viruses

MDPI

Article

An Ancient Lineage of Highly Divergent Parvoviruses Infects both Vertebrate and Invertebrate Hosts

Judit J. Pénzes [1,*], William Marciel de Souza [2], Mavis Agbandje-McKenna [1] and Robert J. Gifford [3,*]

1 McKnight Brain Institute and Department of Biochemistry and Molecular Biology, University of Florida, 1149 Newell Dr, Gainesville, FL 32610, USA; mckenna@ufl.edu
2 Virology Research Center, School of Medicine of Ribeirão Preto of the University of São Paulo, Ribeirão Preto, Brazil; wmarciel2@gmail.com
3 Medical Research Council-University of Glasgow Centre for Virus Research, 464 Bearsden Road, Glasgow G61 1QH, UK
* Correspondence: judit.penzes@ufl.edu (J.J.P.); robert.gifford@glasgow.ac.uk (R.J.G.)

Received: 17 April 2019; Accepted: 5 June 2019; Published: 6 June 2019

Abstract: Chapparvoviruses (ChPVs) comprise a divergent, recently identified group of parvoviruses (family *Parvoviridae*), associated with nephropathy in immunocompromised laboratory mice and with prevalence in deep sequencing results of livestock showing diarrhea. Here, we investigate the biological and evolutionary characteristics of ChPVs via comparative in silico analyses, incorporating sequences derived from endogenous parvoviral elements (EPVs) as well as exogenous parvoviruses. We show that ChPVs are an ancient lineage within the *Parvoviridae*, clustering separately from members of both currently established subfamilies. Consistent with this, they exhibit a number of characteristic features, including several putative auxiliary protein-encoding genes, and capsid proteins with no sequence-level homology to those of other parvoviruses. Homology modeling indicates the absence of a β-A strand, normally part of the luminal side of the parvoviral capsid protein core. Our findings demonstrate that the ChPV lineage infects an exceptionally broad range of host species, including both vertebrates and invertebrates. Furthermore, we observe that ChPVs found in fish are more closely related to those from invertebrates than they are to those of amniote vertebrates. This suggests that transmission between distantly related host species may have occurred in the past and that the *Parvoviridae* family can no longer be divided based on host affiliation.

Keywords: chapparvovirus; parvovirus evolution; endogenous viral elements; *Parvoviridae*; densovirus; homology modeling; new viruses

1. Introduction

Parvoviruses (family *Parvoviridae*) are small, non-enveloped viruses with T = 1 icosahedral symmetry and linear, single-stranded DNA (ssDNA) genomes ~4–6 kilobases (kb) in length. The family has historically been divided into two subfamilies, *Parvovirinae* and *Densovirinae*, containing viruses that infect vertebrate and invertebrate hosts, respectively [1]. Despite exhibiting great variation in expression and transcription strategies, they have a relatively conserved overall genome structure: a non-structural (NS) expression cassette is located at the left side of the genome, while the structural viral proteins (VPs) are encoded by the right, and complex, hairpin-like DNA secondary structures are present at both genomic termini [2]. Small satellite proteins and an assembly-activating protein have been discovered as products of open reading frames (ORFs) overlapping the right-hand expression cassette, whereas additional auxiliary protein-encoding ORFs may be positioned between the two major cassettes [3,4].

Numerous novel parvoviruses have been identified in recent years, primarily via approaches based on high throughput sequencing (HTS) [5–11]. In addition, progress in whole genome sequencing of eukaryotes has revealed that sequences derived from parvoviruses occur relatively frequently in animal genomes [12–17]. These endogenous parvoviral elements (EPVs) are derived from the genomes of ancient parvoviruses that were incorporated into the gene pool of ancestral host species. This can presumably occur when infection of a germline cell leads to parvovirus-derived DNA becoming integrated into host chromosomes, and the cell containing the integrated sequences then goes on to develop into a viable organism [18]. Many EPVs are millions of years old, and are genetically "fixed" in the genomes of host species (i.e., all members of the species have the integrated EPV in their genomes). Such ancient EPV sequences are in some ways analogous to "parvovirus fossils", since they preserve information about the ancient parvoviruses that infected ancestral animals.

Among the novel parvovirus groups identified via sequencing, one—provisionally labeled "chapparvovirus"—stands out as being particularly unusual. These viruses, which have been primarily reported via metagenomic sequencing of animal feces, derive their name from an acronym (CHAP), referring to the host groups in which they were first identified (Chiropteran–Avian–Porcine) [15,16,19,20]. Subsequently, several additional chapparvovirus (ChPV) sequences have been reported, including some that were identified in whole genome sequence (WGS) data derived from vertebrates, including reptiles, mammals, and birds [9]. These sequences were picked up by in silico screens designed to detect EPVs. However, since all the ChPV sequences identified in WGS data lack clear evidence of genomic integration, it is likely that they actually derive from infectious ChPV genomes that contaminated WGS samples, rather than from endogenous elements [9].

Until relatively recently, evidence that the ChPVs detected via sequencing actually infected vertebrate hosts has been lacking. However, a recent study has claimed to demonstrate that a ChPV called mouse kidney parvovirus (MKPV) circulates among laboratory mice populations, in which it causes a kidney disease known as inclusion body nephropathy [21]. These findings, as well as their frequent presence in the feces of livestock, imply that ChPVs might be pathogenic and represent a potential disease threat to wildlife and domestic species. In addition, they have raised interest in the use of these viruses as experimental tools. In this study, we perform a comparative analysis of ChPV genomes and ChPV-derived EPVs, revealing new insights into the biology and evolution of this poorly understood group.

2. Materials and Methods

2.1. Genome Screening and Sequence Analysis

All WGS data were obtained from the National Center for Biotechnology Information (NCBI) genomes resource. We obtained all available genomes for eumetazoan animals as of October 2018. These data were screened for ChPV sequences using the database-integrated genome screening (DIGS) tool [22]. ChPV sequences were characterized and annotated using Artemis Genome Browser [23]. The NCBI Basic Local Alignment Search Tool (BLAST) program and its local executables were used to compare sequences and investigate predicted viral ORFs. To determine potential homology and sequence similarity, even between previously undescribed ORFs, we constructed a local database, including all ORFs exceeding 100 amino acids (aa) in length, derived from all the exogenous and endogenous sequences incorporated in this study, and used the local BLAST P and X algorithms to conduct similarity searches in it. Two ORFs were accepted as homologous if they gave a significant hit, in the case of an expectation value threshold of 1.

Promoters were predicted using the neural network-based promoter prediction server of the Berkeley Drosophila Genome Project and further verified by the Promoter Prediction 2.0 server [24,25]. Splice sites were also detected using the neural network-based applications of the Berkeley Drosophila Genome project and SplicePort [25,26]. Polyadenylation signals were predicted by the SoftBerry

application POLYAH [27]. To verify that these applications were be capable of detecting the above-mentioned chapparvoviral transcription elements we ran MKPV through the workflow pipeline.

2.2. Phylogenetic Reconstructions

The derived aa sequences of ORFs disclosing homology to parvoviral NS1 proteins were aligned with at least five representatives of each parvovirus genus, or with one representative of each species of a given genus in cases where the number of species did not exceed five. To ensure the correct identification of the tripartite helicase domain, structural data was also incorporated into alignment construction using T-coffee Expresso [28] and Muscle [29]. The full-length NS1 derived aa sequences of the ChPV clade were aligned by Muscle and the M-coffee algorithm of T-coffee [30]. Model selection was carried out by ProTest and the substitution models RtREV+I+G, in cases of helicase-based inferences, and LG+I+G, for the complete chapparvoviral NS1 tree, were predicted to be the most suitable, based on both Akaike and Bayes information criteria. The PhyML-3.1 program was used to infer a maximum likelihood phylogenetic tree, with 100 bootstrap iterations [31], based on a guide tree previously constructed by the ProtDist and Fitch programs of the Phylip 3.697 package [32].

2.3. Homology Modeling and DNA Structure Prediction

Structural homology was detected by applying the pGenTHREADER and pDomTHREADER algorithms of the PSIPRED Protein Sequence Analysis Workbench [33]. The same workbench was used to map disordered regions using DISOPRED3 and to predict the secondary structure of the complete chapparvoviral VP protein sequences via the PSIPRED algorithm. The selected PDB structures were applied as templates for homology modeling, carried out by the I-TASSER Standalone Package v.5.1 [34]. To guide the modeling, the predicted secondary structures were applied as a restriction. The Oligomer Generator feature of the Viper web database (http://viperdb.scripps.edu/) [35] was used to construct 60-mers of the acquired putative VP monomer structures. Surface images of the capsids were rendered using the PyMOL Molecular Graphics System [36]. Capsid surface maps and VP monomer superposition were carried out by UCSF Chimera [37]. To predict the presence of potential DNA secondary structural elements, the DNA Folding Form algorithm of the mFold web server was utilized [38].

3. Results

3.1. Comparative Analysis of Previously Reported ChPV Genomes

We performed a comparative analysis of nine previously sequenced ChPV genomes so that we could: (i) identify genome features that characterize these viruses, and (ii) make inferences about aspects of ChPV biology and evolution (Figure 1). ChPV genomes tend toward the shorter end of the parvovirus genome size range (~4 kb). They encode a relatively long *rep* gene, and a relatively short *cap* gene. The *rep* gene product (NS) is ~650 amino acids (aa) in length, with the longest example being the 668 aa protein encoded by *Desmodus rotondus* ChPV (DrChPV). ChPV NS proteins contain ATPase and helicase domains, but these are the only regions exhibiting clear homology to those found in other parvovirus groups (Figure 1). Overlapping the *rep*, a predicted minor ORF, ~220 aa in length, is located in a position equivalent to that of the nucleoprotein (NP) ORF found in certain *Parvovirinae* genera (i.e., *Ave-* and *Bocaparvovirus*). However, it should be noted that the protein encoded by this gene—which we tentatively refer to as NP—exhibits no significant similarity to any other parvovirus NP proteins. Secondary structure predictions indicate that the vast majority of the NP protein has a helical structure, with numerous potential phosphorylation sites as well as a potentially protein-binding disordered N-terminus (Figure S1). Together, these observations suggest a non-structural function. The NP ORF, although of similar length in all genomes, has no canonical start codon in the case of porcine parvovirus 7 (PPV7) and simian parvo-like virus 3 (SiPV3). This would imply that in these viruses, splicing of the *rep* RNA is required for expression of the NP protein.

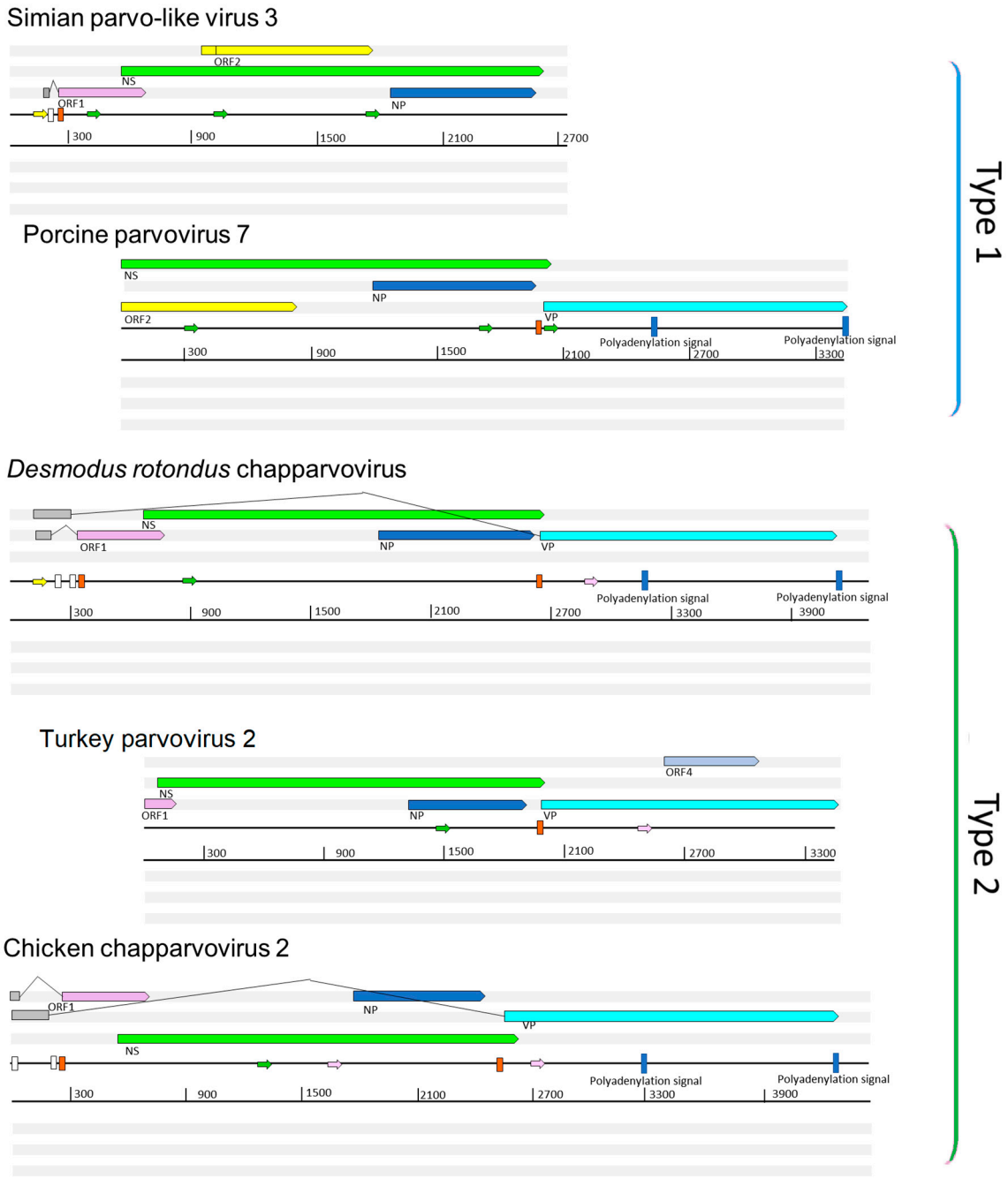

Figure 1. Representative complete coding sequence and partial genome organizations of the two distinct types of exogenous amniote chapparvoviruses (ChPVs). Open reading frames (ORFs) are represented by rectangular arrows, colored according to homology. Splice donor sites are marked by white-colored bars, acceptor sites by orange-colored bars. Blue-colored bars show predicted polyadenylation signals. Small arrows show predicted promoters and are colored according to prediction score (>0.95 = green; 0.9–0.95 = pink: <0.9 = yellow). Grey boxes indicate regions inferred to be transcribed but not translated. Note: ORF4 is unique to turkey parvovirus and is not found in other avian type 2 ChPVs, such as chicken ChPV2.

All ChPVs appear to be characterized by relatively short VP ORFs of 450–500 aa. VP proteins are typically ~650–820 aa in most other parvoviruses, the exception being the brevi- and penstyldensoviruses, which encode an even shorter VP. Notably, the VP proteins encoded by ChPVs share no significant sequence similarity with those of other parvoviruses. In all ChPVs, the first

methionine of the VP ORF is preceded by a potential coding sequence, and in all published ChPV sequences, a canonical splice acceptor site is located immediately upstream. Possibly, the VP ORF encodes only the major capsid protein, and there may be other versions of this VP protein that are elongated at the N-terminus, and are incorporated into the capsid at a lower copy number, as found in the majority of parvoviruses [1]. However, the only splice donor sites we identified are located relatively far upstream. In MKPV, however, there are two large introns present, putting these upstream exons in frame with the VP encoding exon.

In addition to their fundamental NS–NP–VP genome organization, ChPVs encode various additional small ORFs. ORF1 is predicted to encode a small protein of approximately 15 kDa that contains a putative nuclear localization signal (NLS) in its C-terminal region. ORF1, which partially overlaps the N-terminal region of NS, is present in all genomes except PPV7. However, since the PPV7 genome also lacks the corresponding region of NS, this likely reflects a 5′ truncated genome sequence. The same is the case for turkey parvovirus (TPV2), although the C-terminal encoding region of the putative ORF1 protein could be revealed.

A second additional, putative ORF is present in only two of the ChPVs examined here: PPV7 and simian parvo-like virus 3. This ORF, referred to as ORF2, is located downstream from ORF1 in a position completely overlapping the NS ORF. The TPV2 genome also contains a unique, presumably genome-specific additional ORF (ORF4) that overlaps the C-term encoding region of VP, and may encode a predicted 17 kDa protein (Figure 1). Interestingly, this ORF was absent from the other, closely-related avian ChPVs.

Analysis in silico revealed at least three potential promoters in ChPV genomes. One of these is conserved throughout the clade, and is located upstream of all coding features, indicating that it likely drives early expression of virus genes. Moreover, its presence has been confirmed in MKPV by sequencing of cDNA derived from infected mouse tissue. None of the other potential promoters proved to be functional in the case of MKPV. The MKPV transcriptome includes three transcripts confirmed to undergo splicing. Of these, however, only the one with the shortest intron could be confidently predicted in all GenBank sequences with a complete or near complete coding region (Figure 1). Interestingly, DrChPV (similar to rodent-derived ChPVs) and chicken ChPV2 (similar to TPV2, but with a more complete 5′ end) were both predicted to possess the large intron of the putative VP transcript, and therefore appear to utilize a strikingly MKPV-like transcription mechanism, despite missing an acceptor site upstream of the NP start codon. In all ChPVs examined, with the exception of the 3′ truncated entries, we identified two potential polyadenylation signals in positions equivalent to those found in MKPV [21]. This implies that the polyadenylation strategy is a conserved feature of ChPV transcription.

3.2. Identification and Characterization of Novel ChPVs and ChPV-Derived EPVs

We systematically screened published WGS data and identified a total of 15 previously unreported ChPV-derived DNA sequences. Two were identified in vertebrates and 13 in invertebrates (Table 1). The majority of the novel ChPV sequences identified in our screen were derived from the non-structural protein gene (*rep*), but we identified complete sequences derived from both the *rep* and capsid (*cap*) genes in two species: the Gulf pipefish (*Syngnathus scovelli*) and the black widow spider (*Latrodectus hesperus*). Partial *cap* genes were identified in the scarab beetle (*Oryctes borbonicus*), taurus scarab (*Onthophagus taurus*), and Chinese golden scorpion (*Mesobuthus martensii*) elements (Figure 2).

We identified two chapparvoviral sequences in WGS assemblies of syngnathid fish (family Syngnathidae), including the tiger tail seahorse (*Hippocampus comes*) and the Gulf pipefish (*Syngnathus scovelli*). The pipefish sequence occurs in a relatively short scaffold (4002 nt) that is entirely comprised of viral sequence, displaying truncated, but nonetheless detectable, J-shaped terminal hairpin-like structures (Figure S2). This suggests it likely represents a virus contaminant, as suspected for other ChPV sequences recovered from vertebrate WGS data [9]. The virus from which this sequence was presumably derived was designated *Syngnathus scovelli* ChPV (ScChPV).

The seahorse and invertebrate sequences identified in our screen clearly represented EPVs (see below). However, the pipefish sequence lacked flanking genomic sequences and appeared to derive from an exogenous virus, encompassed by truncated hairpin-like secondary structure repeats (Figure S2). None of the ChPV-derived EPVs we identified shared homologous flanking sequences, indicating that each derives from a distinct germline incorporation event.

We identified a total of 13 EPV sequences that disclosed a relatively close phylogenetic relationship to ChPVs. These elements showed varying degrees of degradation. In many cases, only genome fragments were detected (Figure 2), and these usually included multiple nonsense mutations (Table 1). ChPV-derived elements were detected in three major arthropod clades that primarily occupy terrestrial habitats, namely arachnids of Chelicerata, chilopods of Myriapoda, as well as hexapod insects and entognaths.

We used maximum likelihood-based phylogenetic approaches and an alignment spanning the tripartite helicase domain of the NS protein to reconstruct the evolutionary relationships of ChPVs, ChPV-derived EPVs, and previously reported parvoviruses (Figure 3). Strikingly, reconstructions indicated that the family *Parvoviridae* consists of four major clades, rather than the two that have historically been recognized [1]. Of these four lineages, one corresponds to the subfamily *Parvovirinae* as in current taxonomic schemes. However, the subfamily *Densovirinae* is split into two clades; one encompassing all ambisense densoviruses along with viruses of the genus *Iteradensovirus* (which have monosense genomes) and the second, referred to here as HBP, containing the *Hepan-*, *Brevi-*, and *Penstyldensovirus* genera. Moreover, a fourth parvovirus lineage was evident, comprised of the ChPVs and ChPV-derived EPVs.

Table 1. Novel ChPV sequences identified in this study.

Host Common Name	Host Scientific Name	Virus/Element Name [a]	Gene Content	Nonsense Mutations [b]
Vertebrates				
Gulf pipefish	*Syngnathus scovelli*	ScChPV	*rep*+*cap*	0; 0
Tiger tail seahorse	*Hippocampus comes*	ChPV.1-HipCom	*rep*	2; 2
Invertebrates				
Black widow spider	*Latrodectus hesperus*	ChPV.2-LatHes	*rep*+*cap*	4; 1
		ChPV.3-LatHes	*rep*+*cap*	3; 1
		ChPV.4-LatHes	*rep* *	3; 3
		ChPV.5-LatHes	*rep*	4; 2
Chinese scorpion	*Mesobuthus martensii*	ChPV.6-MesMar	*rep*+*cap* *	2; 3
European centipede	*Strigamia maritima*	ChPV.7-StrMar	*rep*	2; 3
Northern forcepstail	*Catajapyx aquilonaris*	ChPV.8-CatAqu	*rep*	0; 0
Emerald ash borer	*Agrilus planipennis*	ChPV.9-AgrPla	*rep**	2; 0
Taurus scarab	*Onthophagus taurus*	ChPV.10-OntTau	*rep*	2; 3
		ChPV.11-OntTau	*rep*	0; 0
		ChPV.12-OntTau	*rep*	2; 1
Rhinocerous beetle	*Oryctes borbonicus*	ChPV.13-OryBor	*rep*	2; 0
		ChPV.14-OryBor	*rep*	0; 0

[a] For sequences that are presumed to derive from viruses, the proposed name of the virus is shown. For endogenous parvoviral elements (EPV) the locus name is given, following the standard nomenclature proposed for endogenous retrovirus (ERV) loci [39], except using the classifier "EPV" in the place of "ERV". The table shows a shortened version of the ID, used in the text of this manuscript, wherein the "EPV" classifier is omitted, and an abbreviated version of the species name is used within the taxonomic component of the ID (derived from the first three letters of each component of the Latin binomial scientific name of the host species). [b] Stop codons; frameshifts. * Asterisks indicate contigs that were truncated within the virus-derived portion of the sequence. Underlined names indicate the presence of the complete ORF.

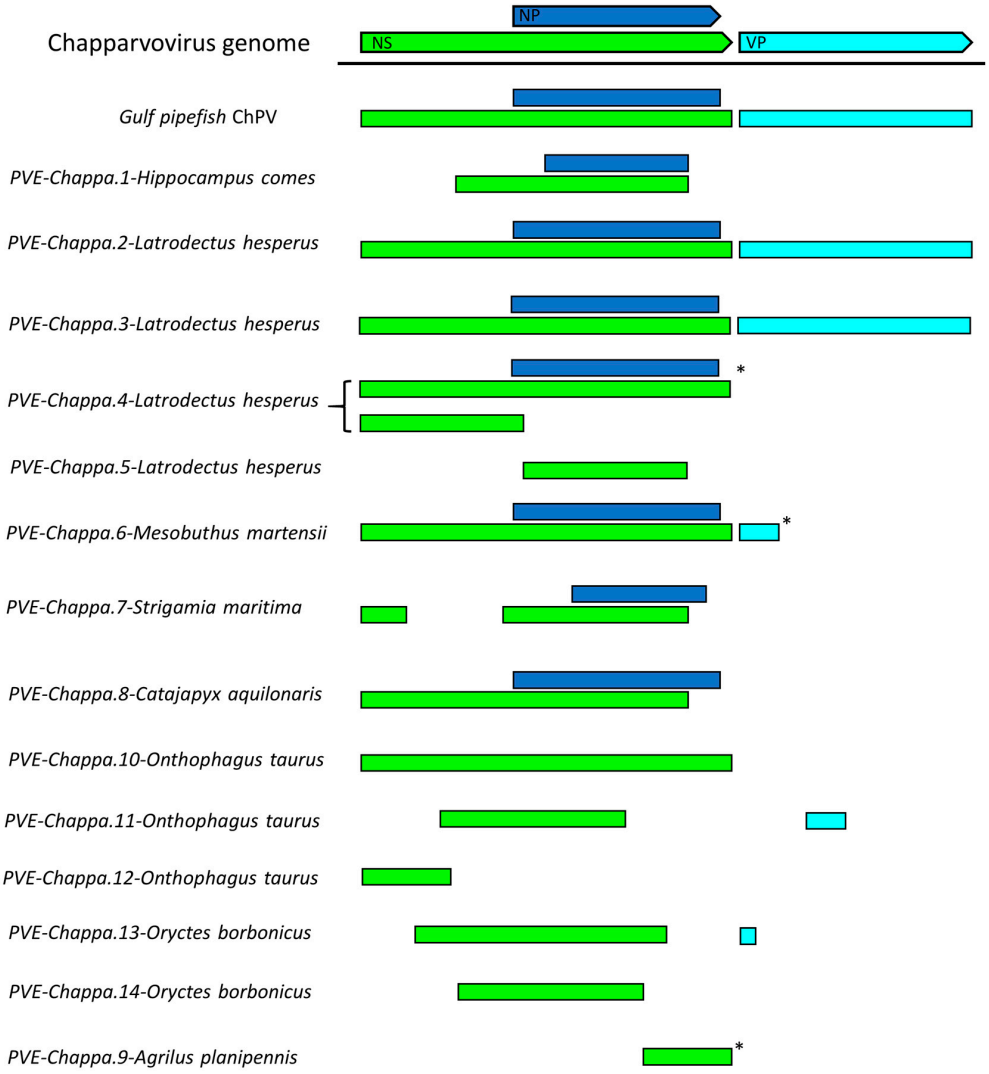

Figure 2. Basic gene content of newly identified chapparvoviruses (ChPVs) and ChPV-derived endogenous parvoviral element (EPV) sequences, shown in relation to a representative ChPV genome (mouse kidney parvovirus). Asterisks indicate contigs that were truncated within the virus-derived portion of the sequence. Abbreviations: non-structural protein (NS); capsid protein (VP); nucleoprotein (NP).

The branching relationships between ChPVs were not fully resolved by phylogenetic analysis of the helicase domain. The putative large non-structural proteins (NS1) of ChPVs displayed a high degree of amino acid variability, particularly toward their N- and C-term. However, a region ~500-aa-long could be aligned reliably throughout all complete and partial entries previously proven to cluster within the ChPV lineage in the case of the NS helicase-based inference. Phylogenies reconstructed from this alignment reveal the ChPV-related viruses to be comprised of three robustly supported monophyletic lineages (Figure 4). One of these includes ChPVs sampled from amniotes (reptiles, birds, and mammals), in which two robustly supported sublineages (labeled type 1 and 2) were observed, corroborating the helicase-based phylogeny. The amniote ChPVs form a sister clade to EPVs found in the arthropod subphyla Chelicerata (arachnids, camel spiders, scorpions, whip scorpions, harvestmen, horseshoe crabs, and kin) and Myriapoda (millipedes, centipedes, and kin) as well as syngnathe fish. A third lineage was also observed, containing sequences from the arthropod subphylum Hexapoda (insects, springtail, and forcepstail). Within this lineage, the beetle EPVs formed a well-supported monophyletic clade.

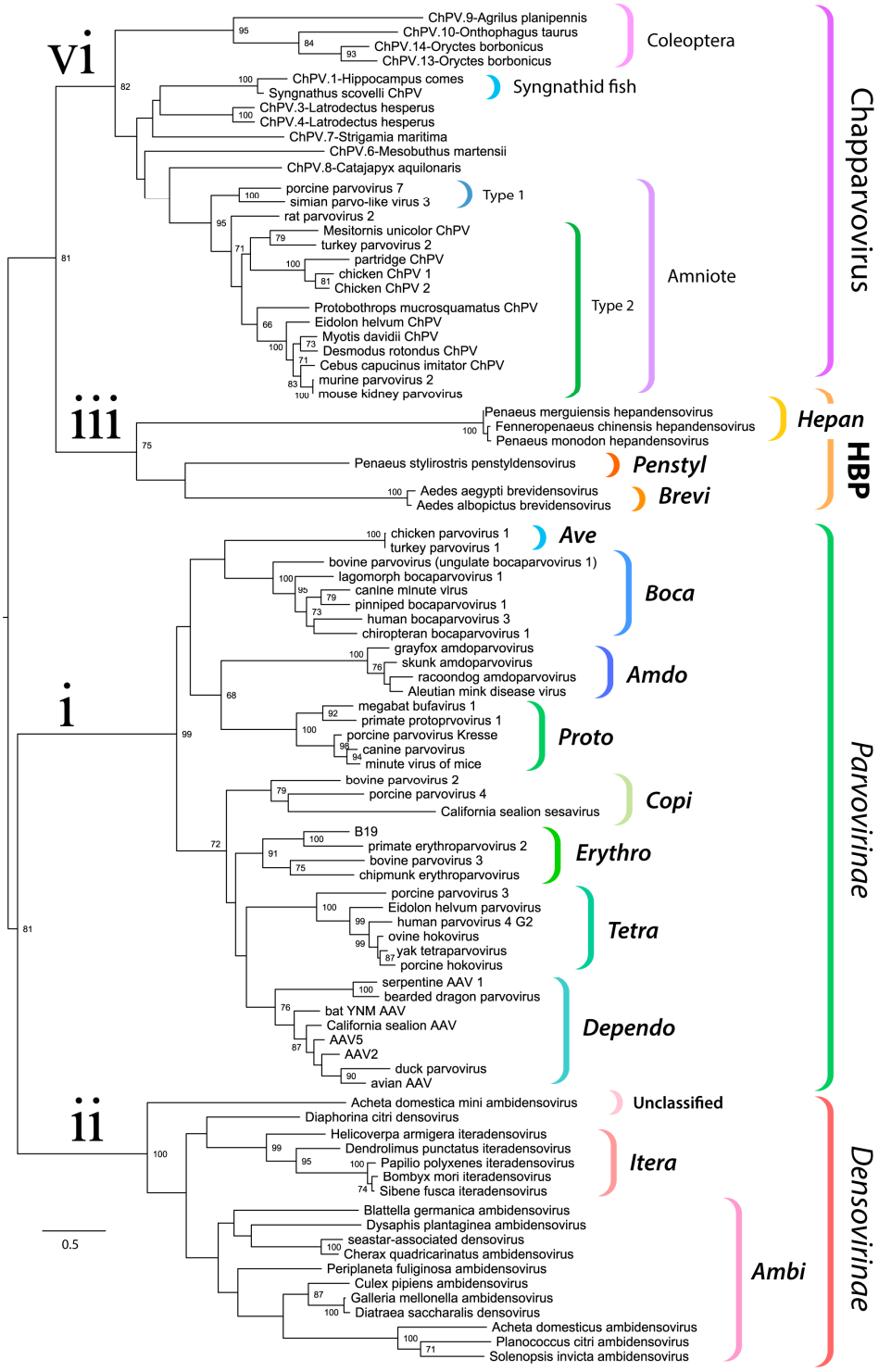

Figure 3. Evolutionary relationships within the family *Parvoviridae* reconstructed via phylogenetic analysis of the tripartite helicase domain. The four major splits within the *Parvoviridae* are indicated in the tree as follows: (i) *Parvovirinae* (ii) *Densovirinae* (excluding genera *Hepan-*, *Brevi-*, and *Penstyldensovirus*, abbreviated as HBP); (iii) HBP; (iv) Chapparvovirus (ChPV-related viruses and EPVs). Brackets to the right indicate taxonomic groups. The names of established genera are shown in bold italics in the abbreviated form (i.e., with the suffix "parvovirus" omitted). The scale bar shows evolutionary distance in substitutions per site. Numbers adjacent to tree nodes show bootstrap support (based on 100 bootstrap replicates) where >70%.

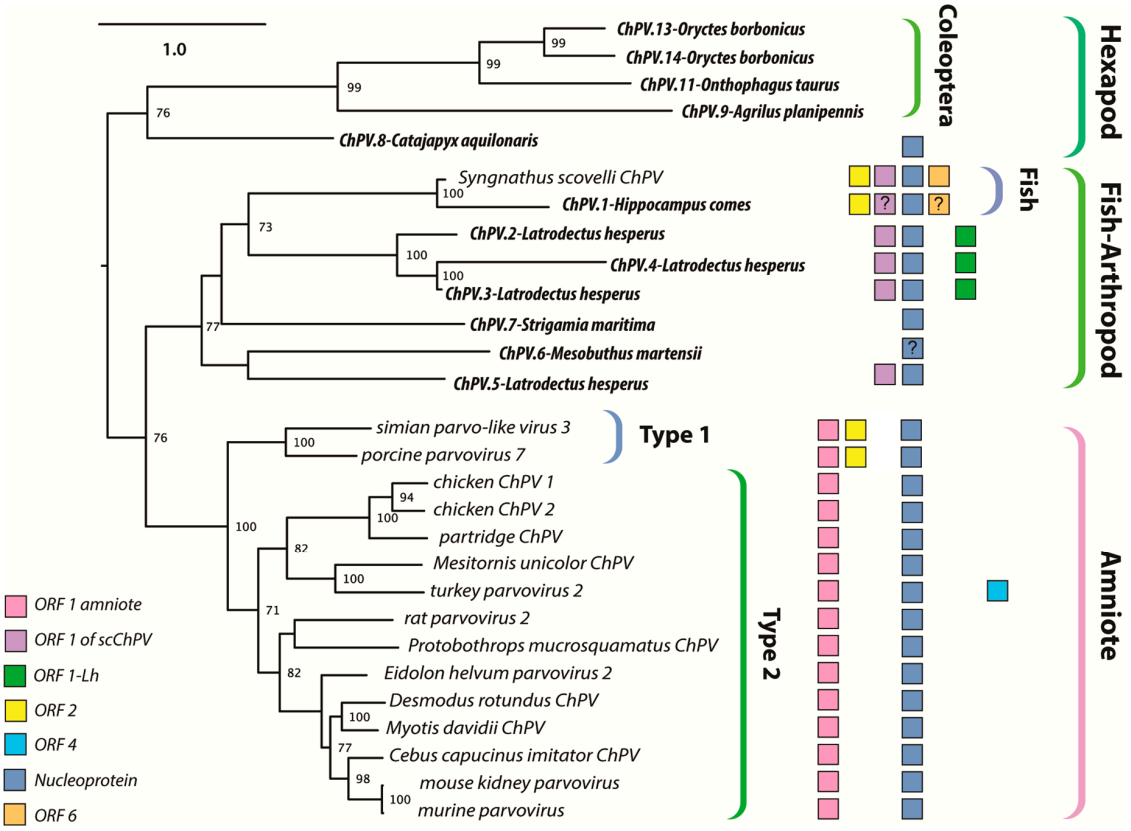

Figure 4. Maximum likelihood phylogenetic reconstructions of the ChPV clade based on the complete aligned amino acid sequences of the NS1. Colored boxes indicate the presence of auxiliary open reading frames (ORFs), as shown in the inset key. The "?" character indicates that the presence of an ORF is suspected but not confirmed. Taxa labels in bold italics indicate endogenous sequences, whereas italics indicate sequences known or believed to derive from viruses. Brackets to the right indicate taxonomic groups. The scale bar (top right) shows evolutionary distance in substitutions per site. Numbers adjacent to tree nodes show bootstrap support (based on 100 bootstrap replicates) where >70%.

3.3. Characterization of Syngnathid ChPVs and EPVs

The ScChPV genome encodes a long NS ORF (807 aa), a strikingly short VP (367 aa), and a ChPV-like NP (Figure 5). Furthermore, a homologue of the ORF2 protein found in the amniote parvoviruses PPV7 and SPV3 was present. A predicted ORF was present in a genomic position equivalent to that of ORF1, found in amniote ChPVs. However, the predicted protein sequence did not disclose any detectable similarity to its amniote counterpart. ORF6, identified in partial overlap with the VP C-term encoding region, encodes a small protein of 27.2 kDa (239 aa), exhibiting no detectable similarity to any other sequence in GenBank. Fold recognition, however, revealed a potential structural similarity to viral structural proteins, including the major envelope glycoprotein of the Epstein–Barr virus (PDB ID: 2H6O chain A, $p = 0.012$), the minor viral protein of the Sputnik virophage (PDB ID: 3J26, chain N, $p = 0.017$) and the surface region of *Galleria mellonella* ambidensovirus (PDB ID: 1DNV, $p = 0.021$). These findings imply ORF6 may encode an auxiliary structural protein.

The partial ChPV-like sequence identified in the genome of the tiger seahorse (*Hippocampus comes*) was flanked by extensive stretches of host genomic sequence, establishing that, unlike the ScChPV sequence identified in the Gulf pipefish genome assembly, it likely represents an EPV rather than a virus. Interestingly, however, phylogenies showed that both sequences obtained from syngnathe fish are relatively closely related, and cluster together with high bootstrap support (Figures 3 and 4).

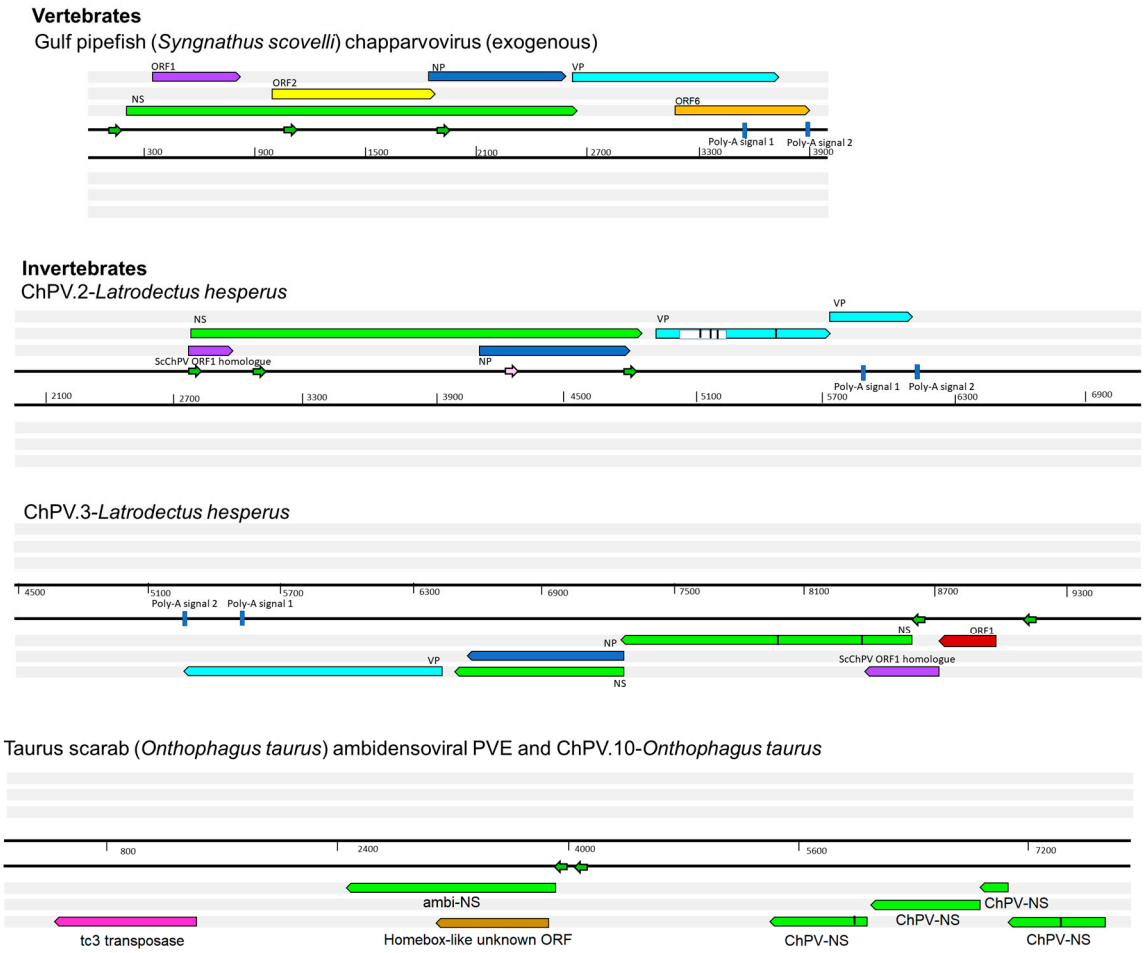

Figure 5. Genomic structures of newly identified chapparvoviruses (ChPVs) and ChPV-derived endogenous parvoviral element (EPV) sequences. The positions of putative open reading frames (ORFs) and predicted cis transcription elements of ChPVs are shown. ChPV.2-LatHes contains a previously unidentified repetitive element, present as multiple copies scattered in the *Latrodectus* genome, marked by the white box within the VP gene. The element ChPV.10-OntTau shares its integration site with another endogenous element, disclosing similarity to ambidensoviruses. ORFs are represented by rectangular arrows, colored according to homology. In-frame stop codons are shown as vertical lines. Splice donor sites are marked by white-colored bars, acceptor sites by orange-colored bars. Blue-colored bars show predicted polyadenylation signals. Small arrows show predicted promoters and are colored according to prediction score (>0.95 = green; 0.9–0.95 = pink: <0.9 = yellow). Grey boxes indicate regions inferred to be transcribed but not translated.

3.4. Characterization of ChPV-Derived EPVs in Invertebrate Genomes

Among the ChPV-derived EPVs we identified in invertebrates, the most complete were identified in the western black widow spider (*Latrodectus hesperus*) (Figure 5). Two of these elements spanned near complete genomes, including *rep*, *cap,* and NP genes, and a homologue of the ORF1 gene found in ScChPV. In addition, the ChPV.2-LatHes element encodes an apparently complete NS protein (690 aa), while ChPV.3-LatHes discloses an undisrupted ORF1 (113 aa) as well as an apparently intact *cap* gene encoding a 386-aa-long VP. The putative NS1 proteins encoded by these elements displayed only 62% identity at aa level. The disrupted *cap* gene of ChPV.2-LatHes was found to include an insertion of 74 aa, suspected to originate from a yet unknown repetitive element (revealed by sequence comparisons to be interspersed throughout the *L. hesperus* genome).

ChPV.3-LatHes, on the other hand, appeared to include an intact upstream region of the genome, revealing an additional small ORF of 81 aa length directly upstream of the ScChPV ORF1 homologue,

designated ORF1-Lh. This ORF disclosed no detectable homology to any sequences to date. Upstream of this ORF, a potential promoter sequence could be identified with high confidence (0.98 of 1). Both elements included complete, NP-encoding ORFs of 233 aa, although a canonical ATG start codon could only be identified in one element (ChPV.3-LatHes).

We identified two further elements in the western black widow spider genome, although these only spanned disrupted *rep* genes. ChPV.4-LatHes encodes nearly complete NS and NP genes, as well as a complete homologue of the ScChPV ORF1 gene. The true extent of preservation could not be assessed for this EPV as it occurs on a short scaffold that terminated within the EPV *rep* sequence. The putative NS1 of ChPV.4-LatHes was 80% identical to its counterpart in ChPV.2-LatHes at aa level. Interestingly, this element contains additional, *rep*-encoding regions directly upstream of a larger, NS1- and ScChPV ORF1-encoding region. This second region encodes only the first 221 aa of the putative NS1 protein, together with the putative ScChPV ORF1 homologue and ORF1-Lh genes. The ORF1-Lh gene encoded by ChPV.4-LatHes lacks an ATG start codon. The upstream promoter was weakly predicted, with a score of 0.6. ChPV.5-LatHes displayed a highly divergent, partial *rep* of 216 aa, with only 42% identity to the ChPV.2-LatHes NS1 at aa level (Figure 5). This element clustered outside the monophyletic clade defined by the three other *Latrodectus* EPVs (Figure 4).

A single ChPV-derived EPV was identified in a second arachnid species—the Chinese golden scorpion (*Mesobuthus martensii*). This element was identified in a relatively short, unplaced scaffold, and comparison to the *Latrodectus* elements indicated that the contig was truncated within the EPV sequence, consequently the true extent of its preservation could not be assessed. Nevertheless, ORFs disclosing homology to the NS, NP, and VP proteins could be identified (Figure 2). While the first 100 or so codons of the NS ORF were absent, a complete NP ORF was detected, along with the first 46 codons of VP. All three ORFs were disrupted by frameshifts and stop codons. No homologues of any alternative ORFs identified in other ChPV genomes could be identified.

We identified a ChPV-derived EPV in the genome of a myriapod—the European centipede (*Strigamia maritima*). This element displayed partial homologues of the NS and NP encoding ORFs, both of which contained large deletions (Figure 2) as well as numerous nonsense mutations (Table 1). Moreover, the NS ORF was disrupted by an extensive stretch of an insertion of unknown origin. No homologues of any of the alternative ORFs found in other ChPVs could be identified in this endogenous sequence.

Seven ChPV-derived EPVs were identified in hexapod arthropods (subphylum Hexapoda). One occurs in the genome of a bristletail species—the Northern forcepstail (*Catajapyx aquillonaris*)—belonging to the entognath order Diplura. The other six were identified in three species belonging to the vast insect order Coleoptera: the emerald ash borer (*Agrilus planipennis*), the taurus scarab (*Onthophagus taurus*), and the scarab beetle (*Oryctes borbonicus*). The bristletail element contains a C-terminal truncated *rep* of at least 250 aa and a near full-length NP ORF. The partial *rep* was intact, but the NP ORF is disrupted and highly divergent, showing significant sequence similarity only in the conserved core region of the putative protein. The ash borer element ChPV.9-AgrPla occurs in a scaffold that is ~1 kb in length. One end of this scaffold contains a 592 nt region exhibiting homology to the NS ORF, which harboured an N-terminal deletion of at least 200 aa.

In ChPV.10-OntTau, a disrupted but almost complete NS ORF could be identified (Figure 5). Interestingly, a second EPV insertion was detected at the same locus. This element encodes an intact, potentially fully-expressible NS gene, homologous to the NS1 of ambidensoviruses (genus *Ambidensovirus*) and disclosing similarity to a recently reported ambidensovirus sequence that has been detected only at cDNA level in the transcriptome of two bumble bee species (*Bombus cryptarum* and *B. terrestris*) [40]. An additional intact, potentially expressible ORF was present in this ambidensoviral element, overlapping the putative NS1 gene, which harboured no significant similarity to any sequences deposited in GenBank to date. In its derived aa sequence, however, a homeobox domain could be revealed. The other two elements of the taurus scarab genome were located together in another assembly scaffold, only 2540 nts apart from each other. Both EPVs consisted of only a partial ORF,

which disclosed similarity to chapparvoviral *rep*s. None of these elements encompassed the tripartite helicase domain, hence they were not included in the phylogenetic inference.

Two EPVs were identified in the scarab beetle genome. One of these, designated ChPV.13-OryBor, harboured a near complete *rep* at 402 aa, as well as a short, partial *cap*, capable of encoding only the first 33 aa of the putative VP. The region of *rep* homology occurred within an ORF that was not disrupted by any frameshifts and could be extended without disruption upstream and downstream, suggesting that a longer gene product—potentially encoding a longer, divergent NS protein—may be present. However, these regions did not disclose sequence similarity to any proteins hitherto deposited to GenBank. The ChPV.14-OryBor element included only a heavily truncated NS of 254 aa.

3.5. Structural Characteristics of ChPV Capsids

We built 3D homology models to facilitate the comparison of ChPV capsid structures to those found in other parvoviruses. Interestingly, structural similarity with erythro-, proto-, and bocaparvoviruses can be detected for VP using fold recognition, even though the VP proteins of ChPVs share no significant sequence similarity with those of other parvoviruses (Figure 6a).

The derived polypeptide sequence of the complete VP ORF encoded by DrChPV was subjected to fold recognition, to identify suitable templates for homology modeling. This comparative analysis showed that the VP2 protein of parvovirus H1 (genus *Protoparvovirus*) (PDB ID: 4G0R) could potentially harbor the most structural similarity ($p = 9 \times 10^{-5}$), and this sequence was therefore used as the template for homology modeling. Due to the lack of sequence identity and the non-homologous nature of the ChPV VP genes to other parvoviral VPs, we used the final model obtained in this analysis as a template to construct homology models for four further ChPV VPs—rat parvovirus 2, PPV7, TPV2, and pit viper ChPV. This allowed us to overcome the stochastic aspect of model construction. Although the pitfalls of using models as templates have to be noted, this approach ensured that only those regions showed structural variability which would likely do so in the actual capsid structures.

We examined the VP sequences of two representatives of the second major ChPV clade (see Figure 4)—one derived from a presumably exogenous virus (ScChPV) and one from an EPV (ChPV.3-LatHes). For the VP protein encoded by ScChPV, fold recognition identified the following dependoparvovirus VP3 proteins as potential templates: adeno-associated virus 8, PDB ID: 2QA0, $p = 9 \times 10^{-4}$; Adeno-associated virus rh32.33, PDB ID: 4IOV, $p = 9 \times 10^{-4}$, while for the VP encoded by the black widow spider EPV the most reliable hit was the VP4 protein of an iteradensovirus (*Bombyx mori* densovirus 1, PDB ID: 3P0S, $p = 8 \times 10^{-4}$). When superimposing the obtained models with the VPs of AAV8 and BmDV1, however, structural similarity only covered the jelly roll core and the αA helix, and of the surface loops traditionally considered more variable, only the BC loop.

Modeling indicated that the ChPV VP monomer harbors an eight-stranded β-barrel "jelly roll" core and the αA helix at the two-fold symmetry axis, as found in all members of the family *Parvoviridae* to date [41] (Figure 6a). Equivalents of all short strands were present (β-C, H, E, F) as well, for four out of the five longer strands (β-B, D, I, G). However, no structural analogue to the outmost β-A could be identified (Figure 6). Examining the secondary structure prediction confirmed that a β-A analogue was not present, indicating β-B to be the closest to the N-term. The first strand of the *Syngnathus scovelli* ChPV VP appeared to fold outside of the jelly roll, leaving the longer sheet of the barrel without a β-B, comprised of only three strands—namely D, I, and G—despite a complete upper, CHEF sheet (Figure 6a). When modeling the complete $T = 1$ capsid polymer, this manifested as a hole, which is normally covered by β-B, even in the case of the smallest parvoviral capsids (Figure 6b). All VPs encoded by ChPVs and ChPV-derived EPVs displayed two canonical loops surrounding their five-fold axes, linking sheets D with E at the five-fold channel and sheets H with I on the floor surrounding the channel. In case of the amniote ChPVs, the pore displayed a tight opening. The sequence of the DE loop varied to some extent among these seven sequences, which also manifested in the models. The HI loop was, however, highly conserved throughout, containing only one variable position between the amniote ChPVs.

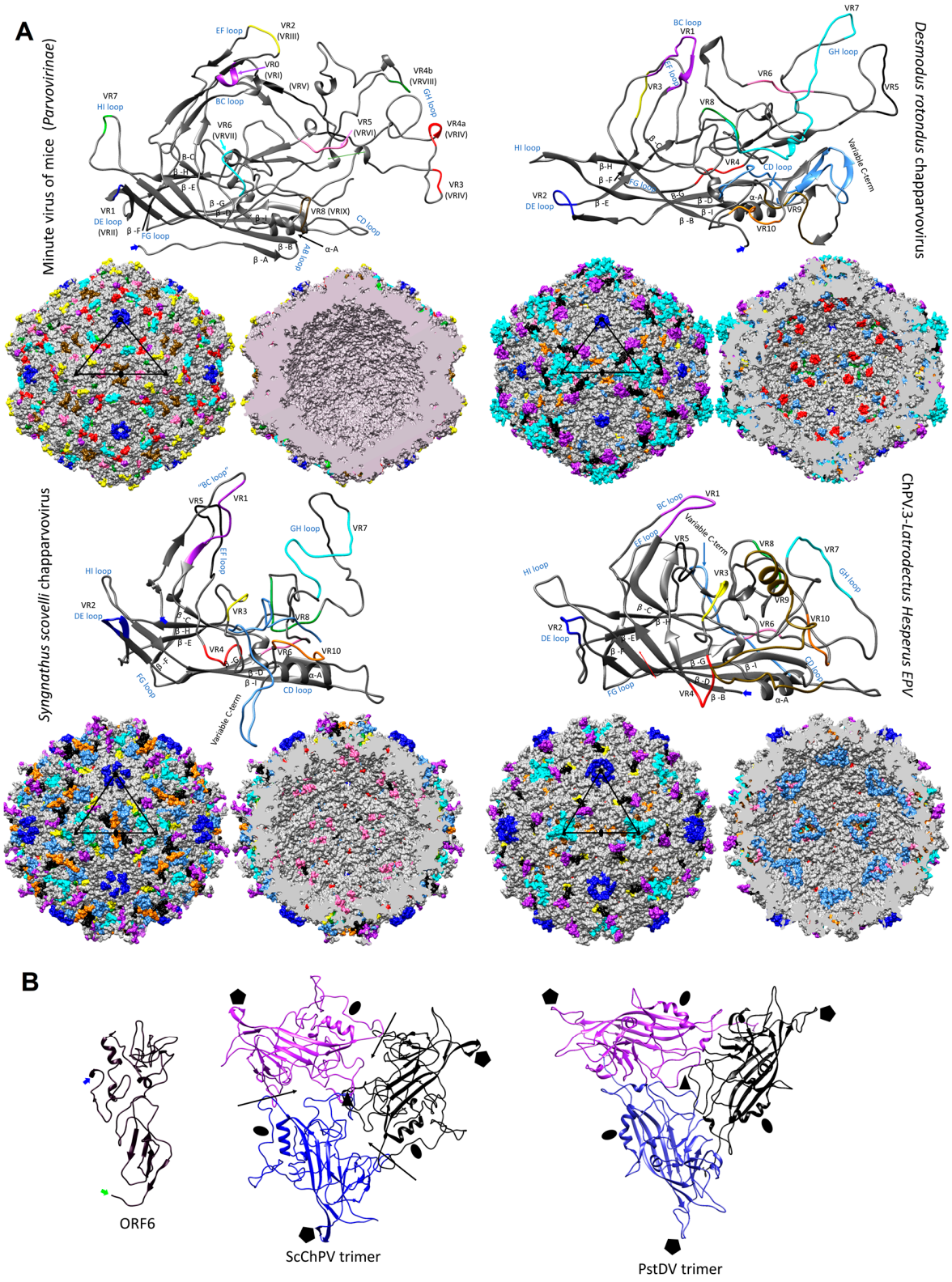

Figure 6. Structural variation and assembly interfaces of chapparvoviruses (ChPVs). (**A**) Comparison of VP monomer ribbon diagrams of the protoparvovirus minute virus of mice (PDB ID: 1Z14) from subfamily *Parvovirinae* to homology models of an amniote, a fish, and a ChPV-derived EPV from an arthropod genome (ChPV.3-LatHes). Variable regions (VRs) of the same number are marked by the

same color and mapped to the surface and luminal area of the $T = 1$ icosahedral capsid model constructed of 60 monomers. In the case of the minute virus of mice, the VRs are marked by both the traditional numbering established for dependoparvoviruses (Roman numerals) and by the special numbering applied for protoparvoviruses only (Arabic numerals). Blue signs indicate the names of the loops linking the beta strands of the conserved jelly roll core. Triangles mark the position of an asymmetric unit within the capsid, the five-fold symmetry axis is marked by a pentagon, the three-fold with the black filled triangles, and the two-fold with an ellipsoid. (**B**) Homology model of ORF6, the hypothetical structural protein of *Syngnathus scovelli* ChPV (ScChPV). The trimer of the ScChPV monomer model reveals a gap at each subunit interaction (arrows), unlike in the case of the trimer of even the hitherto smallest parvoviral capsid protein, *Penaeus stylirostris* densovirus. The gap might accommodate ORF6 in the assembled ScChPV capsid. Symmetry axes are marked by the same symbols as for panel A.

We mapped the chapparvoviral VRs identified by VP alignments (Figure S3) to both VP monomers and complete capsids, to examine how they manifest on the virion surface and make comparisons to parvoviruses of known structure, represented by the minute virus of mice (MVM), the prototypic member of subfamily *Parvovirinae* (PDB ID: 1Z14) (Figure 5). Out of ten chapparvoviral VRs identified (VR 1 to 10), shown in Figure S3a, only VR1, VR2, and VR9 proved to be similarly positioned and hence likely analogous to their counterparts in the MVM capsid. Some VRs (VR4 in all ChPVs examined, VR8 of the amniote ChPVs, and VR6 in ChPV.3-LatHes and ScChPV) appeared to be positioned at the luminal surface of the ChPV capsid, distinct from all parvoviruses studied to date. The only exception, however, is bovine parvovirus, a bocaparvovirus [42] in which VR8 is also located on the luminal surface of the capsid. Since the ChPV VRs appeared to be non-homologous to those established for either proto- or dependoparvoviruses, we re-defined them by numbering from N to C-term.

In addition to their distinctive VRs, ChPVs ubiquitously appeared to harbor a highly variable C-terminal region, with a length varying between 12 and 62 residues. The ChPV VP variable C-term appears to be buried in most cases, with the exception of ScChPV, where it is probably exposed. In the VP encoded by ChPV.3-LatHes it forms the luminal surface of the three-fold, whereas in the case of fish and amniote ChPVs it is located at the two-fold (Figure 6a).

The ScChPV and ChPV.3-LatHes VP lacked a VR6 homologous to that of the amniote ChPVs, albeit displayed variation in another position instead, still in the sixth-place counting from the N-term (Figure S3b). Moreover, both of them displayed truncated VRs 3, 5, and 7, compared to their amniote counterparts. VR9, furthermore, was absent from the ScChPV VP, whereas VR10 was missing from the VP of ChPV.2-LatHes (Figure S3b). As for the surface, the largest variable region for amniote ChPVs, namely DrChPV, is VR7, forming the entire three-fold protrusions, with VRs 1, and 9 forming small protrusions surrounding the aforementioned peaks.

The complete capsid models of non-amniote ChPVs were observed to harbor surface features that are strikingly distinct from those of the amniote ones, more closely resembling the capsids of the *Ambidensovirus-Iteradensovirus* clade of *Densovirinae* (see Figure 3), with a surface that is less spikey (Figure 7a). The homology model of ORF6 of ScCHPV, constructed based on the minor viral protein of the Sputnik virophage (PDB ID: 3J26) indicates that this potentially structural protein harbors multiple beta strands close to its C-term, out of which the outermost could potentially fill in the aforementioned gap caused by the lack of a β-B (Figure 6b).

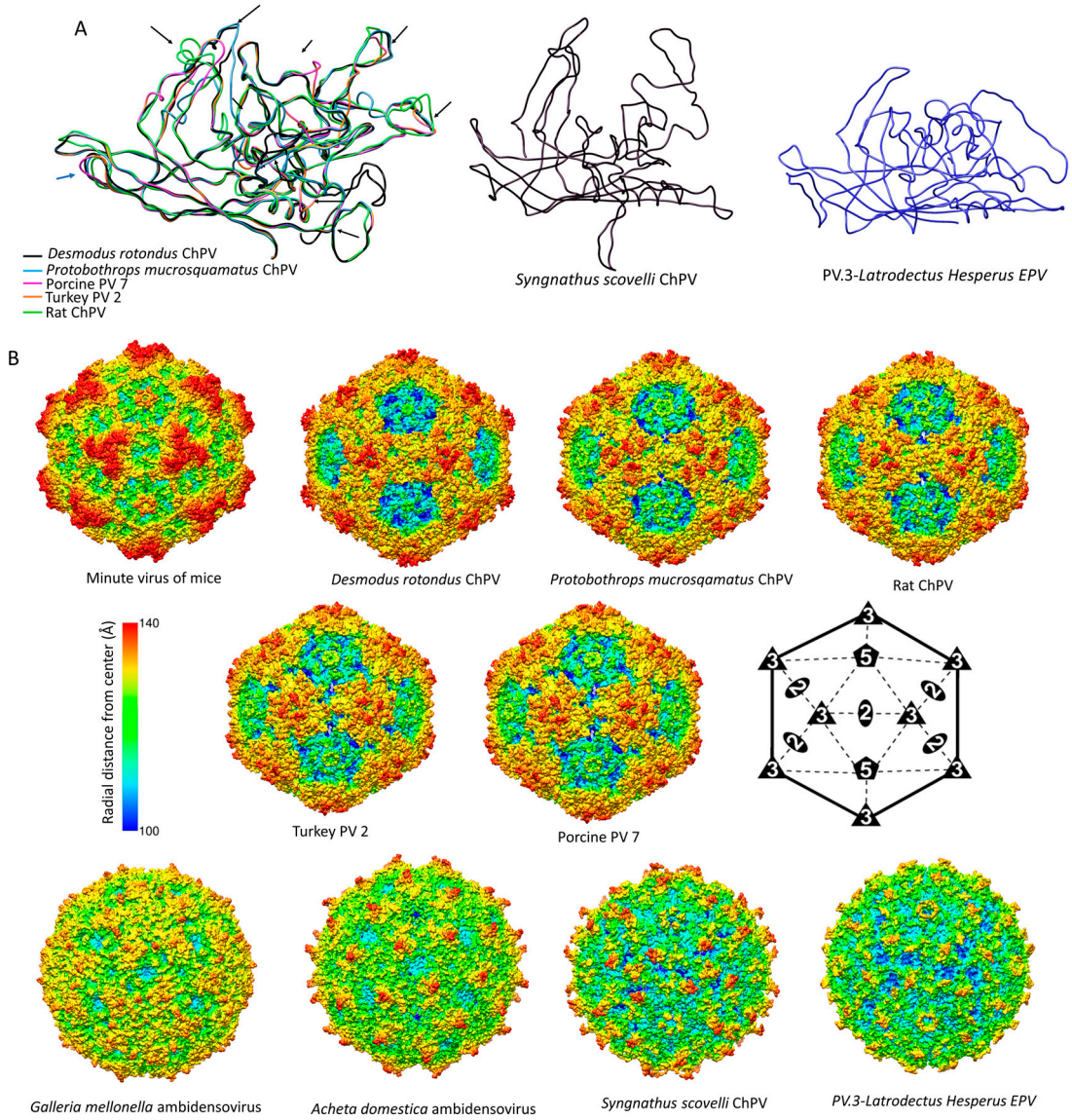

Figure 7. Comparison of chapparvoviruses (ChPV) capsid models of various host affiliations. (**A**) Homology models, shown as ribbon diagrams, representing the probable three different ChPV structural protein types. The first panel shows superposition of VP monomer homology models of amniote ChPV capsids, including reptilian, avian, rodent, chiropteran, and ungulate representatives. Black arrows show variable regions (VRs) previously identified by aligning the VP protein sequences. The next two panels show homology models of capsid monomers from a fish ChPV and an endogenous chapparvoviral element from an arthropod genome. (**B**) Capsid surface morphology of amniote ChPV homology models compared to that of the polymer structure of a prototypic parvovirus, the minute virus of mice (MVM) (PDB ID: 1Z14 at 3.25 Å resolution). Capsids are orientated by their two-fold symmetry axes, as shown in the line diagram, and are radially colored. Below, the comparison of homology models of complete viral capsid surface morphology of the newly identified fish ChPV and arachnid endogenous chapparvoviral element is shown, with that of the actual capsid structure of two densoviruses (subfamily *Densovirinae*, genus *Ambidensovirus*) (PDB ID: 4MGU at 3.5 Å resolution for *Acheta domestica* densovirus and 1DNV at 3.7 Å for *Galleria* densovirus).

4. Discussion

Historically, the family *Parvoviridae* has always been comprised of two subfamilies, with specificity for vertebrate or invertebrate hosts being the major demarcation criterion [2]. This division was initially

supported by phylogenetic inference. However, as the number of densoviral genera increased, the heterogeneity of densoviruses, specifically their segregation into two clades, has not gone unnoticed [1]. Our study provides further evidence that the traditional division of parvoviruses into vertebrate-specific and invertebrate-specific subfamilies no longer holds, rather, it supports the division of the *Parvoviridae* into four major subgroups: the *Parvovirinae*, a split *Densovirinae*, and the ChPVs, as illustrated in Figure 3.

The data presented here show that ChPVs infect an exceptionally broad range of hosts, including both vertebrates and invertebrates. We show that ChPVs found in fish are more closely related to those that infected ancestral arachnoid arthropods than they are to those that infect amniote vertebrates (Figure 4), suggesting that ChPVs may have been transmitted between distantly related host species in the past. Furthermore, phylogenies indicate that all amniote ChPVs have a common origin (Figure 3), consistent with the overall conservation of their genome organization and some aspects of predicted transcriptional strategy (Figure 1).

While previous studies have suggested that ChPVs broadly co-diverged with host species [9], the present, expanded data set reveals that some transmission of ChPVs between vertebrate classes may have occurred (Figure 4). However, it should be kept in mind that almost all amniote ChPVs have been identified via metagenomic sequencing of environmental samples (mostly fecal viromes) and their true host affiliations remain uncertain.

The EPV sequences found in animal genomes overwhelmingly derive from a small proportion of parvovirus lineages [13,14,17,43,44]. For example, ambidensovirus-derived EPVs dominate invertebrate genomes [14], whereas vertebrate EPVs almost exclusively derive from the *Dependoparvovirus* and *Protoparvovirus* genera [12,13,43,44]. In this study, we found no trace of ChPV-derived EPVs in amniote genomes, despite recent evidence that ChPVs infect this host group [21,45]. By contrast, ChPV-derived EPVs are relatively common in arthropods, with some species harboring multiple, independently acquired elements, occasionally even in close proximity within the host genome (Table 1, Figure 4). The tendency of EPVs to derive from a subset of parvovirus genera likely has biological underpinnings. For example, in vertebrates it may reflect the ability of dependoparvoviruses to integrate into host DNA, and/or the requirement of protoparvoviruses to initiate DNA damage response (DDR) during replication [46,47]. Similar features of the viral life cycle could account for the biased distribution of ChPV-related sequences in animal genomes, i.e., arthropod and fish ChPVs might have adopted a replication strategy that favors germline integration, whereas that of amniote ChPVs precludes it. Notably, some arthropod species have integration sites containing multiple independently acquired EPVs of both ChPV and ambidensovirus origin, suggesting that hotspots of parvovirus integration and/or fixation might exist in their genomes.

Our discovery of ChPV-derived elements in fish and arthropod genomes establishes that ChPVs can infect these species in addition to amniotes [21,45]. Moreover, it provides evidence that the ChPVs are likely an ancient lineage of parvoviruses. Though we did not identify any orthologous ChPV insertions, the EPVs described here show extensive evidence of germline degradation. Through comparison to studies of EPVs in mammals (in which several orthologous EPVs have been described [13,48]), it appears likely that ChPVs have been present in animals for many millions of years. Moreover, as the hexapod EPVs appear to be monophyletic and mirror the evolution of their host species, the age of ChPVs could possibly correlate with the Insecta–Entognatha split, suggesting a minimum age of 400 million years [49].

Through comparative analysis of EPVs and ChPVs, we show that ChPV genomes exhibit a number of defining characteristics. Firstly, all possess a short, monosense genome, encoding a relatively large NS and a relatively short VP. The short VP proteins of ChPVs are clearly homologous to one another, but show no similarity to those found in other parvovirus lineages. Similar to those found in the penstyl-, hepan-, and brevidensoviruses, the VP proteins of ChPVs lack the phospholipase A2 (PLA2) domains that are required for infectivity in most other parvoviruses. Notably, these are also the genera to which ChPVs are most closely related in NS-based phylogenies (Figure 3).

Secondly, ChPVs typically encode multiple additional gene products besides the NS and VP. To begin with, almost all encode a nucleoprotein (NP) gene in an overlapping frame with *rep*. In this report, we show that putative NP ORFs are present in ChPV-derived EVEs, suggesting it is an ancestral, conserved feature of these viruses. However, its absence from the coleopteran lineage is intriguing, as it is still present in the EPV of the hexapod stem group Diplura of Entognatha. Phylogenetic reconstructions (and the extensive overlap with *rep*) imply it was acquired ancestrally and independently lost in the lineage derived from members of the hexapod crown group, Coleoptera (Figure 4).

A functional role for auxiliary ORF1 is supported by: (i) its conservation across the entire amniote ChPV clade; and (ii) limited experimental data indicating it is expressed in MKPV via a spliced transcript. Auxiliary ORF2 was only identified in a small subset of ChPV genomes, but a functional role for this ORF is suggested by the presence of homologues in distantly related ChPVs of amniotes and fish (see Figure 4). Interestingly, although all ChPVs appear to express ORF1 via splicing of a small intronic sequence (Figure 1), those harboring an ORF2 homologue are predicted to lack the peculiar large introns found in the expression of MKPV NP and VP transcripts [21]. ScChPV lacks an ORF1 homologue, but contains a predicted reading frame in the corresponding position. Homologues of this ScChPV ORF1 variant are present in all three arachnid EPVs, although not in the first, but in the second position. As only the three *Latrodectus* EPVs possess a homologue of ORF1-Lh, it is possible that this small ORF originated after the split from the syngnathe fish lineage, whereas the ScChPV ORF1 originates earlier. The distribution of homologous auxiliary genes across phylogenetic lineages of ChPVs implies that distinct lineages have acquired and/or lost these genes on multiple, independent occasions.

MKPV has been reported to possess only one promoter and two polyadenylation signals, as well as an extensive number of spliced transcripts. This transcription pattern, however, appears to be unique to only one of the two hitherto amniote ChPV lineages, comprising of rodent, chiropteran, New World primate, avian, and reptilian entries. As members of the "type 1" lineage, including PPV7, appear to display a genome organization specific for this clade and different from that of MKPV, they may utilize distinct transcription strategies as well.

Despite the potential pitfalls of homology modeling, and the use of distinct templates to reconstruct both the VP monomer and capsid structures, we obtained remarkably similar predicted structures for VP sequences found in closely related viruses/EPVs. Since the viral capsid plays an important role in mediating the interactions between parvoviruses and their hosts, comparisons of capsid structures can potentially reveal insights into parvovirus biology. Our analysis indicates that ChPV VPs would assemble into a complete, $T = 1$ icosahedral capsid, despite their relatively small size. Furthermore, their predicted structures are remarkably similar to those found in other parvoviruses, despite the lack of any detectable similarity in the sequences of their VP proteins. Structural similarities include the presence of a conserved jelly roll core and α-A helix, the existence of the D–E and H–I loops, and the presence of identifiable VRs. Interestingly, the amniote ChPV capsids appear to possess the same number of VRs as most of the vertebrate parvoviruses of subfamily *Parvovirinae*, even if only a few of them (namely VRs 1, 2, and 9) proved to be analogous features. In these virus capsids, variations were most prominent among the three-fold peaks and protrusions, as well as the two-fold depression, as observed in members of the *Parvovirinae* (Figure 7). The tendency of some VRs to manifest at the luminal surface of the capsid in models suggests these regions could play a role in intracellular host–virus interactions. For these regions to become accessible to intracellular signaling pathways would require either uncoating or conformational changes. Based on previous findings, however, the parvovirus capsid appears to traffic into the nucleus intact [50,51]. Considering this, these buried regions might play a role in processes linked to the nucleus. Interestingly, bovine parvovirus, the only other parvovirus in which buried VRs have previously been observed [52] is an enteric pathogen, and the association of amniote ChPVs with fecal viromes suggests these viruses might also be largely enteric.

In addition to the VRs, all ChPVs seem to harbor highly variable VP C-terms. A similar phenomenon has been observed in the case of iteradensoviruses, in which the last 40 C-terminal residues are disordered, hence the structure of this region cannot be resolved [53]. Although the location of the ChPV C-term appears to vary, its association with regions that are overtly involved in parvovirus–host interactions (e.g., the two- and three-fold peaks) is certainly intriguing.

MKPV is associated with the pathology of the urogenital system, whereas a related virus, murine ChPV, has been detected at a very high prevalence in murine liver tissue, suggesting it is a gastrointestinal agent [45]. The VPs of the two, however, only differ in six aa residues, located within VR3 and near VR2 on the surface and in the buried VR4, as well as in the similarly buried variable C-term (Figure S3a). Thus, these positions could constitute potential determinants of tissue tropism in murine ChPVs.

Parvovirus subfamilies *Parvovirinae* and *Densovirinae* utilize distinct strategies to stabilize their icosahedral capsids [54]. Vertebrate parvoviruses extend the longer side of the jellyroll fold with an additional, N-terminal strand by folding back β-A to interact with the two-fold axis of the very same monomer, hence creating an extended ABDIG sheet [55,56]. By contrast, the densovirus capsid preserves the symmetric arrangement of the jellyroll fold, and possesses a β-A which is a direct elongated N-terminal extension of the β-B instead, interacting with the β-B strand of the neighboring monomer toward the five-fold axis [57,58]. Strikingly, our data show that ChPV capsids lack β-A strands (and also the β-B strand, in the case of ScChPV). The functional implications of this are unclear—possibly ChPV capsids are stabilized in the absence of β-A via a yet unknown, additional VP. If ChPVs express additional structural proteins, they are presumably encoded by spliced transcripts (given the unusually small size of the *cap* gene). Alternatively, the ChPV capsid might assemble without the incorporation of an additional β strand, perhaps at the cost of losing the stability and resilience typical of parvoviruses in general. Potentially, this could account for the apparent presence of buried VRs. Interestingly, in studies of MKPV, viral proteins could be detected in the kidneys of infected mice, even though no assembled particles could be observed in inclusion body-affected tubular cells [21]. This, along with our structural predictions, suggests that the ChPV strategy for uncoating and cellular trafficking might be very different from that found in the *Parvovirinae* and *Densovirinae*.

Uniquely, the genome of ScChPV appears to include a putative additional structural protein (ORF6), in addition to the above-mentioned alternative ORFs. All parvoviruses to date—except those of genus *Penstyldensovirus*, with only one VP comprising the capsid [58]—have been reported to incorporate up to three additional minor VPs into the virion, which share a common C-terminal region. To encode a structural protein on an entirely separate ORF sharing, no mutual coding sequence with *cap* would be unique. Possibly, this unusual feature could be connected to the predicted lack of a β-B strand in the ScChPV VP monomer.

Taken together, the data presented here establish that the ChPVs belong to a parvovirus lineage that comprises a distinct lineage from all other parvoviruses, and infects an exceptionally broad range of host species, including both vertebrates and invertebrates. Consistent with this, their relatively complex genomes exhibit numerous unique features, implying that their life cycle might significantly differ from what has been established in the case of other members of the family. These findings underscore the need for further basic and comparative studies of ChPVs, to assess their potential impact on animal health, both wildlife and livestock. Furthermore, this is the first study to imply that vertebrate parvoviruses are not monophyletic, and that members of the family must have evolved to infect vertebrates on at least two separate occasions.

Supplementary Materials: The following are available online at http://www.mdpi.com/1999-4915/11/6/525/s1, Figure S1: Predictions of secondary structure, disordered regions and potential phosphorylation sites in case of an amniote exogenous and an endogenous invertebrate ChPV nucleoprotein (NP), Figure S2: Secondary structure predictions of the *Syngnathus scovelli* ChPV genome termini. Figure S3: Variable regions at the derived amino acid sequence level identified among ChPV capsid proteins.

Author Contributions: Conceptualization, R.J.G. and J.J.P.; Methodology, R.J.G. and J.J.P.; Software, R.J.G.; Validation, J.J.P., R.J.G and M.A.-M.; Resources, R.J.G. and M.A.-M.; Data Curation, J.J.P., R.J.G and W.M.d.S.; Writing—Original Draft Preparation, J.J.P and R.J.G.; Writing—Review & Editing, M.A.-M.; Funding Acquisition, R.J.G. and M.A.-M.

Funding: J.J.P. and M.A.-M. are funded by NIH R01 GM109524. R.J.G. was funded by the Medical Research Council of the United Kingdom (MC_UU_12014/12). W.M.d.S. is supported by the Fundação de Amparo à Pesquisa do Estado de São Paulo, Brazil (Scholarships No. 17/13981-0).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cotmore, S.F.; Agbandje-McKenna, M.; Chiorini, J.A.; Mukha, D.V.; Pintel, D.J.; Qiu, J.; Soderlund-Venermo, M.; Tattersall, P.; Tijssen, P.; Gatherer, D.; et al. The family Parvoviridae. *Arch. Virol.* **2014**, *159*, 1239–1247. [CrossRef] [PubMed]
2. Tijssen, P.; Agbandje-McKenna, M.; Almendral, J.M.; Bergoin, M.; Flegel, T.W.; Hedman, K.; Kleinschmidt, J.; Li, Y.; Pintel, D.J.; Tattersall, P. Family Parvoviridae. In *Virus Taxonomy—Ninth Report of the International Committee on Taxonomy of Viruses*; King, A.M., Lefkowitz, E., Adams, M.J., Carstens, E.B., Eds.; Elsevier/Academic Press: London, UK, 2011; pp. 405–425.
3. Zádori, Z.; Szelei, J.; Tijssen, P. SAT: A late NS protein of porcine parvovirus. *J. Virol.* **2005**, *79*, 13129–13138. [CrossRef] [PubMed]
4. Sonntag, F.; Kother, K.; Schmidt, K.; Weghofer, M.; Raupp, C.; Nieto, K.; Kuck, A.; Gerlach, B.; Böttcher, B.; Müller, O.J.; et al. The assembly-activating protein promotes capsid assembly of different adeno-associated virus serotypes. *J. Virol.* **2011**, *85*, 12686–12697. [CrossRef] [PubMed]
5. Siqueira, J.D.; Ng, T.F.; Miller, M.; Li, L.; Deng, X.; Dodd, E.; Batac, F.; Delwart, E. Endemic infection of stranded southern sea otters (*Enhydra lutris* nereis). *J. Wildl. Dis.* **2017**, *53*, 532–542. [CrossRef] [PubMed]
6. Väisänen, E.; Fu, Y.; Hedman, K.; Söderlund-Venermo, M. Human protoparvoviruses. *Viruses* **2017**, *9*, 354. [CrossRef]
7. Geoghegan, J.L.; Pirotta, V.; Harvey, E.; Smith, A.; Buchmann, J.P.; Ostrowski, M.; Eden, J.-S.; Harcourt, R.; Holmes, E.C. Virological sampling of inaccessible wildlife with drones. *Viruses* **2018**, *10*, 300. [CrossRef]
8. De Souza, W.; Dennis, T.; Fumagalli, M.; Araujo, J.; Sabino-Santos, G.; Maia, F.; Acrani, G.; Carrasco, A.; Romeiro, M.; Modha, S.; et al. Novel parvoviruses from wild and domestic animals in Brazil provide new insights into parvovirus distribution and diversity. *Viruses* **2018**, *10*, 143. [CrossRef]
9. De Souza, W.M.; Romeiro, M.F.; Fumagalli, M.J.; Modha, S.; de Araujo, J.; Queiroz, L.H.; Durigon, E.L.; Figueiredo, L.T.M.; Murcia, P.R.; Gifford, R.J. Chapparvoviruses occur in at least three vertebrate classes and have a broad biogeographic distribution. *J. Gen. Virol.* **2017**, *98*, 225–229. [CrossRef]
10. Phan, T.G.; Gulland, F.; Simeone, C.; Deng, X.; Delwart, E. Sesavirus: Prototype of a new parvovirus genus in feces of a sea lion. *Virus Genes* **2015**, *50*, 134–146. [CrossRef]
11. Phan, T.G.; Dreno, B.; Da Costa, A.C.; Li, L.; Orlandi, P.; Deng, X.; Kapusinszky, B.; Siqueira, J.; Knol, A.-C.; Halary, F.; et al. A new protoparvovirus in human fecal samples and cutaneous T cell lymphomas (mycosis fungoides). *Virology* **2016**, *496*, 299–305. [CrossRef]
12. Belyi, V.A.; Levine, A.J.; Skalka, A.M. Sequences from ancestral single-stranded DNA viruses in vertebrate genomes: The parvoviridae and circoviridae are more than 40 to 50 million years old. *J. Virol.* **2010**, *84*, 12458–12462. [CrossRef] [PubMed]
13. Katzourakis, A.; Gifford, R.J. Endogenous viral elements in animal genomes. *PLoS Genet.* **2010**, *6*, e1001191. [CrossRef]
14. Liu, H.; Fu, Y.; Xie, J.; Cheng, J.; Ghabrial, S.A.; Li, G.; Peng, Y.; Yi, X.; Jiang, D. Widespread endogenization of densoviruses and parvoviruses in animal and human genomes. *J. Virol.* **2011**, *85*, 9863–9876. [CrossRef] [PubMed]
15. Reuter, G.; Boros, Á.; Delwart, E.; Pankovics, P. Novel circular single-stranded DNA virus from turkey faeces. *Arch. Virol.* **2014**, *159*, 2161–2164. [CrossRef] [PubMed]
16. Yang, S.; Liu, Z.; Wang, Y.; Li, W.; Fu, X.; Lin, Y.; Shen, Q.; Wang, X.; Wang, H.; Zhang, W. A novel rodent chapparvovirus in feces of wild rats. *Virol. J.* **2016**, *13*, 133. [CrossRef] [PubMed]
17. Kapoor, A.; Simmonds, P.; Lipkin, W.I. Discovery and characterization of mammalian endogenous parvoviruses. *J. Virol.* **2010**, *84*, 12628–12635. [CrossRef] [PubMed]

18. Holmes, E.C. The evolution of endogenous viral elements. *Cell Host Microbe* **2011**, *10*, 368–377. [CrossRef] [PubMed]
19. Baker, K.S.; Leggett, R.M.; Bexfield, N.H.; Alston, M.; Daly, G.; Todd, S.; Tachedjian, M.; Holmes, C.E.; Crameri, S.; Wang, L.-F.; et al. Metagenomic study of the viruses of African straw-coloured fruit bats: Detection of a chiropteran poxvirus and isolation of a novel adenovirus. *Virology* **2013**, *441*, 95–106. [CrossRef] [PubMed]
20. Palinski, R.M.; Mitra, N.; Hause, B.M. Discovery of a novel *Parvovirinae* virus, porcine parvovirus 7, by metagenomic sequencing of porcine rectal swabs. *Virus Genes* **2016**, *52*, 564–567. [CrossRef]
21. Roediger, B.; Lee, Q.; Tikoo, S.; Cobbin, J.C.; Henderson, J.M.; Jormakka, M.; O'Rourke, M.B.; Padula, M.P.; Pinello, N.; Henry, M.; et al. An atypical parvovirus drives chronic tubulointerstitial nephropathy and kidney fibrosis. *Cell* **2018**, *175*, 530–543. [CrossRef]
22. Zhu, H.; Dennis, T.; Hughes, J.; Gifford, R.J. Database-integrated genome screening (DIGS): Exploring genomes heuristically using sequence similarity search tools and a relational database. *bioRxiv* **2018**, 246835. [CrossRef]
23. Carver, T.; Harris, S.R.; Berriman, M.; Parkhill, J.; McQuillan, J.A. Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **2012**, *28*, 464–469. [CrossRef] [PubMed]
24. Knudsen, S. Promoter2.0: For the recognition of PolII promoter sequences. *Bioinformatics* **1999**, *15*, 356–361. [CrossRef] [PubMed]
25. Reese, M.G.; Eeckman, F.H.; Kulp, D.; Haussler, D. Improved splice site detection in genie. *J. Comput. Boil.* **1997**, *4*, 311–323. [CrossRef] [PubMed]
26. Dogan, R.I.; Getoor, L.; Wilbur, W.J.; Mount, S.M. SplicePort—An interactive splice-site analysis tool. *Nucl. Acids Res.* **2007**, *35* (Suppl. 2), W285–W291. [CrossRef]
27. Salamov, A.; Solovyev, V. Recognition of 3′ -processing sites of human mRNA precursors. *Bioinformatics* **1997**, *13*, 23–28. [CrossRef]
28. Armougom, F.; Moretti, S.; Poirot, O.; Audic, S.; Dumas, P.; Schaeli, B.; Keduas, V.; Notredame, C. Expresso: Automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Res.* **2006**, *34*, W604–W608. [CrossRef]
29. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797. [CrossRef]
30. Wallace, I.M.; O'Sullivan, O.; Higgins, D.G.; Notredame, C. M-Coffee: Combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.* **2006**, *34*, 1692–1699. [CrossRef]
31. Guindon, S.; Dufayard, J.F.; Lefort, V.; Anisimova, M.; Hordijk, W.; Gascuel, O. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Boil.* **2010**, *59*, 307–321. [CrossRef]
32. Felsenstein, J. *PHYLIP (Phylogeny Inference Package), Version 3.6*; Department of Genome Sciences, University of Washington: Seattle, WA, USA, 2005.
33. Lobley, A.; Sadowski, M.I.; Jones, D.T. pGenTHREADER and pDomTHREADER: New methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics* **2009**, *25*, 1761–1767. [CrossRef] [PubMed]
34. Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y. The I-TASSER suite: Protein structure and function prediction. *Nat. Methods* **2015**, *12*, 7–8. [CrossRef] [PubMed]
35. Carrillo-Tripp, M.; Shepherd, C.M.; A Borelli, I.; Venkataraman, S.; Lander, G.C.; Natarajan, P.; E Johnson, J.; Brooks, C.L.; Reddy, V.S. VIPERdb2: An enhanced and web API enabled relational database for structural virology. *Nucleic Acids Res.* **2009**, *37*, D436–D442. [CrossRef] [PubMed]
36. Schrödinger, L. *The PyMOL Molecular Graphics System, Version 2.0*; Wiley: Hoboken, NJ, USA, 2002.
37. Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Couch, G.S.; Greenblatt, D.M.; Meng, E.C.; Ferrin, T.E. UCSF Chimera? A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605–1612. [CrossRef] [PubMed]
38. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **2003**, *31*, 3406–3415. [CrossRef] [PubMed]
39. Gifford, R.J.; Blomberg, J.; Coffin, J.M.; Fan, H.; Heidmann, T.; Mayer, J.; Stoye, J.; Tristem, M.; Johnson, W.E. Nomenclature for endogenous retrovirus (ERV) loci. *Retrovirology* **2018**, *15*, 59. [CrossRef]

40. Schoonvaere, K.; Smagghe, G.; Francis, F.; De Graaf, D.C. Study of the metatranscriptome of eight social and solitary wild bee Species reveals novel viruses and bee parasites. *Front. Microbiol.* **2018**, *9*, 177. [CrossRef]
41. Chapman, M.S.; Agbandje-McKenna, M. Atomic structure of viral particles. In *Parvoviruses*; Kerr, J.R., Cotmore, F.C., Bloom, M.E., Linden, R.M., Parrish, C.R., Eds.; Hodder Arnold, Ltd.: London, UK, 2006; pp. 107–123.
42. Kailasan, S.; Agbandje-McKenna, M.; Parrish, C.R. Parvovirus family conundrum: What makes a killer? *Annu. Rev. Virol.* **2015**, *2*, 425–450. [CrossRef]
43. Pénzes, J.J.; Marsile-Medun, S.; Agbandje-McKenna, M.; Gifford, R.J. Endogenous amdoparvovirus-related elements reveal insights into the biology and evolution of vertebrate parvoviruses. *Virus Evol.* **2018**, *4*, vey026. [CrossRef]
44. Arriagada, G.; Gifford, R.J.; Beemon, K.L. Parvovirus-derived endogenous viral elements in two south American rodent genomes. *J. Virol.* **2014**, *88*, 12158–12162. [CrossRef]
45. Williams, S.H.; Che, X.; Garcia, J.A.; Klena, J.D.; Lee, B.; Muller, D.; Ulrich, W.; Corrigan, R.M.; Nichol, S.; Jain, K.; et al. Viral diversity of house mice in New York city. *mBio* **2018**, *9*, e01354-17. [CrossRef] [PubMed]
46. Deyle, D.R.; Russell, D.W. Adeno-associated virus vector integration. *Curr. Opin. Mol. Ther.* **2009**, *11*, 442–447. [PubMed]
47. Majumder, K.; Etingov, I.; Pintel, D.J. Protoparvovirus interactions with the cellular DNA damage response. *Viruses* **2017**, *9*, 323. [CrossRef] [PubMed]
48. Valencia-Herrera, I.; Cena-Ahumada, E.; Faunes, F.; Ibarra-Karmy, R.; Gifford, R.J.; Arriagada, G. Molecular properties and evolutionary origins of a parvovirus-derived myosin fusion gene in guinea pigs. *bioRxiv* **2019**, 572735. [CrossRef]
49. Willmann, R. Phylogenetic relationships and evolution of insects. In *Assembling the Tree of Life*; Cracraft, J., Donoghue, M.J., Eds.; Oxford University Press: Oxford, UK, 2004; pp. 330–344.
50. Cohen, S. Pushing the envelope: Microinjection of *Minute virus* of mice into *Xenopus* oocytes causes damage to the nuclear envelope. *J. Gen. Virol.* **2005**, *86*, 3243–3252. [CrossRef] [PubMed]
51. Sonntag, F.; Bleker, S.; Leuchs, B.; Fischer, R.; Kleinschmidt, J.A. Adeno-associated virus type 2 capsids with externalized VP1/VP2 trafficking domains are generated prior to passage through the cytoplasm and are maintained until Uncoating OCCURS in the nucleus. *J. Virol.* **2006**, *80*, 11040–11054. [CrossRef] [PubMed]
52. Kailasan, S.; Halder, S.; Gurda, B.; Bladek, H.; Chipman, P.R.; McKenna, R.; Brown, K.; Agbandje-McKenna, M. Structure of an enteric pathogen, bovine parvovirus. *J. Virol.* **2015**, *89*, 2603–2614. [CrossRef] [PubMed]
53. Kaufmann, B.; El-Far, M.; Plevka, P.; Bowman, V.D.; Li, Y.; Tijssen, P.; Rossmann, M.G. Structure of Bombyx mori Densovirus 1, a Silkworm Pathogen. *J. Virol.* **2011**, *85*, 4691–4697. [CrossRef] [PubMed]
54. Drouin, L.M.; Lins, B.; Janssen, M.; Bennett, A.; Chipman, P.; McKenna, R.; Chen, W.; Muzyczka, N.; Cardone, G.; Baker, T.S.; et al. Cryo-electron microscopy reconstruction and stability studies of the wild type and the R432A variant of adeno-associated virus type 2 reveal that capsid structural stability is a major factor in genome packaging. *J. Virol.* **2016**, *90*, 8542–8551. [CrossRef]
55. Simpson, A.A.; Hébert, B.; Sullivan, G.M.; Parrish, C.R.; Zádori, Z.; Tijssen, P.; Rossmann, M.G. The structure of porcine parvovirus: Comparison with related viruses. *J. Mol. Biol.* **2002**, *315*, 1189–1198. [CrossRef]
56. Xie, Q.; Bu, W.; Bhatia, S.; Hare, J.; Somasundaram, T.; Azzi, A.; Chapman, M.S. The atomic structure of adeno-associated virus (AAV-2), a vector for human gene therapy. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 10405–10410. [CrossRef] [PubMed]
57. Simpson, A.A.; Chipman, P.R.; Baker, T.S.; Tijssen, P.; Rossmann, M.G. The structure of an insect parvovirus (*Galleria mellonella* densovirus) at 3.7 å resolution. *Structure* **1998**, *6*, 1355–1367. [CrossRef]
58. Kaufmann, B.; Li, Y.; Szelei, J.; Tijssen, P.; Bowman, V.D.; Waddell, P.J.; Rossmann, M.G. Structure of Penaeus stylirostris densovirus, a shrimp pathogen. *J. Virol.* **2010**, *84*, 11289–11296. [CrossRef] [PubMed]