs from which 269 and 54 had an E-value lower than 0.1 and 0.01, respectively. For the pine tree species *Pinus contorta* (SRR073389/90/91; E = 0.00013) we succeeded in assembling a 3467 nt contig (PCPLAV-1 in Fig. 1) that showed two longer partially overlapping ORFs resembling the ORF1a/b architecture seen in astroviruses and several other virus families. The most similar sequence in RefSeq (Pruitt et al., 2007), which included both host and viral sequences, was found to be the RdRp of a rat astrovirus (Genbank accession ADJ38390). This Blast hit was highly significant with an E-value of 6×10^{-14} and showed a local amino acid sequence identity of only 27%, indicating that this contig represented the partial genome sequence of an unknown astro-like virus. In addition, several SRA runs from a project about another pine species, *Pinus taeda*, obtained significant hits against the query sequences. From four of these runs (SRR1276184/88/90/91) we succeeded in assembling two additional contigs with astrovirus-like ORF1a/b-ORF2 genomic organization (PTPLAV-1 and PTPLAV-2 in Fig. 1). Although similar in architecture, the three discovered partial viral genomes showed only low degrees of sequence similarity between each other (amino acid identity less than 28%), indicating that we had discovered several unknown lineages of astro-like viruses.

To further increase the sensitivity of our SRA screening we constructed a multiple sequence alignment of predicted RdRp sequences of these three viruses and a multiple alignment of astrovirus RdRp sequences. These two alignments were used as query in a profile search against 6683 SRA runs from studies of seed plants (*Spermatophyta*, txid:58024) excluding two highly sampled taxa of flowering plants, the *Gunneridae* (txid:91827) and the *Poales* (txid:38820). This search resulted in 1893 hits from which 822 and 575 showed an E-value better than 0.1 and 0.01, respectively. We then attempted an assembly for the most promising hits ranked by E-value and the number of detected sequencing reads. From this analysis it became evident that we had accidentally discovered numerous viruses from known plant virus families, like for instance potyviruses, partitiviruses and picorna-like viruses (data not shown). However, among the assembled sequences were also eight additional genome sequences with astrovirus-like ORF1a/b-ORF2 architecture (Fig. 1). These included a third virus from a *P. taeda* run, three different sequences from data sets of the banana species *Musa acuminata*, one sequence each from studies of the fir tree species *Abies nordmanniana* and *Cunninghamia lanceolata*, and the flowering plants *Chelidonium majus*, and *Clematis apiifolia*. Two of the eleven genome sequences were found to be complete, covered the full-length coding region, and were 7646 and 6498 nucleotides in length, while the remaining nine sequences were partial. The eleven viral genome sequences are provided in **Supplementary data file 1** and their detailed genomic characterization is provided in Section 3.2 below. The mean coverage with sequencing reads per nucleotide position varied between the 11 assemblies and was in the range of 12 to 443. The high or very high coverage of > 20 reads per position observed for all except two genomes suggested that the viral genomes were actively replicating. Moreover, we did not find inactivating mutations resulting in premature stop codons or insertion/deletions and we also did not obtain significant hits in a comprehensive Blast search against the WGS database for any of the 11 discovered plastrovirus genomes (see Section 2.3 for details), showing that we had discovered exogenous viruses.

We were unable to assemble any contigs longer than few hundred nucleotides from another two dozen or so SRA runs that obtained promising hits during our profile search, showing that sufficient sequence coverage is essential for successful assembly. Moreover, we did not include an additional 100,000 SRA transcriptome data sets from flowering plants available from the SRA in our screens. Together, this indicated that the number and diversity of members of this new viral group are likely much larger than reported here.

The low degree of amino acid sequence identity among the new viruses, i.e. only 20.4–34.7% in the most conserved genome regions, let us to conclude that we had discovered several highly divergent astro-like viral lineages. For the sake of brevity, we refer to them as *plastroviruses* (plant-associated astro-like viruses, PLAVs). Although we found 9 of the 11 viruses in samples of tree species, we cannot exclude that these viruses infect members of the mycorrhiza or other microorganisms rather than the plants, as all samples in which we detected plastroviruses included material from the roots of the plants. A summary of the discovered plastrovirus genomes and associated samples is shown in Table 1.

3.2. Genomic and phylogenetic characterization of plastrovirus genomes

All plastrovirus genomes except for the incomplete genome of CLPLAV-1 showed a full-length ORF1b-like coding region that partially overlapped with an upstream ORF1a-like region whose reading frame was shifted one nucleotide downstream relative to that of ORF1b for eight of the ten genomes (Fig. 1). Sizes of the overlapping regions of the eight genomes were in the range of 49 to 310 nucleotides and contained one or several putative RFS signals (slippery sequences). The two

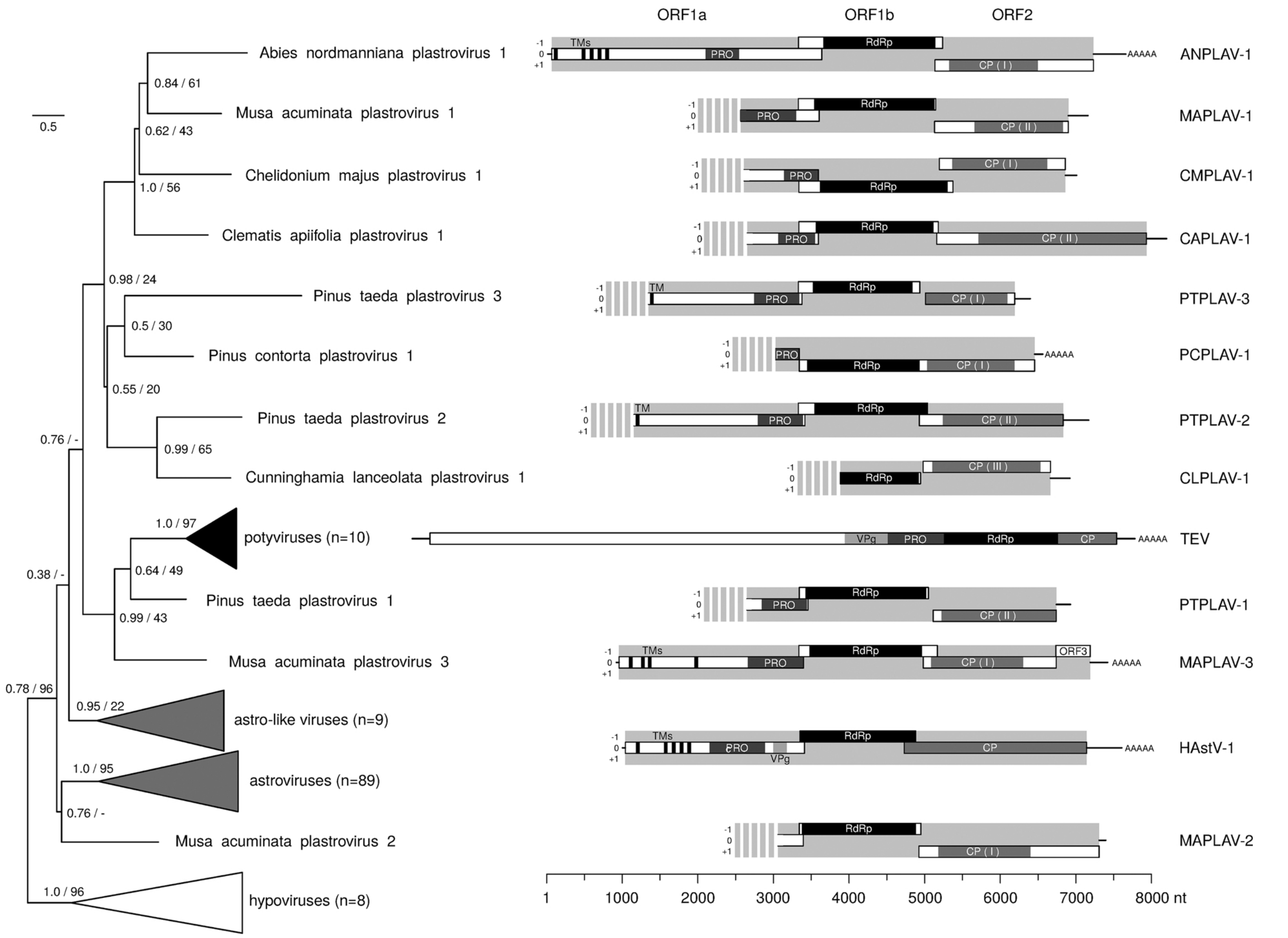

Fig. 1. Phylogenetic clustering and genomic organization of plastroviruses. The phylogeny of astro-, plastro- and potyviruses based on conserved RdRp regions was reconstructed using a Bayesian approach and rooted by inclusion of hypoviruses as outgroup. Numbers at branching points indicate posterior probability support values (left, decimal number) and bootstrap support values from an independent maximum likelihood (ML) reconstruction (right, integer) if the corresponding subtree was present in the ML phylogeny; minus signs otherwise. Internal branches are collapsed for four viral reference groups and the number of included viruses per group is shown in brackets. The scale bar is in average amino acids substitutions per site. To the right of the tree the genomic organizations of the 11 plastroviruses and a potyvirus (Tobacco etch virus, TEV) and astrovirus (Human astrovirus 1, HAstV-1) reference is shown. The light-gray background indicates coding regions, and three reading frames are discriminated except for the potyvirus genome. Striped background indicates plastrovirus genomes with incomplete 5′ termini. White rectangles show ORFs, which are defined here as coding regions flanked by stop codons. Predicted or known (in case of TEV and HAstV-1) protease (PRO), replicase (RdRp), VPg and capsid proteins (CPs) are highlighted using different shadings of gray, and transmembrane domains (TMs) as black bars. Other protein domains of astro- and potyviruses are not shown for the sake of clarity. Note that borders of plastrovirus protein domains reflect most conserved regions as predicted by the sequence homology searches and may actually differ for full-length proteins. Roman numerals indicate three plastrovirus CP types with respect to similarities with single jelly roll CPs of other viruses.

remaining partial genomes, CMPLAV-1 and PCPLAV-1, respectively, showed an unusual shift of ORF1a by one nucleotide upstream (instead of downstream) relative to ORF1b. We identified additional genomic features that deviate from those seen in astroviruses for both viruses (see below).

The 5’ portions of ORF1a were missing for eight of the assembled sequences, while ANPLAV-1 and MAPLAV-3 showed full-length ORF1a regions. Downstream of ORF1b, all plastrovirus genomes displayed ORF2-like coding regions that ended with a termination codon, indicating that the ORF2 regions were of full length. Most notably, ORF1b and ORF2 counterparts were found to be fused into a single large ORF in the case of PCPLAV-1. For three of the assembled plastrovirus genomes (ANPLAV-1, MAPLAV-3 and PCPLAV-1) we were able to retrieve polyA-tailed 3′ genome termini that were separated from ORF2 by short untranslated regions. In summary, this showed that we had discovered two nearly complete genomes (ANPLAV-1 and MAPLAV-3) with full-length coding regions and seven partial PLAV genomes with a genomic architecture which mirrored that of astroviruses, as well as two partial genomes with a non-canonical yet similar genomic organization compared to the former.

In the case of CMPLAV-1, we could verify correctness of the assembly in the unusual 260 nt ORF1a/b overlapping region (contig positions 727–986) which was covered by > 10 sequencing reads per position including properly paired read mates. It is tempting to speculate that this virus utilizes transcriptional instead of translational slippage for ORF1b expression, a mechanism that has been reported for potyviruses (Olspert et al., 2015) and different negative-sense RNA viruses (Cattaneo et al., 1989; Sanchez et al., 1996; Thomas et al., 1988). In this respect, we identified a putative slippage sequence (AAGAAAA) similar to that used by the potyviruses (GAAAAAA). Alternatively, +1 frameshifting during translation could be mediated through a second putative signal sequence (CCCCAAAA) in the overlap.

For the second virus with non-canonical ORF1a/b overlap, PCPLAV-1, we could not verify the assembly in this region, as it was covered only by two sequencing reads. The most striking feature of its genome, however, was the fusion of ORF1b- and ORF2-like regions. Through alignment of the predicted polypeptide (pp) sequence encoded by this fused ORF with predicted pp1b sequences of the other plastroviruses,

Table 1
Discovered plastrovirus genomes and associated SRA data sets and samples.

virus[a]	length[b]	coverage[c]	complete CDS	1a/b FS[d]	overlap length[b]	putative slippery sequences	SRA identifier	host type	sampled tissue
Abies nordmanniana plastrovirus 1	7646	44.7	yes	−1	310	CCCGGGG, CCCAAAAA	SRR5583210/1/3/6	fir tree	root
Chelidonium majus plastrovirus 1	4400	60.1	no	+1	260	CCCCAAAA, AAGAAAA	SRR341947	flowering plant	root
Clematis apiifolia plastrovirus 1	5550	54.4	no	−1	259	GGUUUUU, TATAAAA	SRR5657903/4	flowering plant	root, stem, leaf
Cunninghamia lanceolata plastrovirus 1	3038	14.9	no	/	/	/	SRR5282558	fir tree	root
Musa acuminata plastrovirus 1	4592	21.8	no	−1	277	CAAAAAA, ACCCCCC, AAGGGG, AGAAAAA	SRR2132844	banana tree	root
Musa acuminata plastrovirus 2	4336	25.9	no	−1	55	CAAAAAA	SRR2132844	banana tree	root
Musa acuminata plastrovirus 3	6506	442.7	yes	−1	64	CTTTAAAC	SRR2132843	banana tree	root
Pinus contorta plastrovirus 1	3467	12.4	no	+1	2	/	SRR073389/90/91	pine tree	root, stem, leaf
Pinus taeda plastrovirus 1	4280	29.4	no	−1	118	TTTTGGGG, AGAAAAA, GGGGGAAA	SRR1276184/88/90/91	pine tree	root
Pinus taeda plastrovirus 2	6016	42.3	no	−1	88	GAAAAGG, CAAAAAAC	SRR1276188/90/91	pine tree	root
Pinus taeda plastrovirus 3	5050	56.5	no	−1	49	TTTAAAA	SRR1276201/2/4	pine tree	root

[a] virus name composed of scientific name of putative host species from which the sample was taken + plastrovirus + number of.

[b] in nucleotides.

[c] average number of sequencing reads mapping to a contig position.

[d] shown is the frameshift (FS) of ORF1b relative to ORF1a.

we could narrow down the putative junction site to a region of about 100 nucleotides (contig position 1957 to 2058) in the PCPLAV-1 genome. We then assessed for this region the quality of the assembly, which was built from sequencing reads produced with 454 technology that had an average length of 543 nucleotides. Although the average coverage was only four to five reads for this region, these reads showed 100% sequence identity, making an assembly artifact as cause for the fusion unlikely. PCPLAV-1 and other plastroviruses may thus form a valuable system for gaining new insights into major changes of genome architecture during virus evolution (see also Section 3.3 below).

Besides genomic organization, we additionally identified many similarities between astro- and plastrovirus proteomes. We could confidently assign RdRp domains to the ORF1b region of all 11 plastrovirus genomes using HHpred (probability of 100% for each PLAV, Fig. 1). We then constructed a multiple RdRp sequence alignment which included representative sequences from astroviruses as well as related poty- and hypoviruses (Fig. 2A). This alignment was subsequently used to reconstruct a Bayesian phylogeny for the four groups of viruses (Fig. 1, **Supplementary data file 2**). The fungi-infecting hypoviruses formed a well-supported outgroup to the plastro-, astro- and potyviruses in the RdRp phylogeny. The plastroviruses constituted at least three highly divergent lineages. Eight of them formed a monophyletic lineage with high support that we believe to prototype one or two new viral families. MAPLAV-2 clustered with the astroviruses that together were positioned at the root of the astro-/plastro-/potyvirus group. The astro-like viruses recently reported by Shi and colleagues (Shi et al., 2016, 2018) branched off next. Most notably, the potyviruses formed a well-supported monophyletic lineage together with PTPLAV-1 and MAPLAV-3 (Fig. 1) that grouped next to the large plastrovirus cluster composed of the remaining eight novel viruses. Most of the major branching points in the phylogeny were supported by an additional tree reconstructed using the maximum likelihood method, although mostly with only moderate to low support (Fig. 1). These relative uncertainties can be attributed to the high divergence of the plastroviruses with only one genome available per viral lineage. Notably, basal astroviruses (bastroviruses) that have recently been discovered from faecal samples of humans, pigs, bats and rats are most closely related to hepeviruses and the bastrovirus RdRp in particular shows only remote sequence similarity to the RdRps of plastro-, astro-, poty- and hypoviruses (Fig. S1) (Oude Munnink et al., 2016). The bastrovirus CP, on the other hand, is closely related to the CP of astroviruses but shows no significant sequence similarity to plastrovirus CPs (see below). In sum, the phylogenetic relationship between bastro- and plastroviruses is only remote, and hence does not provide evidence supporting the hypothesis that bastroviruses represent contaminants in stool originating from food plants (Oude Munnink et al., 2016).

An affinity of astro- and potyvirus polymerases has been noted before (Jiang et al., 1993), but the previous virus sampling limited to these two viral lineages with no close relatives available by that time did not allow for drawing major conclusions about a possible common evolutionary history of these two major ssRNA + virus families. Our findings brought a first indication that this ancestral virus might have been more similar to extant astroviruses and that the monocistronic potyvirus genome encoding a single large polyprotein may have evolved from a genome with multiple, potentially overlapping ORFs.

In line with this reasoning, we argued that the products encoded by ORF2 sequences (p2) of the 11 plastroviruses constitute capsid proteins (CPs), as do their counterparts in astroviruses. Although we were unable to identify significant sequence similarities to astro- or potyvirus CPs in a protein BLAST search, the p2 sequences of plastroviruses resembled those of astroviruses in terms of size, while the potyvirus capsid sequences were approximately half as long (Fig. 1). In a subsequent HHpred analysis of plastrovirus p2, we obtained hits against CPs from different reference viruses with moderate to high support for seven of the eleven p2 sequences (probabilities 66.8–100%). For four plastroviruses (CMPLAV-1, MAPLAV-3, PCPLAV-1, PTPLAV-3) the most

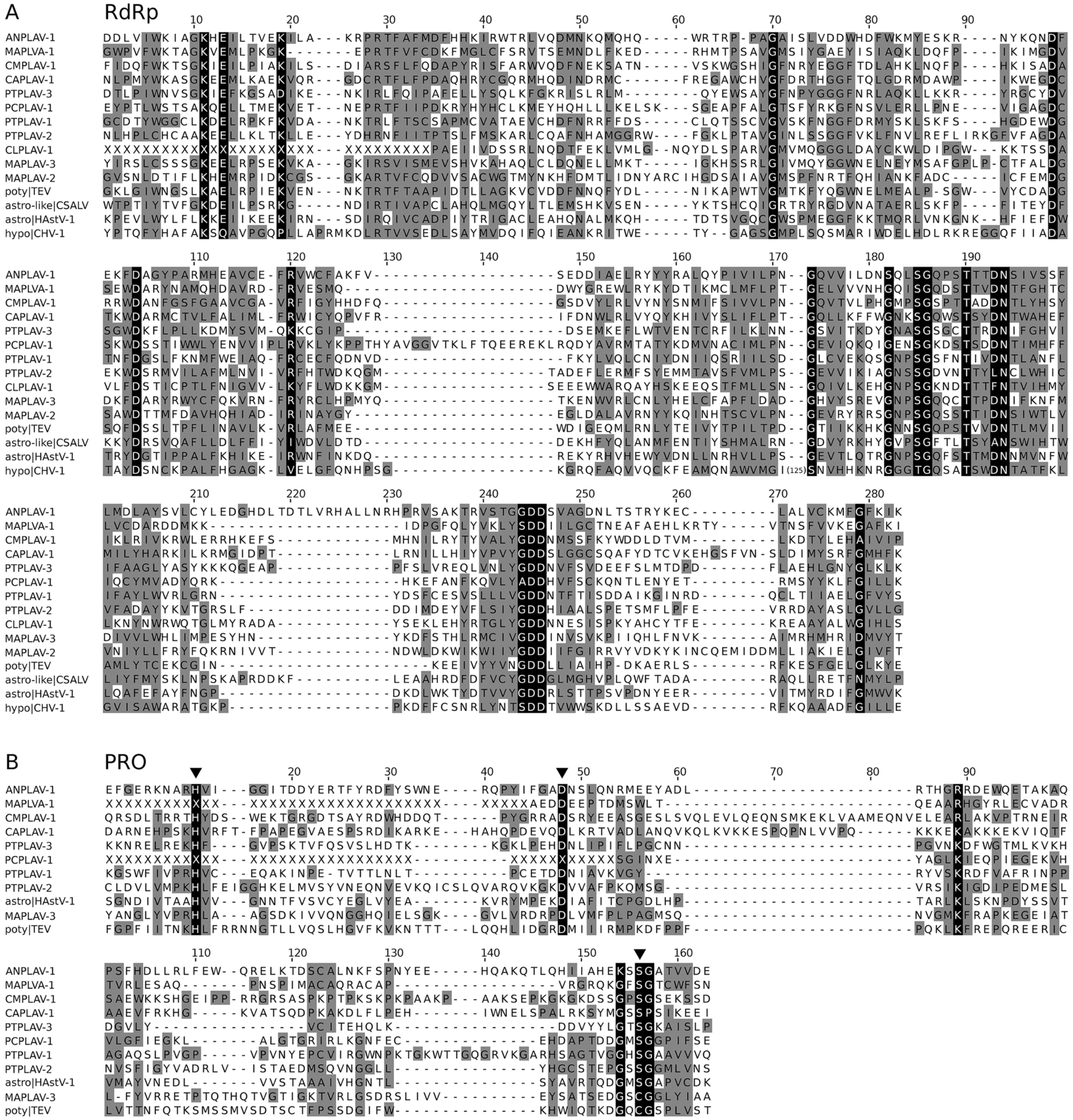

Fig. 2. Sequence alignments of RdRp and PRO_{Ser} core regions. Residues are colored according to amino acid properties and site conservation. Missing sequence is shown as X. (A) The RdRp region shown in the alignment corresponds to positions 1094–1318 in pp1ab of HAstV-1 and positions 2441–2656 in the polyprotein of TEV. Genbank accessions of four reference viruses are NC_001555 (poty/TEV), MG599877 (astro-like/CSALV), NC_030922 (astro/HAstV-1) and NC_041091 (hypo/CHV-1). An insertion of 125 amino acids in CHV-1 after alignment position 171 was omitted. (B) The protease region shown in the alignment corresponds to positions 452–558 in pp1a of HAstV-1 and positions 2074–2195 in the polyprotein of TEV. The astro- and potyvirus references are the same as in A. Arrow heads indicate the three catalytic sites of the protease.

similar reference capsids were from tombusviruses, while ANPLAV-1, MAPLAV-2 and PTPLAV-2 putative CPs, respectively, had an unclassified RNA virus (Orsay virus, PDB accession 4NWV_B), a sobemovirus and a nodavirus as top hit. We could verify theses assignments and additionally predict the expression of CPs by ORF2 for the remaining four plastroviruses (CAPLAV-1, CLPLAV-1, MAPLAV-1, PTPLAV-1) through an all-against-all BLAST search of p2 protein sequences in which each of the 11 plastroviruses obtained hits against at least two other viruses (Fig. 3). Eight of them showed one or several strong hits with an E-value of 1 × 10^{-9} or better, while the putative CPs of CLPLAV-1, MAPLAV-2 and PCPLAV-1, on the other hand, received only moderate to low support in the Blast analysis, with E-values in the range of 0.01 to 0.003.

With the aim to identify additional structural proteins from known viruses with remote similarities to the putative plastrovirus CPs, we used PSI-BLAST with each plastrovirus p2 sequence as query to iteratively retrieve such entries from RefSeq. This analysis yielded 1241 protein sequences that showed various degrees of pairwise sequence similarities between each other and to the plastrovirus p2 sequences (Fig. 4, **Supplementary data file 2**). Two large clusters - tombus-like

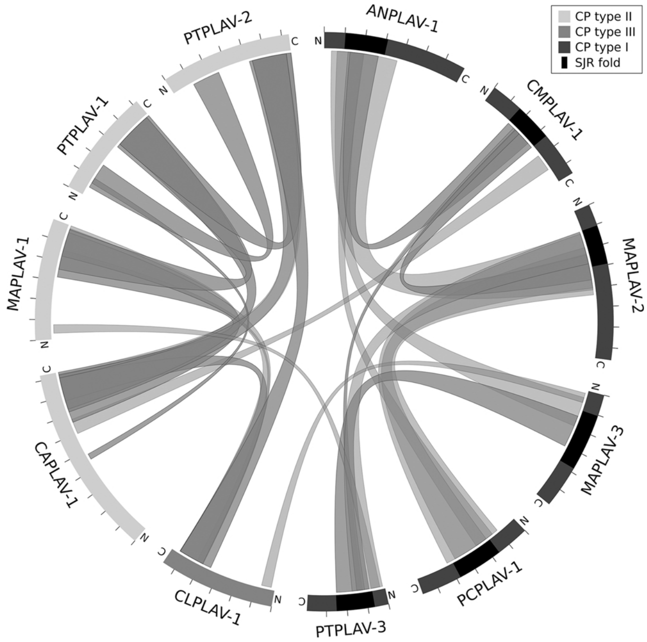

Fig. 3. Sequence similarities between plastrovirus CPs. The 11 plastrovirus CPs are shown as rectangles that are arranged in a circle; amino (N) and carboxy (C) sequence termini are indicated. Three CP types are discriminated using different shadings of gray. CP regions predicted by homology modelling to adopt the single jelly roll (SJR) fold are in black. Curved ribbons connect regions of the CPs that obtained significant hits in an all-against-all BLAST search of the plastrovirus p2 sequences with an E-value of 0.1 or below. Note that for the six CPs of type I most hits involved the predicted SJR domains, while the other plastrovirus CPs showed the strongest sequence conservation at their C-termini. The figure was generated using Circos (Krzywinski et al., 2009).

and picorna-like CPs that are known to adopt the single jelly roll (SJR) fold (Krupovic and Koonin, 2017) - as well as several isolated points or small groups without any connections to other CPs were evident from this similarity network. Importantly, six of the plastrovirus p2 sequences showed well-supported links with various tombus-like CPs that also included those from noda-, and recently reported cruciviruses (Diemer and Stedman, 2012; Quaiser et al., 2016). Also astro- and related hepevirus as well as luteovirus CPs are distantly related to this cluster (Fig. 4). These relations indicated that the p2 sequences of six plastroviruses (ANPLAV-1, CMPLAV-1, MAPLAV-2, MAPLAV-3, PCPLAV-1 and PTPLAV-3) contain CP domains that adopt the SJR fold, which we could verify through a homology modelling approach that utilizes information from viral CPs with resolved ultrastructures.

In contrast, the BLAST-based network analysis could not detect any sequence similarities to known SJR CPs for the remaining five plastrovirus p2 sequences, and we were also unable to determine a SJR fold for them using homology modelling. Four of them (CAPLAV-1, MAPLAV-1, PTPLAV-1 and PTPLAV-2) showed significant sequence similarities between each other while the fifth (CLPLAV-1) constituted a single point with no connections to other CPs in the similarity network (Fig. 4).

These results indicated that plastroviruses employ at least three highly divergent types of capsids from which at least one type adopts the SJR fold (Fig. 3). Notably, the distribution of these capsid types among the plastroviruses did not reflect the phylogenetic relationships of their RdRp sequences (Fig. 1), suggesting different trajectories of replication and structural modules and a relatively frequent replacement of capsids during the course of plastrovirus evolution, similar to what has been observed in many other viruses (Koonin et al., 2015; Krupovic and Koonin, 2017; Roux et al., 2013; Stedman, 2015; Welch et al., 2018).

The genome sequence of MAPLAV-3 showed an additional putative ORF3 immediately downstream of ORF2 for which we were unable to

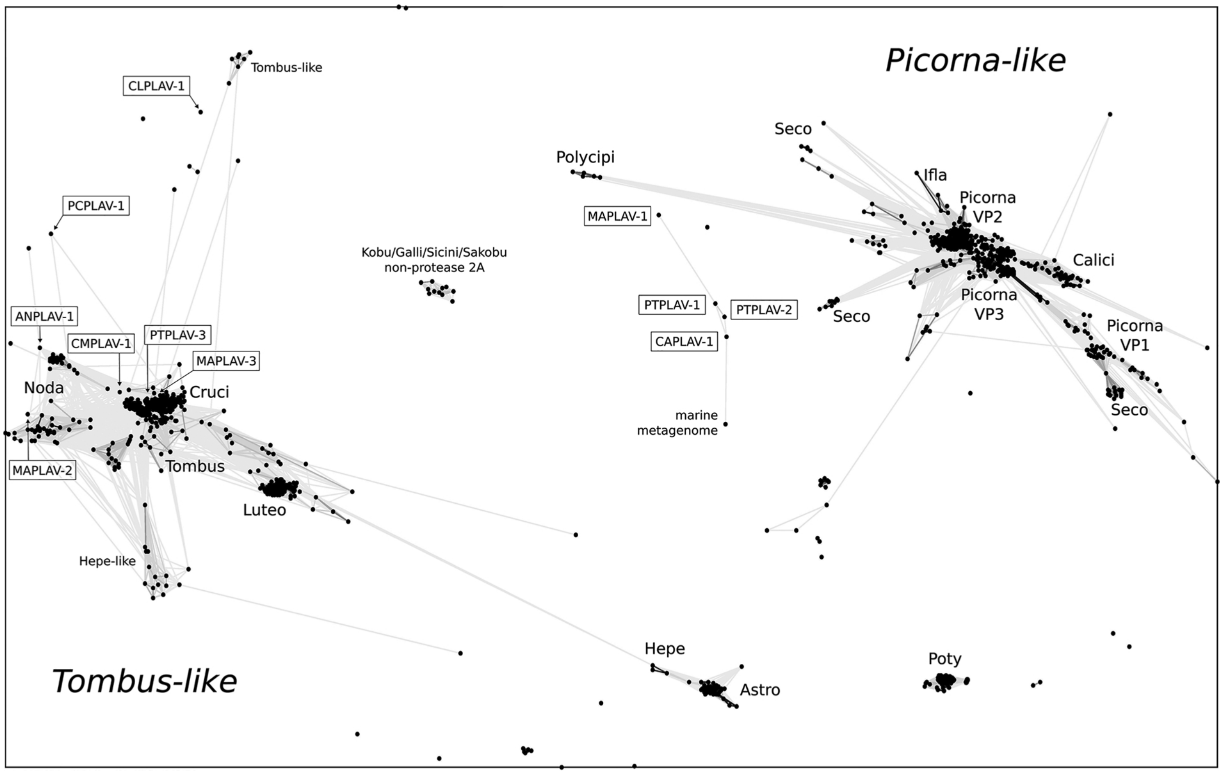

Fig. 4. Sequence similarity network of viral CPs. The viral CPs are shown as black dots and include those from the 11 PLAVs (names in open rectangles), 1241 sequences from RefSeq that showed significant sequence similarity with at least one PLAV CP in a PSI-BLAST search, and 96 potyvirus CPs. Gray edges indicate significant pairwise sequence similarities with a p-value of 0.001 or below. Major viral groups are indicated by labels.

detect any similarities in the databases with probability > 50% using HHpred (Fig. 1). To test whether ORF3 encodes a viral movement protein of the 30 K superfamily seen in many different plant virus families (Melcher, 2000) we performed a secondary structure prediction for the p3 sequence of MAPLAV-3. In contrast to members of the 30 K superfamily that are characterized by a core structure composed of a series of beta sheets with a nearly invariant aspartic acid at its center (Melcher, 2000; Mushegian and Elena, 2015), p3 was found to be predominantly composed of alpha helices. This showed that p3 is not a 30 K-like movement protein but likely has a different function that is yet to be determined. We note, however, that this does not exclude that plastroviruses express other movement proteins which do not belong to the 30 K superfamily. In the case of potyviruses, for instance, several proteins have been associated with viral movement, including the viral capsid (Deng et al., 2015; Dolja et al., 1994; Vijayapalani et al., 2012).

Although the ORF1a regions were incomplete for nine and missing for one of the assemblies, we could confidently assign the expression of a proteinase domain in this genomic region. Six of the plastroviruses (MAPLAV-1, MAPLAV-3, PCPLAV-1, PTPLAV-1, PTPLAV-2, PTPLAV-3) received strong hits against the Serine-like Cysteine (NIa) protease (PRO_{Ser}) of Tobacco vein mottling virus from the *Potyviridae* family in an HHpred search (probabilities 84.2–99.9%). Through expert-guided sequence alignment of the most conserved region spanning the catalytic triad (His-Asp-Cys/Ser) that involved astro- and potyvirus references we also detected highly divergent PRO_{Ser}-like core regions for ANPLAV-1, CMPLAV-1 and CAPLAV-1 (Fig. 2B). Only for CLPLAV-1 (missing ORF1a) and MAPLAV-2 (short 112 nt partial ORF1a presumably not covering the catalytic triad) we were unable to detect a protease domain. The plastrovirus PRO_{Ser} sequences were approximately equidistant to the astro- and potyvirus counterparts and eight out of nine had a Ser residue at the third catalytic position, as observed for astroviruses (Jiang et al., 1993). In sharp contrast, the MAPLAV-3 sequence showed a substitution of the catalytic Ser to a Cys residue, thus resembling potyviruses (Hwang et al., 2000) and indicating a shift from a canonical Serine to a Serine-like Cysteine protease in the MAPLAV-3 lineage. Because a change of Ser to Cys could be obtained by a single point mutation at the first codon position, we sought to rule out an assembly artifact as the cause for the observed substitution. Indeed, the Cys codon in MAPLAV-3 (TGT, contig position 2244–2246) was supported by > 181 sequencing reads, showing that this substitution was genuine.

Besides the catalytic Ser residue of PRO_{Ser}, the ORF1a sequences of the two plastroviruses with complete coding regions (ANPLAV-1 and MAPLAV-3) resembled the astroviruses by encoding, respectively, five and four transmembrane domains upstream of the protease domain (TMHMM probabilities 60–100%). In addition, the other two plastroviruses for which a major portion of ORF1a was covered by the assembly (PTPLAV-2 and PTPLAV-3) each also showed a transmembrane domain upstream of PRO_{Ser} (probability 100%). In contrast, the location of the PRO_{Ser} domain in ORF1a differed from that in astroviruses by being encoded in the 3′-terminal portion of ORF1a for eight of the nine plastroviruses for which we detected its catalytic core region (Fig. 1). Only ANPLAV-1 may express an additional protein domain of about 240–340 amino acids downstream of PRO^{Ser}, but we were unable to define a respective domain border with confidence. The most unusual putative protease domain was again that of CMPLAV-1. It ended just 8 residues after the predicted catalytic Ser, while the C-terminus of PRO_{Ser} was considerably longer in the other plastroviruses and the astro- and potyviruses. Notably, we identified weak sequence similarity between the N-terminus of pp1b of CMPLAV-1 and the PRO_{Ser} C-terminus of other plastroviruses, indicating that production of a full-length protease might require ORF1a/b frameshifting. The encoding of ORF1b in the +1 instead of -1 reading frame relative to ORF1a (see above) already indicated that the pp1ab expression mechanism may be very unusual for CMPLAV-1.

For five plastroviruses (ANPLAV-1, CMPLAV-1, PCPLAV-1, PTPLAV-1 and PTPLAV-3) we obtained hits against astro- or potyvirus VPg sequences in a Blast search involving the plastrovirus pp1a region upstream of PRO_{Ser}. We note, however, that these predictions should be treated with caution, due to weak E-values in the range of 0.085 to 0.8, a biased amino acid composition of astro- and potyvirus VPg sequences that increases the risk for false-positive hits, and because we could not verify the inferences in a profile search using an astrovirus and a potyvirus VPg alignment as query. Notwithstanding that, the predicted positions of putative plastrovirus VPgs upstream of PRO_{Ser} would correspond to that of VPg in the potyvirus polyprotein (Shahabuddin et al., 1988).

From the analysis of plastrovirus genomes and proteomes described above, it became apparent that the highly divergent plastrovirus lineages may have reached different intermediate positions on an evolutionary trajectory from astro-like to poty-like viruses. If true, this argues for an emergence of potyviruses from an ancestral virus with (pl) astro-like genomic features.

3.3. A model about the emergence of potyviruses from viruses with (pl) astro-like genomes

Based on our results we propose a model in which potyviruses emerged from an ancestral virus with astro-like genomic features

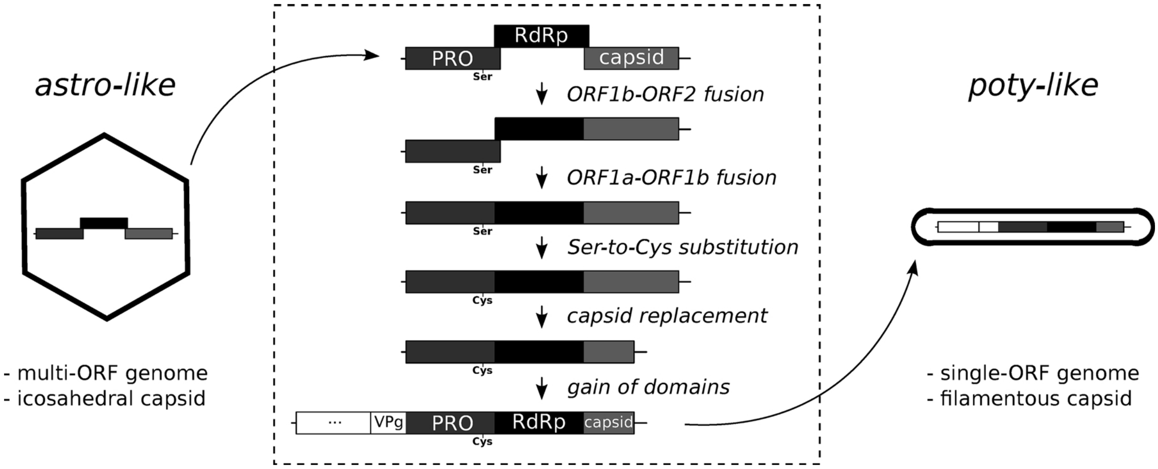

Fig. 5. Theoretical model of potyvirus emergence from astro-like viruses. The hypothesized evolutionary transition from a virus with astro-like features (left) to a poty-like virus (right) involves multiple steps (middle) that do not necessarily need to occur in the presented order. Some of the discovered plastrovirus genomes resemble genomes at intermediate steps of the model.

(Fig. 5). Although verification or disqualification of this model will require further experiments including the discovery of additional viruses intermediate to the astro-, plastro- and potyviruses as well as functional assays of proteins expressed by different of these intermediates, the proposed model stimulates future research in these directions. Our model is based on the following key observations and conclusions:

First, the high similarity of RdRp sequences of astro-, plastro- and potyviruses shows that these viral groups are relatively closely related to each other compared to other RNA viruses. We explicitly do not consider a scenario in which the ancestral potyvirus acquired its polymerase from an astro-like virus (or *vice versa*) as we are not aware of any example of horizontal gene transfer involving a RdRp between members from different virus families.

Second, potyviruses are positioned within a major cluster of plastroviruses in the RdRp phylogeny (Fig. 1).

Third, we discovered the plastrovirus genomes in plant transcriptome projects, indicating that they likely infect plants or microorganisms tightly associated with plants, which would form the basis for an evolutionary transition to poty-like viruses.

Fourth, an array of three major proteins (PRO_{Ser}-RdRp-CP) appears in the same order in the genomes of all three groups of viruses. Importantly, PRO_{Ser} and RdRp are separated by additional domains in astroviruses but not in plastroviruses for which PRO_{Ser} is encoded in the very 3′ terminus of ORF1a, thus positioning PRO_{Ser} and RdRp adjacent to each other in the expressed 1ab polyprotein and resembling the situation in the potyvirus polyprotein (Fig. 1).

Fifth, two plastroviruses can be considered as intermediates on a possible transition from astro-like to poty-like genomes. The fusion of ORF1b with the subsequent ORF2 in PCPLAV-1 could resemble an early step towards expression of a single large polyprotein, while the major change from Ser to Cys at the third catalytic site of PRO_{Ser} in MAPLAV-3, which is phylogenetically close to the potyviruses, is also seen in the potyvirus PRO_{Ser} (Fig. 2B).

Sixth, the deviation of PCPLAV-1 from other plastroviruses by fusion of ORF1b and ORF2 resembles the differences in genomic organization between genera of the ssRNA + virus family *Caliciviridae*. While members of two calicivirus genera (*Vesivirus* and *Norovirus*) encode their non-structural proteins and the major CP in two different, adjacent ORFs (ORF1 and ORF2, respectively), viruses from the three other genera (*Lagovirus*, *Sapovirus* and *Nebovirus*) use a single ORF for the expression of these proteins (resembling a putative fusion of ORF1 and ORF2) (Clarke and Lambden, 2000; Liu et al., 1995). Notably, caliciviruses share with (pl)astroviruses the expression of the structural genes via subgenomic RNAs and their genomes encode the characteristic array of VPg-PRO-RdRp-CP domains in the same order seen in poty- and plastroviruses. Moreover, caliciviruses are relatively closely related to picornaviruses and other families of the order *Picornavirales* some of which encode their proteins using a single large ORF, resembling potyviruses in this respect (Le Gall et al., 2008; Zell et al., 2017). The evolution of these viral groups from a common ancestral genome may thus have followed a trajectory that is very similar to that proposed here for (pl)astro- and potyviruses.

And seventh, the relative orientation of ORF1a and ORF1b reading frames is altered in two of the plastroviruses, indicating a change from -1 to +1 translational frameshifting or, in the case of CMPLAV-1, possibly even to transcriptional frameshifting. In the case of transcriptional slippage, which is assumed to add nucleotides to the transcript (Olspert et al., 2015), it is tempting to speculate that such an elongated transcript was accidentally encapsidated and has replaced the original genome, essentially giving rise to a viral variant whose genome encodes former ORF1a and ORF1b in the same reading frame.

According to our model (Fig. 5), a transition from an astro-like to a poty-like genome requires at least five steps that do not necessarily need to occur in the order presented below. Perhaps early on, a fusion of ORF1b and ORF2 happened and was accompanied by a loss of subgenomic RNA production, a hallmark of astrovirus ORF2 expression (Monroe et al., 1993). This putative intermediate virus is resembled by PCPLAV-1. Next, a similar fusion process at or close to the ORF1a/b overlap resulted in the expression of PRO_{Ser} and RdRp domains from the new elongated ORF. The resulting intermediate virus, for which no extant virus is present in our data set, lost the ability to regulate the relative amounts of mature proteins encoded in ORF1a and ORF1b through ribosomal frameshifting. This may have triggered a switch from a catalytic Ser to a Cys in the protease active site in a third step, with MAPLAV-3 resembling this putative intermediate virus. The changed catalytic residue likely affected cleavage specificity and might have enabled fine-tuning the relative amounts of mature proteins released from the polyprotein through regulated proteolytic processing (instead of frameshifting) whose rate is expected to vary depending on the sequence context at a cleavage site. It followed a replacement of the viral coat protein resulting in a switch from icosahedral to filamentous capsids with helical symmetry. Lastly, additional protein domains, including VPg, were inserted or emerged *de novo* upstream of the PRO_{Ser} domain to give rise to the extant potyviruses we see today.

We emphasize that the putative emergence of potyviruses from astro-like viruses would not be inconsistent with a proposed transition of an ancestral astrovirus to the progenitor of nidoviruses, the latter resembling astroviruses in terms of a multi-ORF organization of their genomes as well as RFS- and subgenomic RNA-mediated gene expression (Gorbalenya et al., 2006). In this respect, poty- and nidoviruses may represent the outcomes of two different and independent evolutionary trajectories, highlighting once more the extraordinarily high capacity of RNA viruses to produce genetic diversity.

4. Conclusions

Here we have shown that the SRA database is a rich source of unknown viral sequences that can be screened in high-throughput and with high sensitivity. Consistent with our earlier study (Lauber et al., 2017) we have demonstrated that our approach to virus discovery is inexpensive as it does not rely on the acquisition of samples and subsequent sequencing, and it is more comprehensive than any other current approach due to the millions of data sets from a large variety of potential host species available from the SRA. By targeting a single virus family, the *Astroviridae*, we discovered nearly a dozen highly divergent viral genomes as well as short sequence fragments from an even larger number of viruses in a minor fraction of available transcriptomes of diverse plants, indicating that many more related viruses are yet to be discovered. Two of the viral genomes that we assembled from SRA data sets had full-length coding regions while the remaining sequences were partial as we were unable to extend the respective assemblies due to insufficient sequence coverage. By using a combinatorial approach these discovered sequences provided novel insights into the evolution of two major and very different virus families, the astroviruses and potyviruses, and allowed us to reveal so far unappreciated evolutionary connections between them. This gain of knowledge would not have been possible without considering the partial genomes. Future extension of the search to additional plant taxa, including economically important species, will show whether the discovered viruses, which we tentatively named plastroviruses, are a burden for society.

Acknowledgements

We thank the Center for Information Services and High Performance Computing (ZIH) at TU Dresden for the generous allocation of computer time and excellent technical support. R.B. is supported by the Deutsche Forschungsgemeinschaft, TRR179, TP9.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.virusres.2018.11.009.

References

Abascal, F., Zardoya, R., Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. Bioinforma. Oxf. Engl. 21, 2104–2105. https://doi.org/10.1093/bioinformatics/bti263.

Ali, M., Hameed, S., Tahir, M., 2014. Luteovirus: insights into pathogenicity. Arch. Virol. 159, 2853–2860. https://doi.org/10.1007/s00705-014-2172-6.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Alves, J.M.P., de Oliveira, A.L., Sandberg, T.O.M., Moreno-Gallego, J.L., de Toledo, M.A.F., de Moura, E.M.M., Oliveira, L.S., Durham, A.M., Mehnert, D.U., Zanotto, P.M., de, A., Reyes, A., Gruber, A., 2016. GenSeed-HMM: a tool for progressive assembly using profile HMMs as seeds and its application in alpavirinae viral discovery from metagenomic data. Front. Microbiol. 7, 269. https://doi.org/10.3389/fmicb.2016.00269.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. J. Comput. Mol. Cell Biol. 19, 455–477. https://doi.org/10.1089/cmb.2012.0021.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2013. GenBank. Nucleic Acids Res. 41, D36–D42. https://doi.org/10.1093/nar/gks1195.

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M., 1977. The Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol. 112, 535–542.

Brierley, I., Vidakovic, M., 2003. V, 2.Ribosomal frameshifting in astroviruses. Perspectives in Medical Virology. Elsevier, pp. 587–606. https://doi.org/10.1016/S0168-7069(03)09035-9.

Brierley, I., Boursnell, M.E., Binns, M.M., Bilimoria, B., Blok, V.C., Brown, T.D., Inglis, S.C., 1987. An efficient ribosomal frame-shifting signal in the polymerase-encoding region of the coronavirus IBV. EMBO J. 6, 3779–3785.

Brister, J.R., Ako-Adjei, D., Bao, Y., Blinkova, O., 2015. NCBI viral genomes resource. Nucleic Acids Res. 43, D571–577. https://doi.org/10.1093/nar/gku1207.

Carrington, J.C., Cary, S.M., Parks, T.D., Dougherty, W.G., 1989. A second proteinase encoded by a plant potyvirus genome. EMBO J. 8, 365–370.

Cattaneo, R., Kaelin, K., Baczko, K., Billeter, M.A., 1989. Measles virus editing provides an additional cysteine-rich protein. Cell 56, 759–764.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., Thompson, J.D., 2003. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. 31, 3497–3500.

Clarke, I.N., Lambden, P.R., 2000. Organization and expression of calicivirus genes. J. Infect. Dis. 181 (Suppl. 2), S309–316. https://doi.org/10.1086/315575.

Deng, P., Wu, Z., Wang, A., 2015. The multifunctional protein CI of potyviruses plays interlinked and distinct roles in viral genome replication and intercellular movement. Virol. J. 12, 141. https://doi.org/10.1186/s12985-015-0369-2.

Diemer, G.S., Stedman, K.M., 2012. A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. Biol. Direct 7 (13). https://doi.org/10.1186/1745-6150-7-13.

Dolja, V.V., Koonin, E.V., 2018. Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer. Virus Res. 244, 36–52. https://doi.org/10.1016/j.virusres.2017.10.020.

Dolja, V.V., Haldeman, R., Robertson, N.L., Dougherty, W.G., Carrington, J.C., 1994. Distinct functions of capsid protein in assembly and movement of tobacco etch potyvirus in plants. EMBO J. 13, 1482–1491.

Drexler, J.F., Corman, V.M., Müller, M.A., Maganga, G.D., Vallo, P., Binger, T., Gloza-Rausch, F., Cottontail, V.M., Rasche, A., Yordanov, S., Seebens, A., Knörnschild, M., Oppong, S., Sarkodie, Y.A., Pongombo, C., Lukashev, A.N., Schmidt-Chanasit, J., Stöcker, A., Carneiro, A.J.B., Erbar, S., Maisner, A., Fronhoffs, F., Buettner, R., Kalko, E.K.V., Kruppa, T., Franke, C.R., Kallies, R., Yandoko, E.R.N., Herrler, G., Reusken, C., Hassanin, A., Krüger, D.H., Matthee, S., Ulrich, R.G., Leroy, E.M., Drosten, C., 2012. Bats host major mammalian paramyxoviruses. Nat. Commun. 3. https://doi.org/10.1038/ncomms1796.

Drozdetskiy, A., Cole, C., Procter, J., Barton, G.J., 2015. JPred4: a protein secondary structure prediction server. Nucleic Acids Res. 43, W389–W394. https://doi.org/10.1093/nar/gkv332.

Drummond, A.J., Ho, S.Y.W., Phillips, M.J., Rambaut, A., 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4, e88. https://doi.org/10.1371/journal.pbio.0040088.

Drummond, A.J., Suchard, M.A., Xie, D., Rambaut, A., 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29, 1969–1973. https://doi.org/10.1093/molbev/mss075.

Eddy, S.R., 2011. Accelerated profile HMM searches. PLoS Comput. Biol. 7, e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797. https://doi.org/10.1093/nar/gkh340.

Frickey, T., Lupas, A., 2004. CLANS: a Java application for visualizing protein families based on pairwise similarity. Bioinforma. Oxf. Engl. 20, 3702–3704. https://doi.org/10.1093/bioinformatics/bth444.

Gorbalenya, A.E., Donchenko, A.P., Blinov, V.M., Koonin, E.V., 1989. Cysteine proteases of positive strand RNA viruses and chymotrypsin-like serine proteases. A distinct protein superfamily with a common structural fold. FEBS Lett. 243, 103–114.

Gorbalenya, A.E., Enjuanes, L., Ziebuhr, J., Snijder, E.J., 2006. Nidovirales: evolving the largest RNA virus genome. Virus Res. 117, 17–37. https://doi.org/10.1016/j.virusres.2006.01.017.

Gouy, M., Guindon, S., Gascuel, O., 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. 27, 221–224. https://doi.org/10.1093/molbev/msp259.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652. https://doi.org/10.1038/nbt.1883.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59, 307–321. https://doi.org/10.1093/sysbio/syq010.

Hellmann, G.M., Shaw, J.G., Rhoads, R.E., 1988. In vitro analysis of tobacco vein mottling virus NIa cistron: evidence for a virus-encoded protease. Virology 163, 554–562.

Hwang, D.C., Kim, D.H., Lee, J.S., Kang, B.H., Han, J., Kim, W., Song, B.D., Choi, K.Y., 2000. Characterization of active-site residues of the NIa protease from tobacco vein mottling virus. Mol. Cells 10, 505–511.

Jiang, B., Monroe, S.S., Koonin, E.V., Stine, S.E., Glass, R.I., 1993. RNA sequence of astrovirus: distinctive genomic organization and a putative retrovirus-like ribosomal frameshifting signal that directs the viral replicase synthesis. Proc. Natl. Acad. Sci. U. S. A. 90, 10539–10543.

Koonin, E.V., 1991. Genome replication/expression strategies of positive-strand RNA viruses: a simple version of a combinatorial classification and prediction of new strategies. Virus Genes 5, 273–281.

Koonin, E.V., Dolja, V.V., Krupovic, M., 2015. Origins and evolution of viruses of eukaryotes: the ultimate modularity. Virology 479–480, 2–25. https://doi.org/10.1016/j.virol.2015.02.039.

Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L., 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. 305, 567–580. https://doi.org/10.1006/jmbi.2000.4315.

Krupovic, M., Koonin, E.V., 2017. Multiple origins of viral capsid proteins from cellular ancestors. Proc. Natl. Acad. Sci. 114, E2401–E2410. https://doi.org/10.1073/pnas.1621061114.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., Marra, M.A., 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19, 1639–1645. https://doi.org/10.1101/gr.092759.109.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923.

Lauber, C., Seitz, S., Mattei, S., Suh, A., Beck, J., Herstein, J., Börold, J., Salzburger, W., Kaderali, L., Briggs, J.A.G., Bartenschlager, R., 2017. Deciphering the origin and evolution of hepatitis B viruses by means of a family of non-enveloped fish viruses. Cell Host Microbe 22, 387–399. https://doi.org/10.1016/j.chom.2017.07.019. e6.

Le Gall, O., Christian, P., Fauquet, C.M., King, A.M.Q., Knowles, N.J., Nakashima, N., Stanway, G., Gorbalenya, A.E., 2008. Picornavirales, a proposed order of positive-sense single-stranded RNA viruses with a pseudo-T = 3 virion architecture. Arch. Virol. 153, 715–727. https://doi.org/10.1007/s00705-008-0041-x.

Leinonen, R., Sugawara, H., Shumway, M., International Nucleotide Sequence Database Collaboration, 2011. The sequence read archive. Nucleic Acids Res. 39, D19–21. https://doi.org/10.1093/nar/gkq1019.

Liu, B.L., Clarke, I.N., Caul, E.O., Lambden, P.R., 1995. Human enteric caliciviruses have a unique genome structure and are distinct from the Norwalk-like viruses. Arch. Virol. 140, 1345–1356.

Madeley, C.R., Cosgrove, B.P., 1975. Letter: viruses in infantile gastroenteritis. Lancet Lond. Engl. 2, 124.

Maia, I.G., Séron, K., Haenni, A.L., Bernardi, F., 1996. Gene expression from viral RNA genomes. Plant Mol. Biol. 32, 367–391.

Mäkinen, K., Tamm, T., Naess, V., Truve, E., Puurand, U., Munthe, T., Saarma, M., 1995. Characterization of cocksfoot mottle sobemovirus genomic RNA and sequence comparison with related viruses. J. Gen. Virol. 76 (Pt 11), 2817–2825. https://doi.org/10.1099/0022-1317-76-11-2817.

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C.J., Lu, S., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Lu, F., Marchler, G.H., Song, J.S., Thanki, N., Wang, Z., Yamashita, R.A., Zhang, D., Zheng, C., Geer, L.Y., Bryant, S.H., 2017. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. 45, D200–D203. https://doi.org/10.1093/nar/gkw1129.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. Embnet J. 17, 10. https://doi.org/10.14806/ej.17.1.200.

Melcher, U., 2000. The "30K" superfamily of viral movement proteins. J. Gen. Virol. 81, 257–266. https://doi.org/10.1099/0022-1317-81-1-257.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., Marshall, D., 2010. Tablet–next generation sequence assembly visualization. Bioinforma. Oxf. Engl. 26, 401–402. https://doi.org/10.1093/bioinformatics/btp666.

Monroe, S.S., Jiang, B., Stine, S.E., Koopmans, M., Glass, R.I., 1993. Subgenomic RNA sequence of human astrovirus supports classification of Astroviridae as a new family of RNA viruses. J. Virol. 67, 3611–3614.

Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536–540. https://doi.org/10.1006/jmbi.1995.0159.

Mushegian, A.R., Elena, S.F., 2015. Evolution of plant virus movement proteins from the 30K superfamily and of their homologs integrated in plant genomes. Virology 476, 304–315. https://doi.org/10.1016/j.virol.2014.12.012.

Olspert, A., Chung, B.Y.-W., Atkins, J.F., Carr, J.P., Firth, A.E., 2015. Transcriptional slippage in the positive-sense RNA virus family Potyviridae. EMBO Rep. 16, 995–1004. https://doi.org/10.15252/embr.201540509.

Oude Munnink, B.B., Cotten, M., Canuti, M., Deijs, M., Jebbink, M.F., van Hemert, F.J., Phan, M.V.T., Bakker, M., Jazaeri Farsani, S.M., Kellam, P., van der Hoek, L., 2016. A Novel Astrovirus-Like RNA Virus Detected in Human Stool. Virus Evol. 2https://doi.org/10.1093/ve/vew005. vew005.

Phan, T.G., Kapusinszky, B., Wang, C., Rose, R.K., Lipton, H.L., Delwart, E.L., 2011. The fecal viral flora of wild rodents. PLoS Pathog. 7, e1002218. https://doi.org/10.1371/journal.ppat.1002218.

Pruitt, K.D., Tatusova, T., Maglott, D.R., 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35, D61–D65. https://doi.org/10.1093/nar/gkl842.

Quaiser, A., Krupovic, M., Dufresne, A., Francez, A.-J., Roux, S., 2016. Diversity and comparative genomics of chimeric viruses in Sphagnum-dominated peatlands. Virus Evol. 2https://doi.org/10.1093/ve/vew025. vew025.

Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: the european molecular biology open software suite. Trends Genet. TIG 16, 276–277.

Rodamilans, B., Shan, H., Pasin, F., García, J.A., 2018. Plant viral proteases: beyond the role of peptide cutters. Front. Plant Sci. 9, 666. https://doi.org/10.3389/fpls.2018.00666.

Rosario, K., Breitbart, M., 2011. Exploring the viral world through metagenomics. Curr. Opin. Virol. 1, 289–297. https://doi.org/10.1016/j.coviro.2011.06.004.

Roux, S., Enault, F., Bronner, G., Vaulot, D., Forterre, P., Krupovic, M., 2013. Chimeric viruses blur the borders between the major groups of eukaryotic single-stranded DNA viruses. Nat. Commun. 4, 2700. https://doi.org/10.1038/ncomms3700.

Rybicki, E.P., 2015. A Top ten list for economically important plant viruses. Arch. Virol. 160, 17–20. https://doi.org/10.1007/s00705-014-2295-9.

Sanchez, A., Trappier, S.G., Mahy, B.W., Peters, C.J., Nichol, S.T., 1996. The virion glycoproteins of Ebola viruses are encoded in two reading frames and are expressed through transcriptional editing. Proc. Natl. Acad. Sci. U. S. A. 93, 3602–3607.

Shahabuddin, M., Shaw, J.G., Rhoads, R.E., 1988. Mapping of the tobacco vein mottling virus VPg cistron. Virology 163, 635–637.

Shi, M., Lin, X.-D., Tian, J.-H., Chen, L.-J., Chen, X., Li, C.-X., Qin, X.-C., Li, J., Cao, J.-P., Eden, J.-S., Buchmann, J., Wang, W., Xu, J., Holmes, E.C., Zhang, Y.-Z., 2016. Redefining the invertebrate RNA virosphere. Nature. https://doi.org/10.1038/nature20167.

Shi, M., Lin, X.-D., Chen, X., Tian, J.-H., Chen, L.-J., Li, K., Wang, W., Eden, J.-S., Shen, J.-J., Liu, L., Holmes, E.C., Zhang, Y.-Z., 2018. The evolutionary history of vertebrate RNA viruses. Nature 556, 197–202. https://doi.org/10.1038/s41586-018-0012-7.

Stedman, K.M., 2015. Deep Recombination: RNA and ssDNA Virus Genes in DNA Virus and Host Genomes. Annu. Rev. Virol. 2, 203–217. https://doi.org/10.1146/annurev-virology-100114-055127.

Suttle, C.A., 2007. Marine viruses–major players in the global ecosystem. Nat. Rev. Microbiol. 5, 801–812. https://doi.org/10.1038/nrmicro1750.

Thomas, S.M., Lamb, R.A., Paterson, R.G., 1988. Two mRNAs that differ by two nontemplated nucleotides encode the amino coterminal proteins P and V of the paramyxovirus SV5. Cell 54, 891–902.

Tözsér, J., Tropea, J.E., Cherry, S., Bagossi, P., Copeland, T.D., Wlodawer, A., Waugh, D.S., 2005. Comparison of the substrate specificity of two potyvirus proteases. FEBS J. 272, 514–523. https://doi.org/10.1111/j.1742-4658.2004.04493.x.

van Marle, G., Luytjes, W., van der Most, R.G., van der Straaten, T., Spaan, W.J., 1995. Regulation of coronavirus mRNA transcription. J. Virol. 69, 7851–7856.

Verchot, J., Koonin, E.V., Carrington, J.C., 1991. The 35-kDa protein from the N-terminus of the potyviral polyprotein functions as a third virus-encoded proteinase. Virology 185, 527–535.

Vijayapalani, P., Maeshima, M., Nagasaki-Takekuchi, N., Miller, W.A., 2012. Interaction of the trans-frame potyvirus protein P3N-PIPO with host protein PCaP1 facilitates potyvirus movement. PLoS Pathog. 8, e1002639. https://doi.org/10.1371/journal.ppat.1002639.

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., Lepore, R., Schwede, T., 2018. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 46, W296–W303. https://doi.org/10.1093/nar/gky427.

Welch, N.L., Yutin, N., Dill, J.A., Camus, A.C., Pang, Y.-Y.S., Schiller, J.T., An, P., Cantalupo, P.G., Pipas, J.M., Delwart, E., Koda, S., Subramaniam, K., Waltzek, T.B., Bian, C., Shi, Q., Ruan, Z., Koonin, E.V., Buck, C.B., Ng, T.F.F., 2018. Adomaviruses: an Emerging Virus Family Provides Insights Into DNA Virus Evolution. https://doi.org/10.1101/341131.

Whelan, S.P.J., Barr, J.N., Wertz, G.W., 2004. Transcription and replication of nonsegmented negative-strand RNA viruses. Curr. Top. Microbiol. Immunol. 283, 61–119.

Willcocks, M.M., Brown, T.D., Madeley, C.R., Carter, M.J., 1994. The complete sequence of a human astrovirus. J. Gen. Virol. 75 (Pt 7), 1785–1788. https://doi.org/10.1099/0022-1317-75-7-1785.

Wylie, S.J., Adams, M., Chalam, C., Kreuze, J., López-Moya, J.J., Ohshima, K., Praveen, S., Rabenstein, F., Stenger, D., Wang, A., Zerbini, F.M., ICTV Report Consortium, 2017. ICTV virus taxonomy profile: potyviridae. J. Gen. Virol. 98, 352–354. https://doi.org/10.1099/jgv.0.000740.

Zell, R., Delwart, E., Gorbalenya, A.E., Hovi, T., King, A.M.Q., Knowles, N.J., Lindberg, A.M., Pallansch, M.A., Palmenberg, A.C., Reuter, G., Simmonds, P., Skern, T., Stanway, G., Yamashita, T., ICTV Report Consortium, 2017. ICTV virus taxonomy profile: picornaviridae. J. Gen. Virol. 98, 2421–2422. https://doi.org/10.1099/jgv.0.000911.

Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A.N., Alva, V., 2017. A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. J. Mol. Biol. https://doi.org/10.1016/j.jmb.2017.12.007.