Long title (250 characters):

Comparative analysis reveals the long-term co-evolutionary history of parvoviruses and vertebrates.

Short title (70 characters):

Origin and evolution of vertebrate parvoviruses

Matthew A. Campbell*, Shannon Loncar*, Robert Kotin and Robert J. Gifford

*Equal contributions

Matthew A. Campbell: *University of Alaska Museum of the North, Fishes and Marine Invertebrates, 1962 Yukon Drive, Fairbanks, AK 99775 USA*

Shannon Loncar: *University of Massachusetts Medical School, Department of Microbiology and Physiological Systems, Gene Therapy Center, 55 Lake Ave. North, Worcester, MA 01655, USA*
Current address: Sanofi Genzyme

Robert Kotin: *University of Massachusetts Medical School, Department of Microbiology and Physiological Systems, Gene Therapy Center, 55 Lake Ave. North, Worcester, MA 01655, USA*
And Synteny Therapeutics, Cambridge, MA 02138

Robert J. Gifford: *MRC-University of Glasgow Centre for Virus Research, 464 Bearsden Rd, Bearsden, Glasgow, UK, G61 1QH*

Corresponding author: Robert J. Gifford: robert.gifford@glasgow.ac.uk

SUMMARY

Parvoviruses (family *Parvoviridae*) are small, non-enveloped DNA viruses that infect a broad range of animal species. Comparative studies, supported by experimental evidence, show that many vertebrate species contain sequences derived from ancient parvoviruses embedded in their genomes. These 'endogenous parvoviral elements' (EPVs), which arose via recombination-based mechanisms in infected germline cells of ancestral organisms, constitute a form of 'molecular fossil record' that can be used to investigate the origin and evolution of the parvovirus family. Here, we use comparative approaches to investigate 198 EPV loci, represented by 470 EPV sequences identified in a comprehensive *in silico* screen of 752 published vertebrate genomes. We investigated EPV loci by constructing an open resource that contains all of the data items required for comparative sequence analysis of parvoviruses and uses a relational database to represent the complex semantic relationships between them. We used this standardised framework to implement reproducible comparative phylogenetic analysis of combined EPV and virus data. Our analysis reveals that viruses closely related to contemporary parvoviruses have circulated among vertebrates since the Late Cretaceous epoch (100-66 million years ago). We present evidence that the subfamily *Parvovirinae*, which includes ten vertebrate-specific genera, has evolved in broad congruence with the emergence and diversification of major vertebrate groups. Furthermore, we infer defining aspects of evolution within individual parvovirus genera - mammalian vicariance for protoparvoviruses (genus *Protoparvovirus*), and inter-class transmission for dependoparvoviruses (genus *Dependoparvovirus*) - thereby establishing an ecological and evolutionary perspective through which to approach analysis of these virus groups. We also identify evidence of EPV expression at RNA level and show that EPV coding sequences have frequently been maintained during evolution, adding to a growing body of evidence that EPV loci have been co-opted or exapted by vertebrate species, and especially by mammals. Our findings offer fundamental insights into parvovirus evolution. In addition, we establish novel genomic resources that can advance the development of parvovirus-related research - including both therapeutics and disease prevention efforts - by enabling more efficient dissemination and utilisation of relevant, evolution-related domain knowledge.

INTRODUCTION

Parvoviruses (family *Parvoviridae*) are a diverse group of small, non-enveloped DNA viruses that infect a broad and phylogenetically diverse range of animal species [1, 2]. The family includes numerous important pathogens of humans and domesticated species, including erythroparvovirus B19 (fifth disease), carnivore protoparvovirus 1 (canine parvovirus) and carnivore amdoparvovirus 1 (Aleutian mink disease). Parvoviruses are also being developed as next-generation therapeutic tools - rodent protoparvoviruses (RoPVs) are promising anticancer agents that show natural oncotropism and oncolytic properties [3, 4], while adeno-associated virus (AAV), a non-autonomously replicating dependoparvovirus, has been successfully adapted as a gene therapy vector, and parvoviruses are leading candidates for the further development of human gene therapy [5, 6].

Parvoviruses have highly robust, icosahedral capsids (T=1) that contain a linear, single-stranded DNA genome typically ~5 kilobases (kb) in length. Parvovirus genomes are typically very compact and generally exhibit the same basic genetic organization comprising two major gene cassettes, one (Rep/NS) that encodes the non-structural proteins, and another (Cap/VP) that encodes the structural coat proteins of the virion [2]. However, some genera contain additional open reading frames (ORFs) adjacent to these genes or overlapping them in alternative reading frames. The genome is flanked at the 3' and 5' ends by palindromic inverted terminal repeat (ITR) sequences that are the only *cis* elements required for replication.

Recent years have seen many important advances in understanding of parvovirus evolution and diversity, driven primarily by dramatic increases in the availability of DNA sequence data and investments in deriving novel adeno-associated virus capsid for gene therapy applications. Metagenomic sequencing has enabled the discovery of numerous novel parvovirus species, which in turn has led to the taxonomic re-organization of the *Parvoviridae* to include additional subfamilies and genera [1]. In addition, whole genome DNA sequencing has revealed that DNA sequences derived from parvoviruses are widespread in animal genomes [7-13]. These endogenous parvoviral elements (EPVs) are thought to have arisen when parvovirus infection of germline cells (i.e., gametes, gamete producing cells, or early-stage embryos) led to integration of parvovirus-derived DNA into chromosomal DNA so that it was subsequently inherited as a newly acquired allele. Integration of parvovirus DNA can occur via cell-mediated, non-homologous recombination but may also be mediated by the activities of virus-encoded proteins [14, 15]. Comparative genomic studies have shown that EPV sequences often occur as orthologous loci in multiple related host species, demonstrating that they were incorporated into the germline of a common ancestor. Thus, species divergence times – which are in part based

on evidence from the fossil record - provide a robust method of deriving minimum age estimates for EVE insertions. Many EVEs represent virus lineages that have not been described previously and may be extinct [7, 16]. However, others clearly represent members of contemporary virus groups, and age estimates obtained for these EVEs provide insights into their long-term evolutionary history [7, 17-20].

The extent to which EPVs have reached fixation through positive selection - as opposed to incidental factors such as founder effects, population bottlenecks, and genetic hitchhiking - remains unclear. Potentially, EPV genetic information might sometimes be co-opted or "exapted" as has been reported for EVEs derived from other virus groups, including retroviruses (family *Retroviridae*) [21, 22] and polydnaviruses (family *Polydnaviridae*) [23]. Recent studies have revealed that two distinct, fixed EPVs in the germline of (i) the degu (*Octodon degus*) and (ii) family Elephantidae (elephants) – both of which encode an intact Rep protein ORF - exhibit similar patterns of tissue-specific expression in the liver [24, 25]. These observations suggest that expression of Rep protein or mRNA might – in some way - be physiologically relevant. More broadly, incorporation of parvovirus-derived DNA into animal germlines may provide a novel DNA substrate for the evolution of new genes – for example, guinea pigs (*Cavia porcellus*) encode a predicted polypeptide gene product comprising a partial myosin9-like (M9l) gene fused to a 3' truncated, EPV-encoded replicase [26].

Comparative genomic analysis can reveal key insights into the biology and evolution of viral species. EVE data are critically important components of these studies as they provide calibrations in geologic time. Unfortunately, making effective use of these data is challenging for a variety of reasons. This reflects a general lack of reproducibility and reusability in computational genomics [27], particularly where rapidly evolving and highly divergent sequences are involved [28, 29]. To address these issues we previously developed GLUE (Genes Linked by Underlying Evolution), a sequence data-centric bioinformatics environment computational genomics, with a focus on variation, evolution, and sequence interpretation [30]. Here, we used GLUE to create 'Parvovirus-GLUE', an extensible, open resource for comparative genomic analysis of parvovirus and EPV sequence data. We catalogue hundreds of EPV sequences in published whole genome sequence (WGS) data using a standardized nomenclature system to systematize the parvovirus fossil record. We capture these data in Parvovirus-GLUE and use reproducible approaches to examine their genomic and phylogenetic characteristics, revealing new insights into parvovirus ecology and evolution.

RESULTS

Creation of open resources for reproducible genomic analysis of parvoviruses

Comparative genomic analyses generally entail the construction of complex data sets comprising molecular sequence data linked to other kinds of information. Usually these include genome feature annotations and multiple sequence alignments (MSAs) as well as other diverse kinds of data. We used GLUE to create Parvovirus-GLUE [31], an open accessible online resource for comparative genomic analysis of parvoviruses and EPVs that preserves the 'state' of our data so that our analyses can be precisely and widely replicated (**Fig. S1a-b**). Furthermore, hosting of the Parvovirus-GLUE project in an openly accessible online version control system (GitHub) provides a platform for ongoing development of this resource by multiple collaborators, following practices established in the software industry (**Fig. S1c**).

The Parvovirus-GLUE project incorporates all of the data items required for broad comparative sequence analysis of parvoviruses, including: (i) a set of reference sequences representing all known parvovirus species (**Supplementary**); (ii) sequence and isolate-specific information (e.g. host species, vector species) in tabular form; (iii) a standardized set of parvovirus genome features and their coordinates within selected reference genome sequences; (iv) a set of MSAs incorporating all sequences in the project. Loading the project into the GLUE 'engine' generates a relational database that not only contains the data items associated with our analysis, but also represents the complex semantic links between them (**Fig. S1a**). Reproducible comparative genomic analysis of parvoviruses can be implemented by using GLUE's command layer to coordinate interactions between the Parvovirus-GLUE database and bioinformatics software tools. The resource can be installed on all commonly-used computing platforms, and is also fully containerised via Docker [32]. It can be used as a local, stand-alone tool or as a robust foundation for the development of genome analysis-based reporting tools for potential use in human and animal health (**Fig. S1b**) - e.g. see HCV-GLUE [30], RABV-GLUE [33].

Systematic recovery of the parvovirus 'fossil record'

We screened *in silico* WGS data of 738 vertebrate species and recovered a total of 595 EPV sequences. EPV sequences ranged from near full-length virus genomes to fragments ~150-300 nucleotides (nt) in length (**Fig. 1**). We classified EPVs into taxonomic groups based on their phylogenetic relationships to contemporary viruses (**Fig. 2**). Among EPVs that were >200nt in length, the majority (192/198=0.97%) were either unambiguous members of contemporary genera or members of groups that emerge as sister clades to these genera. We

used comparative approaches to resolve these sequences into sets of orthologs, revealing that they represent at least 198 distinct germline incorporation events (**Table 1, Tables S1-S6, Supplementary**). As part of these efforts, we investigated the genes flanking each putatively novel locus. We identified common sets of genes flanking most EPV loci (**Fig. 2**). We applied unique identifiers to all EPVs identified in our study using a systematic nomenclature system that was originally developed for endogenous retroviruses (ERVs) but has more recently been applied to EVEs [19, 20]. This nomenclature captures information about orthology, enforcing a higher level of order on our data set by unambiguously associating EPV sequences with genomic loci. The semantic links between these data items are recorded in the Parvovirus-GLUE database, so that they are available for retrieval and manipulation in computational analyses.

EPVs were identified in the genomes of all major groups of terrestrial vertebrates (**Table 1**) except agnathans, crocodiles, or amphibians. However, they occur much more frequently in mammals than in other groups. The majority of the EPVs identified in vertebrate WGS data derived from genera within the subfamily *Parvovirinae*, but we also identified rare examples of EPVs derived from subfamily *Hamaparvovirinae* (genus *Icthamaparvovirus*) (**Table 2**, **Fig. S5**). The identification of icthamaparvovirus-derived EPVs in snakes provides the first evidence that the host range of this viral genus extends to reptiles. Furthermore, orthologous copies of this EPV were identified in multiple snake species, providing a minimum age of 62 My for the *Ichthamaparvovirus* genus, and by extension the *Hamaparvovirinae* subfamily. Among *Parvovirinae*-derived EPVs, those derived from proto- and dependoparvoviruses predominate. Other genera (*Erythyroparvovirus*, *Amdoparvovirus* [34]) are also represented in the mammalian germline, but are relatively rare (<1% of species examined).

Previous studies have shown that some EPV loci express RNA with the potential to encode polypeptide gene products, either as unspliced viral RNA [12, 24, 25], or as fusion genes comprising RNA sequences derived from both host and viral sources [26]. We examined coding potential in EPVs and identified numerous sequences capable of encoding uninterrupted polypeptide sequences of 300 amino acids (aa) or more, with some ranging up to 722aa (**Table S7**). Furthermore, screening of RNA databases revealed evidence for expression of EPV RNA in several previously unreported EPVs) (**Supp. Doc 1).**

EPVs reveal the deep evolutionary origins of the subfamily Parvovirinae

We performed a comprehensive phylogenetic analysis of the *Parvoviridae*, using all available information, including EPVs. Phylogenies revealed three robustly supported sub-

lineages within subfamily *Parvovirinae*, each encompassing multiple genera, as follows: (i) "Amdo-Proto": *Amdo*- and *Protoparvovirus*; (ii) "Ave-Boca": *Ave*- and *Bocaparvovirus*; (iii) "ETDC": *Erythro*-, *Tetra*-, *Dependo*- and *Copiparvovirus* (**Fig. 2**).

Next, we compiled estimates of EPV integration dates (**Table S1-S6**) to provide an overview of parvovirus and vertebrate interactions over the past 100 My (**Fig. 3**). This revealed that germline incorporation of parvovirus DNA occurred throughout the Cenozoic Era in a broad range of vertebrate species (**Fig. 3**). Our results reveal that mammals acquired EPVs at a much higher frequency than other vertebrate groups (**Table 2**). In addition to the previously reported dependoparvovirus-derived elements in "whippomorphs" (cetaceans and hippopotamus) [13], lagomorphs, Old World rodents [7], New World rodents [24, 26], elephants [25], and macropoids [12] we identified numerous other ancient EPV loci diverse range of animal species (**Table S1, Fig 3**). Orthologous sets of EPV sequences demonstrating were also identified in passerine birds (order Passeriformes), establishing the ancestral presence of these viruses among ancestral members of clade Neoaves >85 Mya [35]. In addition, two ancient EPV loci were identified in snakes - an EPV in pit vipers provides a minimum age of 30 My for the Amdo-Proto lineage.

The parvovirus family likely has extremely ancient origins, perhaps dating back to the origin of animal species [11], and the independent formation and fixation of EPVs in such a diverse range of vertebrate groups demonstrates that *Parvovirinae* genera circulated widely among vertebrate fauna throughout the Cenozoic Era. Furthermore, since we know that transmission between distantly related host groups is rare, we can tentatively estimate the age of sublineages within the *Parvovirinae*, based on the assumption that they reflect broad codivergence of viruses and hosts – at least at higher taxonomic levels. For example, the "Ave-Boca" lineage comprises clearly distinct avian (Ave-) and mammalian (Boca-) lineages, suggesting that ancestral members of this virus lineage circulated among the common ancestors of birds and mammals >300 Mya. Furthermore, we identified EPVs in the genomes of cartilaginous and ray-finned fish, suggesting that the subfamily *Parvovirinae* may be as old, if not older, than the vertebrate lineage itself (**Fig. 2**).

Interestingly, roseoloviruses (genus *Roseolovirus*) – the group of betaherpesviruses (subfamily *Betaherpesvirinae*) that includes human herpesvirus 6 (HHV6) – have acquired a homolog of the parvovirus *rep* gene, called U94, presumably when infection of the same host cell led to parvovirus DNA being incorporated the 'germline' of an ancient betaherpesvirus [36]. The presence of U94 in an orthologous position in rodent and bat betaherpesviruses, as well as within a betaherpesvirus-derived EVE in the tarsier genome [37], demonstrates that it arose

through a horizontal gene transfer event that occurred ancestrally, likely before the divergence of (i) eutherian mammal orders and (ii) betaherpesvirus genera [38]. Phylogenetic reconstructions show that this gene derives from the EDTC sub-lineage within the subfamily *Parvovirinae* (**Fig. 2**).

Parvovirus genomes have palindromic ITR sequences at both the 3' and 5' ends which can fold back on themselves to form "hairpin" structures that are stabilized by intramolecular base-pairing. These "hairpin" structures are critical for genome replication in all parvoviruses, however, whereas they are heterotelomeric (asymmetrical) in some genera (*Amdo-, Proto-, Boca-*, and *Aveparvovirus*) they are homotelomeric (symmetrical) in others [39]. Interestingly, the distribution of this trait (where it has been described) across sub-lineages within the subfamily *Parvovirinae* suggests that - under the principles of maximum parsimony – the asymmetrical form (which is found across the “Amdo-Proto” and “Ave-Boca” sublineages) would be the ancestral form. Within the “ETDC” lineage, ITRs have only been described for the *Dependoparvovirus* and *Erythroparvovirus* genera, both of which have homotelomeric ITRs (**Fig 2c**), suggesting this that the presence of homotelomeric ITRs is a derived characteristic in subfamily *Parvovirinae*. Similarly, in all *Parvovirinae* groups except genus *Amdoparvovirus*, the N-terminal region of VP1 (the largest of the capsid) contains a phospholipase A2 (PLA2) enzymatic domain that becomes exposed at the particle surface during cell entry and is required for escape from the endosomal compartments. Phylogenetic reconstructions indicate that loss of PLA2 is an acquired characteristic of amdoparvoviruses (**Fig. 2**).

Another variable characteristic found in the *Parvovirinae* is the regulation of gene expression strategies, with members of the *Proto-* and *Dependoparvovirus* genera using two to three separate transcriptional promoters, whereas in the *Amdo-*, *Erythro-*, and *Boca-* genera all genes are expressed from a single promoter and genus-specific read-through mechanisms are used to produce alternative transcripts [2]. The presence of multiple separate promoters in the distantly related *Proto-* and *Dependoparvovirus* genera indicates that this expression strategy is probably ancestral, although the possibility that it evolved convergently in each lineage cannot be formally ruled out.

Mammalian vicariance shaped the evolution of protoparvoviruses

We identified 121 protoparvovirus-related EPV sequences in mammals, which we estimate to represent at least 105 distinct germline incorporation events (**Table S1**). Several near full-length genomes were identified, and many elements spanned >50% of the genome (**Fig 1a**). We reconstructed the evolutionary relationships between protoparvovirus-related

EPVs and contemporary protoparvoviruses, revealing three major subclades within the *Protoparvovirus* genus, which we labelled “Archaeo-”, “Meso-” and “Neo-” protoparvovirus) (**Fig 4a**). The Archaeoprotoparvovirus (ApPV) clade is comprised exclusively of EPVs and is highly represented in the genomes of Australian marsupials (Australidelphia), American marsupials (Ameridelphia) and New World rodents. It includes numerous elements that are near full-length, but none encoding intact open reading frames (ORFs).

The Mesoprotoparvovirus (MpPV) clade is also comprised exclusively of EPVs, and was sparsely represented in the EPV fossil record, being detected in the Southern tamandua (*Tamandua tetradactyla*) – a xenarthran – and the aardvark (*Orycteropus afer*). The EPV locus found in aardvarks is relatively degraded, but the tamandua EPV sequence is nearly full-length and relatively intact (**Table S7**) (**Fig. 1a**). Finally, the Neoprotoparvovirus (NpPV) clade contains EPVs along with all known contemporary protoparvoviruses (**Fig. 4a**). Of the four NpPV-derived EPVs we identified here, three have been reported previously [9, 40], and all were identified in rodents. The novel representative was identified in the steppe mouse (*Mus spicelagus*) and comprises a near complete genome (**Fig. 1**). Notably, the VP gene of this element groups robustly with a bat-derived virus [41] in phylogenetic trees (**Fig. S6a**), whereas those encoded by other NpPV-derived EPVs group separately, in an entirely different subclade, together with VP sequences derived from carnivore and porcine protoparvoviruses. Notably, phylogenetic reconstructions show that none of the rodent EPVs in the NpPV clade groups with contemporary RoPVs, but instead cluster robustly with pPVs found in other mammalian host groups (e.g., carnivores, artiodactyls). This suggests that horizontal transfer from rodents to other mammalian orders may have been a common feature of parvovirus evolution. Tusavirus, a divergent protoparvovirus of uncertain host origin [42] groups basally in the NpPV clade (**Fig. 4a**), but could potentially represent an entirely distinct, under-sampled pPV lineage.

Continental drift over the past 150-200 My is widely accepted to have had a dramatic impact on mammalian evolution [43]. Around 200 Mya, all continents were part of an interconnected landmass (Pangaea) that later separated into two subcomponents (**Fig. 4b**). One (Laurasia) comprised Europe and most of Asia, while the second (Gondwanaland) comprised Africa, South America, Australia, India and Madagascar). Mammalian subpopulations were fragmented by these events, and then fragmented further as Gondwanaland separated into its component continents. The associated genetic isolation due to geographic separation (vicariance) drove the early diversification of major subgroups, including indigenous mammalian lineages in South America (xenarthans and marsupials), Australia (marsupials), and Africa (afrotherians). At points throughout the Cenozoic Era, placental mammal groups that evolved in

Laurasia (Boreoeutherians) expanded into other continental regions. For example, the ancestors of contemporary New World rodents (which include capybaras, chinchillas, and guinea pigs among many other, highly diversified species), are thought to have reached the South American continent ~35 Mya [44].

The reconstructed evolutionary relationships between protoparvoviruses and protoparvovirus-derived EPVs strikingly reveal the impact of mammalian vicariance – and later migration – on the emergence and spread of novel protoparvovirus sublineages (**Fig. 4**). The protoparvovirus phylogeny can readily be mapped onto the phylogeny of mammalian host species so that the three major protoparvovirus lineages emerge in concert with major groups of mammalian hosts. These evolutionary relationships, which are supported by numerous, independently acquired EPV loci and ortholog sets (**Table S1**), are consistent with a parsimonious evolutionary scenario under which: (i) the ancestors of the contemporary protoparvovirus species were present in the ancient supercontinent of Pangaea prior to its breakup; (ii) the vicariance-driven, deep divergences in the mammalian phylogeny drove the emergence of distinct protoparvovirus lineages in distinct biogeographic regions throughout the course of the Cenozoic Era (from 65 Mya to present); (iii) the founder event associated with migration of rodents into the New World allowed this group to escape infection with NpPVs, but presumably, following their colonisation of the South American continent (estimated to have occurred ~50-30 Mya [44]), they were then exposed to infection with ApPVs, to the extent that numerous ApPV-derived EPVs were independently fixed in the germline. A previously reported ApPV-derived EPV in the common opossum (*Monodelphis domestica*) [10] groups intermediate between clades comprised of Australian marsupials EPVs and NW rodent EPVs, consistent with this hypothesis. Biogeographic analysis of host species distributions and ancestral range reconstruction support these findings (**Fig. S6b**).

Ancient origins and inter-order transmission of erythroparvoviruses

Our comprehensive *in silico* screen of vertebrate genomes identified the first reported examples of EPVs derived from genus *Erythroparvovirus*. One was identified in the genome of the Patagonian mara (*Dolichotis patagonum*) - a New World rodent – and another was identified in the genome of the Indri (*Indri indri*), a Malagasy primate. The mara element spans a complete NS gene, whereas the indri element encodes a complete viral genome with intact NS and VP genes and incorporating putative ITR sequences (**Fig. 2**), suggesting it integrated relatively recently. As reported in other erPVs, the viral protein 1 unique region (VP1u) of Erythro.1-Indri is relatively long. Neither contained obvious homologs of the accessory proteins reported in

contemporary erPV genomes. Both erPV-derived EPVs grouped with erPVs derived isolated from rodents in phylogenetic trees, indicating inter-order transmission from rodents to lemuriforme primates (**Fig. S7**). Furthermore, when examined in relation to the biogeographic distribution of host species, these phylogenetic relationships provide tentative age calibrations for the *Erythroparvovirus* genus. based on the parsimonious assumption that the presence of the EPVs derived from rodent erPVs in Madagascar and South America reflects their spread into these isolated geographic regions during the Cenozoic Era (**Table 3**).

Inter-class transmission and the evolution of non-autonomous dependoparvoviruses

We identified 213 dependoparvovirus-related EPV sequences in mammals, which we estimate to represent at least 80 distinct germline incorporation events (**Table S3**). A small number of near full-length genomes were identified, but a large share of these elements spanned only small fragments (i.e. >40%) of the dependoparvovirus (dPV) genome (**Fig 1b**). We reconstructed the evolutionary relationships between dPV-related EPVs and contemporary dPVs (**Fig 5**, **Fig. S9**). In trees that included EPVs, support for internal branching order was typically quite low. This reflects the short length of many dPV-related EPV sequences, and the fact that many parts of the viral genome sequence are relatively degraded [13]. However, when only viruses and longer EPV sequences are included, phylogenies based on *rep* gene sequences disclose several robustly supported subclades within the *Dependoparvovirus* genus (**Fig. 5**). They include clades exclusive to reptilian species (Sauria-), Australian marsupials (Oceania-), and Boreoeutherian mammals (Neo-). A fourth clade, which we named "Shirdaldependoparvovirus" (ShdPV), contains dPV taxa derived from both avian and mammalian hosts. Both the composition of this clade in terms of hosts, and its phylogenetic position relative to other dPV groups, implies a role for interclass transmission between mammals and birds in dPV evolution (**Fig. 5b**). Firstly, the avian viruses in this clade group basally, forming a paraphyletic group relative to a derived subclade - here referred to as 'Lemuria-' - of ancient EPVs obtained from a diverse range of mammalian hosts. This topology suggests that clade Lemuria- may have originated via transfer from birds to mammals. Furthermore, in both midpoint-rooted phylogenies, and in phylogenies rooted on the saurian dependoparvoviruses (SdPVs) (as proposed by Penzes *et al* [45]), the ShdPVs as a whole fall intermediate between two exclusively mammalian groups – the neodependoparvoviruses (NdPVs) found in placental mammals, and the Oceaniadependoparvoviruses (OdPVs) found in Australian marsupials (**Fig. 5a**). This suggests the ShPVs originated via transmission from mammals to birds.

The NdPVs include the non-autonomous parvoviruses (AAVs), which require a helper virus for replication typically a nuclear DNA virus (e.g. herpesvirus, adenovirus [1]). In phylogenies rooted on the RdPVs, the NdPVs emerge as a derived clade with the autonomously replicating avian viruses grouping basal. Fragmentary EPVs found in Cercopithecine primate genomes arose between 23-16 Mya and appear to represent the ancient progenitors of contemporary primate AAVs (**Fig. 5a**).

DISCUSSION

In this study we recovered the complete repertoire of EPV sequences in WGS data representing 738 vertebrate species. While previous studies have reported a sampling of EPV diversity in vertebrates [7-13, 24, 26, 34, 35, 40, 46], the present study is an order of magnitude larger in scale – we identify 595 sequences representing nearly 200 discrete germline incorporation events (**Table S1, Fig 1**). Furthermore, we introduced a higher level of order to these data by: (i) discriminating between unique loci and orthologous copies; (ii) hierarchically arranging MSAs so that phylogenetic analysis (and taxonomic classification) of individual EPVs could utilise the maximum amount of available data; (iii) applying to EPVs a standardised nomenclature that captures information about orthology and taxonomy; (iv) inferring ancestral reference sequences for EPV coding domains.

The EPVs reported here are derived from a diverse array of distinct parvovirus groups (**Fig. 2**). The majority grouped within subfamily *Parvovirinae*, but we also identify rare examples of EPVs derived from *Icthamaparvovirus* (a genus in subfamily *Hamaparvovirinae*) in snakes. Orthologous copies of this element demonstrate that the association between hamaparvoviruses and vertebrates extends to the late Mesozoic Era >100 Mya, reinforces the view that this recently described subfamily is ancient and broadly distributed [46, 47]. Among EPVs derived from subfamily *Parvovirinae*, the majority derived from two genera – *Protoparvovirus* and *Dependoparvovirus*. However, we also identified representatives of other genera (*Amdoparvovirus*, *Erythroparvovirus*) as well as several highly divergent EPVs that likely represent novel genera. Among these sequences, those that were identified in mammals may simply represent mutationally degraded members of the established genera, since they are ancient and relatively short (**Table 3**, **Fig 1f)**. However, those identified in basal vertebrate lineages such as cartilaginous fish (class Chondrichthyes) and lobe-finned fish (clade Sarcopterygii) are likely to represent novel groups. These EPV sequences also demonstrate that the host range of the subfamily *Parvovirinae* extends to basal vertebrates.

We obtained robust minimum age calibrations based on the identification of orthologous genomic flanking sequences for all parvovirus genera represented in the viral fossil record except *Erythroparvovirus*. However, for erythroparvoviruses (erPVs) and many other *Parvovirinae* genera (including those that are not represent in the molecular fossil record) we could infer more tentative calibrations based on the distribution of EPVs and viruses across host groups (**Table 3**). The most striking example of this occurs in the *Protoparvovirus* genus, in which the impact of biogeographic isolation throughout the Cenozoic Era is strongly reflected in the phylogenetic relationships between virus subgroups and the distribution of virus subgroups across host taxonomic groups and biogeographic host ranges. A simple, parsimonious explanation of these relationships is presented in **Fig. 4**, wherein ancestral protoparvoviruses (pPVs) were present in mammalian ancestors inhabiting Pangea ~200 Mya, and distinct pPV lineages emerged as mammalian species were compartmentalised into distinct biogeographic regions by continental drift. Later, the migration of mammalian groups into previously isolated continental regions provided the opportunity for these pPV subgroups to infect new host groups. Thus, the ancient AdPV lineage, which evolved primarily in marsupials spread into placental mammal group (New World rodents) during the later Cenozoic Era (see **Fig. 4**). This extended evolutionary timeline for pPVs is supported by evidence from orthology (**Table 3**), lending credibility to similar, biogeography and distribution-based age estimated inferred for viral lineages in which we did not obtain minimum age estimates based on orthologous EPVs (e.g., the Ave-Boca lineage).

Whereas some genera, such as *Protoparvovirus* and *Dependoparvovirus*, are highly represented in the genomic 'fossil record', others are conspicuously absent. For example, no EPVs derived from the 'Ave-Boca' lineage, or from the *Tetraparvovirus* and *Copiparvovirus* genera, were identified. However, the ancient calibrations obtained for dPVs and pPVs imply that other *Parvovirinae* genera have similarly ancient origins, and thus are consistent with the avian and mammalian components of the Ave-Boca lineage emerging via broad codivergence with vertebrate hosts ~400-300 Mya (**Table 3**). Extending this logic, the identification of *Parvovirinae* lineages in basal vertebrate lineages such as fish suggests that the subfamily *Parvovirinae* may have primordial origins within vertebrates.

While inter-class transmission of parvoviruses appears to be rare overall, we obtained compelling evidence that it has occurred in the *Dependoparvovirus* genus, specifically in the evolution of a lineage that contains parvoviruses and EPVs derived from both avian and mammalian hosts, and which we named "Shirdaldependoparvovirus" (ShDPV). This robustly supported clade contains both the avian autonomous dependoparvoviruses (dPVs) and the

lemuriadependoparvoviruses (LdPVs) – a clade of mammalian dPVs that existed >80 million years ago (**Table 3**) and is so far only represented by EPVs. The topology of NS/Rep phylogenies cannot be reconciled with codivergence and instead implies that both ShDPV and the LdPV subclade it contains arose in separate inter-class transmission events involving mammals and birds (**Fig. 5**).

The non-autonomous dependoparvoviruses – often referred to as "adeno-associated viruses" (AAVs) – are characterised by the requirement for a helper virus to replicate. All of these viruses group within the 'neodependoparvovirus' (NdPV) clade in our analysis. Most recently described AAVs – such as those identified in bats, rodents and carnivores - have only been characterised at sequence-level, and little is known about their phenotypic properties. However, most of these viruses fall within the range of diversity defined by two AAV groups (Dependo-A and Dependo-B) indicating that the requirement for a helper virus ("dependency") is an ancestral characteristic of the NdPV and likely to be shared among most if not all AAV species. Furthermore, the EPV fossil record supports the view that the host range of NdPVs encompasses all placental mammals. Dependency may have evolved as a means of timing replication - some large DNA viruses, such as herpesviruses, are able to establish latent, persistent infections within which they can 'sense' the cellular environment and switch to replicative mode when conditions are optimal [48]. The success of this strategy is reflected in the extremely high prevalence of herpesviruses in mammalian populations (often close to 100%). Possibly, NdPVs can optimise their transmission by tethering their replication cycle to that of these ubiquitous, sophisticated DNA viruses.

It seems extraordinary that so many EPVs have been fixed in the mammalian germline, since the formation of a novel EPVs is almost certainly a rare event - *in utero* virus infections are often lethal and virus-infected gamete cells are unlikely to be viable under most circumstances. Furthermore, the neutrality principle of population genetics predicts the loss of new alleles occurring at low frequency (unless there are selective advantages from the genotype). In most EPVs, ORFs have been disrupted by indels and contain multiple nonsense mutations rendering the ancestral viral ORFs non-translatable, but some retain long regions of intact coding sequence – both NS/Rep and VP/Capsid sequences are among the longest open regions (**Table S7**). The Rep protein is structurally and functionally related to the rolling circle replication (RCR) proteins that are among the oldest replicator proteins known [49, 50]. The RCR proteins play a pivotal role for replication of both circular and linear genomes, and therefore, are inextricably linked with single-stranded DNA viruses. With the exception of the mitochondrial DNA polymerase, RCR proteins are restricted to microbial and viral species. Experimental

studies have shown that dPV Rep protein (over) expression affects healthy cells through a variety of activities including DNA binding, constitutive ATPase, and inhibiting the (cyclic) cAMP -activated protein kinase A (PKA) and protein kinase X (PrKX) [51, 52]. Rep-mediated inhibition of these kinases not only affects the infected cell, but also diminishes the proliferation of adenovirus helper virus, perhaps attenuating the virulence and virus-induced pathogenesis [53, 54]. Conceivably, it could be these properties that have favoured their capture by herpesviruses and by host species genomes. The selective forces that have favoured the retention of open VP/capsid genes in some EPVs (see **Table S7**) are unclear.

The extended evolutionary timescale implied by our analysis raises interesting questions about parvovirus evolution. For example, all members of the subfamily *Parvovirinae* use similar basic mechanisms to achieve specific steps in infection, but the specific details of these processes differ between genera. Our study suggests that these differences could have evolved gradually as distinct parvovirus lineages adapted to distinct ecological niches. It is clear from phylogenetic and genomic analysis that most vertebrate species are infected with multiple distinct parvovirus groups - for example, at least seven distinct genera circulate in mammals. Has each parvovirus genus developed specializations that allow it to occupy a unique ecological niche, or are some or all parvoviruses generalists? Other questions concern the current distribution of parvoviruses – e.g., to what extent does it reflect long-term evolutionary processes versus (possibly) more recent, anthropogenic influences? Also, which parvovirus groups currently only known via EPVs e.g. (**Fig. 4**, **Fig. 5**) are still prevalent, and which are extinct? Our study shows that parvovirus host-associations are relatively stable over time, implying that further sampling of parvovirus and EPV diversity will help address these questions.

EPVs allow the present-day biological properties of parvoviruses to be examined in the light of a time-calibrated evolutionary history. They can also be used to investigate the structural properties of ancient capsids based on molecular modelling [34] and even to reconstruct viable versions of capsids from extinct paleoviruses [40]. However, making effective use of EPV data is often challenging, since high levels of sequence divergence preclude straightforward analysis. In this study we introduced a template for computational genomics studies of viruses that focusses on facilitating the reproduction of comparative analyses and re-use of the complex datasets that underpin them (e.g., MSAs) (**Fig. S1**). This approach can not only scale to accommodate greatly increased quantities of virus species and sequences, but also introduces new levels of reproducibility and re-usability so that researchers working in different areas of parvovirus genomics - but utilizing related data and domain knowledge - can benefit from one another's work. The resources and approaches developed in this study can thus facilitate the

development of a broader understanding of parvovirus biology, covering multiple biological scales, which can be used to mitigate their harmful impacts and inform their development as therapeutic tools.

METHODS

Creation of resources for reproducible comparative analysis of parvovirus genomes

We used the GLUE software environment [30] to create a sequence-oriented resource for comparative genomic analysis of parvoviruses (including extinct paleoviruses). This resource, called 'Parvovirus-GLUE', not only contains the data items required for comparative analysis (i.e., virus genome sequences, multiple sequence alignments (MSAs), genome feature annotations, and other sequence-associated data), it also represents the semantic relationships between these data items via a relational database (**Fig. S1**). A library of parvovirus reference sequences was compiled from the virus reference genomes resource [55] and from recent papers describing novel parvoviruses and parvovirus-derived EVEs (identified via PubMed search). We obtained reference genome sequences for all Parvovirus species recognised by the International Committee for Taxonomy of Viruses (ICTV), as well as a representative set of recently identified parvovirus-related viruses that have not yet been incorporated into official taxonomy. The database is constructed using GLUE's native command layer (**Fig. S3a**). The command layer can be used to interact with the database and with bioinformatics software tools required for comparative sequence analysis, thus establishing a platform for implementing comparative analyses in a reproducible, standardised way.

Parvovirus-GLUE incorporates MSAs representing distinct taxonomic levels within the family *Parvoviridae* and uses a 'constrained alignment tree' data structure to represent the hierarchical relationships between them (**Fig. S2, Table 1**). The root alignment contains all reference sequences, whereas all children of the root inherit at least one reference from their immediate parent. Thus, all alignments are linked to one another via our chosen set of references. This allows a single unified alignment to be used for phylogenetic analyses across distinct taxonomic levels and enables standardised sequence comparisons across the entire parvovirus family. A set of 'master' reference sequences - each representing a distinct clade in the parvovirus phylogeny – was defined. Reference sequences were used to define 'constrained' alignments (i.e., alignments in which the genomic coordinate spaces are constrained to the chosen reference sequence). For the lower taxonomic levels (i.e., genus and below) we aligned complete coding sequences. For the root we aligned a conserved region of

NS protein consistent with the approach used by ICTV [1]. We used GLUE's 'constrained alignment tree' data structure [30] to link MSAs constructed at distinct taxonomic levels, via a set of common reference sequences.

Genome screening in silico

We used the Database-Integrated Genome Screening (DIGS) tool [56] to derive a non-redundant database of EPV loci within published WGS assemblies. The DIGS tool is a Perl-based framework that uses the Basic Local Alignment Search Tool (BLAST) program suite [57] to perform similarity searches and the MySQL relational database management system to coordinate screening and record output data. A user-defined reference sequence library provides (i) a source of 'probes' for searching WGS data using the tBLASTn program [57, and (ii) a means of classifying DNA sequences recovered via screening **Fig. S4**. For the purposes of the present project, we collated a reference library composed of polypeptide sequences derived from representative parvovirus species and previously characterised EPVs. Whole genome sequence (WGS) data of animal species were obtained from the National Center for Biotechnology Information (NCBI) genome database [58]. We obtained all animal genomes available as of March 2020. We extended the core schema of the screening database to incorporate additional tables representing the taxonomic classifications of viruses, EPVs and host species included in our study. This allowed us to interrogate the database by filtering sequences based on properties such as similarity to reference sequences, taxonomy of the closest related reference sequence, and taxonomic distribution of related sequences across hosts. Using this approach, we categorised sequences into: (i) putatively novel EPV elements; (ii) orthologs of previously characterised EPVs (e.g., copies containing large indels); (iii) non-viral sequences that cross-matched to parvovirus probes (e.g., retrotransposons). Sequences that did not match to previously reported EPVs were further investigated by incorporating them into genus-level, genome-length MSAs (see **Table 1**) with representative parvovirus genomes and reconstructing maximum likelihood phylogenies using RAxML (version 8) [59].

Where phylogenetic analysis supported the existence of a novel EPV insertion, we also attempted to: (i) determine its genomic location relative to annotated genes in reference genomes; and (ii) identify and align EPV-host genome junctions and pre-integration insertion sites (see below). Where these investigations revealed new information (e.g., by confirming the presence of a previously uncharacterised EPV insertion) we updated our reference library accordingly. This in turn allowed us to reclassify putative EPV loci in our database and group sequences more accurately into categories. By iterating this procedure, we progressively

resolved the majority of EPV sequences identified in our screen into groups of orthologous sequences derived from the same initial germline incorporation event (**Table S1-S6**).

We applied standard identifiers (IDs) to all EPV loci, following a convention established for endogenous retroviruses [60] that has more recently been applied to EVEs [19, 20]. Each EVE is assigned a unique identifier (ID) constructed from two components. The first component is the classifier 'EPV'. The second component is itself a composite of two distinct subcomponents separated by a period; (i) the name of the lowest level taxonomic group (i.e., species, genus, subfamily, or other clade) into which the element can be confidently placed by phylogenetic analysis; (ii) a numeric ID that uniquely identifies the insertion.

Phylogenetic and Phylogeographic analysis

A process for reconstructing evolutionary relationships across the entire *Parvoviridae* was implemented using GLUE. We used a data structure called a 'constrained MSA tree' to coordinate genomic analyses across the large phylogenetic distances found in parvoviruses **Fig. S7d**. This approach addresses the issue that MSAs constructed at higher taxonomic levels (e.g., above genus-level in the *Parvoviridae*) typically contain far fewer columns than those constructed at low taxonomic levels (e.g., genus, subgenus), meaning that to fully investigate phylogenetic relationships using all available information it is often necessary to construct several separate MSAs each representing a distinct taxonomic grouping (e.g., see **Table 2**). The difficulties encountered in maintaining these MSAs in sync with one another while avoiding irreversible data loss are a key factor underlying the low levels of reproducibility and re-use in comparative genomic analyses [27], particularly those that examine more distant evolutionary relationships genomic [29]. To address these issues the 'constrained MSA tree' data structure represents the hierarchical links between MSAs constructed at distinct taxonomic levels, creating in effect a single, unified MSA that can be used to reconstruct both shallow and deep evolutionary relationships while making use of the maximum amount of available information at each level (**Table 2, Fig. S1d**). This approach also has the effect of standardising the genomic coordinate space to the constraining reference sequences selected for each MSA without imposing any limitations on which references are used (e.g., laboratory strains versus wild-type references), since additional, alternative constraining references can be incorporated, and contingencies such as insertions relative to the constraining reference are dealt with in a standardised way [30].

MSAs partitions derived from the constrained MSA tree were used as input for phylogenetic reconstructions. Nucleotide and protein phylogenies were reconstructed using

maximum likelihood (ML) as implemented in RAxML (version 8.2.12) [59]. To handle coverage-related issues we generated gene coverage data prior to phylogenetic analysis and used this information to condition the way in which taxa are selected into MSA partitions. Protein substitution models were selected via hierarchical maximum likelihood ratio test using the PROTAUTOGAMMA option in RAxML. For multicopy EPV lineages we constructed MSAs and phylogenetic trees to confirm that branching relationships follow those of host species (**Fig S4b**, [31]).

Time-calibrated vertebrate phylogenies were obtained via TimeTree [61]. We used a time-calibrated phylogeny of protoparvovirus host species and the present continental distribution of host organisms to model ancestral biogeographical range of protoparvovirus host [62] (**Fig. 6b**). Country-level distribution information for each host species was obtained via the occ_search function of the rgbif library in R [63]. Country records were consolidated in continent entries with the continents function of the countrycode library and manually curated to ensure accuracy. Within continents, North Africa and Sub-Saharan Africa were considered distinct distributions and coded separately. The Dispersal-Extinction-Cladogenesis (DEC) model implemented in the program Lagrange C++ was applied without constraining the number of ancestral states nor limiting connectivity between biogeographic units [64]. Ancestral states at all nodes in the tree were inferred and the tree visualized in R with the ggtree and ggplot libraries [65, 66]. Input data and configuration files for Lagrange along with the time tree and Lagrange output are provided in the Data Supplement [31].

Genomic analysis of EPVs

ORFs were inferred by manual comparison of sequences to those of reference viruses. For phylogenetic analysis, the putative peptide sequences of EVEs (i.e., the virtually translated sequences of EVE ORFs, repaired to remove frameshifting indels) were aligned with polypeptide sequences encoded by reference genomes. We used PAL2NAL [67] to generate in-frame, DNA alignments of virus coding domains from alignments of polypeptide gene products. Phylogenies were reconstructed using maximum likelihood (ML) as implemented in RAxML [59] and GTR model of nucleotide selection as selected using the likelihood ratio test. The putative peptide sequences of EPVs were aligned with NS and VP polypeptides of representative exogenous parvoviruses using MUSCLE.

Expression and intactness of EPVs

We identified open coding regions of coding sequence in EPVs by using PERL scripts, (included with Parvovirus-GLUE [31]) to process EPV sequence data. To determine if there was evidence of expression of EPVs in host species, we searched the NCBI Reference RNA Sequences (refseq_rna) with Dependoparvovirus VP and Rep sequences (NC_002077). We searched a translated nucleotide query and a translated database using tBLASTx [57] and evaluated alignments found between refseq_rna sequences and Dependoparvovirus VP and Rep sequences. To further verify expression, we determined if the annotations were solely based on computational prediction or if there is RNAseq data alignment to the annotation in support of the feature. For those host species with evidence of expression, we conducted blastn searches within refseq_rna to identify expressed EPVs.

Table 1. Multiple sequence alignments included in Parvoviridae-GLUE

#	Scope/Name	Parent	Children	Constraining reference	Coverage*	Viruses		EVE loci
	Root							
1	*Parvoviridae*	none	3	CPV	NS (13%)	3		0
	Subfamily							
2	*Parvovirinae*	*Parvoviridae*	2	CPV	NS (63%)	13		**4**
3	*Hamaparvovirinae*	*Parvoviridae*	2	PPV7	NS	5		0
4	*Densoparvovirinae*	*Parvoviridae*	0	JcDNV	NS	9	1*	0
	Cross-genus							
5	Boca-Ave	*Parvovirinae*	2	ChPV	Genome (70%)	2		0
6	Amdo-Proto	*Parvovirinae*	2	CPV	Genome (77%)	2		0
7	EDCT	*Parvovirinae*	3	HPV4	Genome (57%)	4		0
8	Chaphama-Icthama	*Hamaparvovirinae*	2	PPV7	Genome	2		0
	Genus							
9	*Aveparvovirus*	Boca-Ave	0	ChPV	Genome (88%)	4		0
10	*Bocaparvovirus*	Boca-Ave	0	BPV	Genome (75%)	7		0
11	*Erythyroparvovirus*	EDCT	0	B19	Genome (80%)	9	2*	**2**
12	*Tetraparvovirus*	EDCT	0	HPV4	Genome (80%)	10		0
13	*Dependoparvovirus*	EDCT	0	AAV2	Genome (84%)	27		**81**
14	*Copiparvovirus*	EDCT	0	BPV2	Genome (62%)	2		0
15	*Amdoparvovirus*	Amdo-Proto	0	AMDV	Genome (85%)	6		**6**
16	*Protoparvovirus*	Amdo-Proto	0	CPV	Genome (90%)	18		**106**
17	*Chaphamaparvovirus*	*Hamaparvovirinae*	0	PPV7	Genome (85%)	12	4**	**0**
18	*Icthamaparvovirus*	*Hamaparvovirinae*	0	SyIPV	Genome (62%)	2		**2**
	Totals					**137**		**201**

Footnote: Linking alignments that represent internal nodes contain only the reference sequences for their 'child' alignments in the constrained alignment tree. *Putatively exogenous parvoviruses identified in this study. **Putatively exogenous parvoviruses identified in previous studies. Abbreviations: EDCT=Erythyro-Dependo-Copi-Tetra group; CPV=canine parvovirus; JcDNV=*Junonia coenia* densovirus; PPV7=porcine parvovirus 7; ChPV=chicken parvovirus; HPV4=human parvovirus 4; B19=Human erythroparvovirus B19; BPV=bovine parvovirus; AMDV=Aleutian mink disease virus; *Syngnathus scovelli* ichthamapaparvovirus.

Table 2. Incorporation of parvovirus DNA into the vertebrate germline

Parvovirus genus	**Host species group**									
	Chondrichthyes *species=5*		Actinopterygii *species =175*		Sauria *species =200*		Mammalia *species=353*		Vertebrata *species =752**	
	loci	***ratio****	***loci***	***ratio***	***loci***	***ratio***	***loci***	***ratio***	***loci***	***ratio***
Ichthamaparvovirus	0	-	1	0.01	2	0.01	0	-	2	0.003
Erythroparvovirus	0	-	0	-	0	-	2	0.01	2	0.003
Amdoparvovirus	0	-	0	-	1	0.01	2	0.01	3	0.004
Dependoparvovirus	0	-	0	-	13	0.07	66	0.19	80	0.108
Protoparvovirus	0	-	0	-	0	-	105	0.3	105	0.142
New clades	1	*0.2*	2	0.01	0	-	3	0.01	6	0.008
Totals	1	*0.2*	3	0.02	16	0.08	178	0.5	198	0.263

Footnote: *Agnathans (n=3), Amphibians (n=15), and Lungfish (*Latimeria chalumnae*) were also screened but results are not shown since no species in these groups were found to have EPVs. ***** Ratio = unique loci / genomes screened.

Table 3. Calibrations of parvovirus evolution

Parvovirus lineage	Host species lineage(s)	High	Low
Ortholog-based*			
Primate AAVs (Dependo-A)	OW Primates	23	16
Neodependo-	Glires	88	76
Neodependo-	Lagomorpha	77	23
Neodependo-	Vespertilionidae	49	38
Neodependo-	Elephantidae	23	9
Neodependo-	Eulemur	9	6
Neodependo-	Hyracoidea	14	7
Lemuriadependo-	Whippomorpha	56	52
Lemuriadependo-	Rhinocerotidae	51	15
Lemuriadependo-	Phyllostomidae	39	35
Oceaniadependo-	Macropus	45	27
Amdo-	Serpentes	111	100
Amdo-	Hyracoidea	14	7
Proto-	Afrotheria		
Proto-	Marsupials		
EDTC	Laurasitheria	84	73
Ichthama	Serpentes	74	49
Archaeo-Proto-	Ctenomyidae-Octodontidae	24	16
Codivergence based			
Proto-Amdo lineage	Sharks/Bony fish		
Neoprotoparvovovirus	Eutherian mammals	200	150
Boca-Ave lineage	Aves/Mammalia	393	294
Copiparvovirus	Eutherian mammals	111	100
Tetraparvovirus	Eutherian mammals	111	100
Erythroparvovirus	Eutherian mammals	111	100
Dependoparvovirus	Euteleostii	446	425
Biogeography-linked			
Archeoproto-	Pangea	200	180
Archeoproto- NW rodent clade	NW Rodents S. America	50	30
Erythroparvo- Rodent clade	Malagasy rodents Madagascar	30	20
U94 gene transfer			
EDTC lineage	Betaherpesviruses	200	80

Footnote: *Not all ortholog-based calibrations are shown, only the oldest for each virus lineage in which we identified orthologous sets of EPV sequences. Complete records of ortholog-based dates can be found in **tables S1-S6**.

Figure 1. Genomic structures of unique EPV loci.

(a) Protoparvovirus-derived EPV loci shown relative to the canine parvovirus (CPV) genome; **(b)** Dependoparvovirus-derived EPVs loci shown relative to the adeno-associated virus 2 (AAV-2) genome; **(c)** EPV loci derived from Amdoparvovirus-like viruses shown relative to the Aleutian mink disease (AMDV) genome; **(d)** Erythroparvovirus-derived loci shown relative to the parvovirus B19 genome; **(e)** EPVs derived from unclassified parvoviruses shown relative to a generic parvovirus genome. **(f)** Icthamaparvovirus-derived loci shown relative to *Syngnathus scovelli* parvovirus (SscPV); EPV locus identifiers are shown on the left. Solid bars to the right of each EPV set show taxonomic subgroupings below genus level. Where numbers are shown to the immediate right, the sequence shown is a consensus and numbers indicate how many individual orthologs sequences were used to create the consensus. Boxes bounding EPV elements indicate either (i) the presence of an identified gene (see **Tables S1-S6**), (ii) an uncharacterised genomic flanking region, or (iii) a truncated contig sequence (see key). EPV locus identifiers use six letter abbreviations to indicate host species (**Table S8**). **Abbreviations**: NS = non-structural protein; VP = capsid protein; ORF = open reading frame. ITR=Inverted terminal repeat; PLA2 = phospholipase A2 motif.

Figure 2. Evolution of subfamily *Parvoviridae*.

A maximum likelihood phylogeny showing the reconstructed evolutionary relationships between contemporary parvoviruses and the ancient parvovirus species represented by endogenous parvoviral elements (EPVs). Panels (A) and (B) show a more detailed view of subclades (labelled I and II) within the phylogeny shown in panel (C). The complete phylogeny, which is midpoint rooted for display purposes, was reconstructed using a multiple sequence alignment spanning 270 amino acid residues positions of the Rep protein and the LG likelihood substitution model. Coloured brackets indicate the established parvovirus genera recognised by the International Committee for the Taxonomy of Viruses. Bootstrap support values (1000 replicates) are shown for deeper internal nodes only. Scale bars show evolutionary distance in substitutions per site. Taxa labels are coloured based on taxonomic grouping as indicated by brackets, unclassified taxa are shown in black. Viral taxa are shown in bold, while EPV taxa are show in regular text. **Abbreviations**: PV=Parvovirus; HHV=Human herpesvirus; AAV=Adeno-associated virus; AMDV=Aleutian mink disease; CPV=canine parvovirus; BPV=bovine parvovirus; BrdPV=Bearded dragon parvovirus; MdPV=Muscovy duck parvovovirus; SlPV=slow loris parvovirus. TS=transcription strategy; MTSP=Multiple transcriptional start

positions;STSP+=single transcription start position plus additional strategies; HOMO=homoteleomeric; HETERO=heteroteleomeric.

Figure 3. Incorporation of EPVs into the vertebrate germline.

A time-calibrated evolutionary tree of vertebrate species examined in this study, illustrating the distribution of germline incorporation events over time. Colours indicate parvovirus genera as shown in the key. Diamonds on internal nodes indicate minimum age estimates for EPV loci endogenization (calculated for EPV loci found in >1 host species). Coloured circles adjacent to tree tips indicate the presence of EPVs in host taxa, with the diameter of the circle reflecting the number of EPVs identified. Brackets to the left show taxonomic groups within vertebrates.

Figure 4. Protoparvovirus evolution has been shaped by mammalian vicariance.

(A) Maximum likelihood-based phylogenetic reconstructions of evolutionary relationships between contemporary parvovirus species and the ancient parvovirus species represented by endogenous parvoviral elements (EPVs). The phylogeny was constructed from a multiple sequence alignment spanning 712 amino acid residues in the Rep protein (substitution model=LG likelihood). The tree is midpoint rooted for display purposes. Asterisks indicate nodes with bootstrap support >70% (1000 replicates). The scale bar shows evolutionary distance in substitutions per site. Coloured brackets to the right indicate (i) subgroups within the *Protoparvovirus* genus (outer set of brackets) and (ii) the host range of each subgroup (inner set of brackets). Terminal nodes are represented by squares (EPVs) and circles (viruses) and are coloured based on the biogeographic distribution of the host species in which they were identified. Coloured diamonds on internal nodes show the inferred ancestral distribution of parvovirus ancestors, using colours that reflect the patterns of continental drift and associated mammalian vicariance shown in the maps in panel (B). **Evidence for the presence of the "Mesoprotoparvovirus" group in Afrotherians is presented in **Fig. 2**. **(B)** Mollweide projection maps showing how patterns of continental drift from 200-35 led to periods of biogeographic isolation for terrestrial mammals in Laurasia (Europe and Asia), South America, Australia Africa and Madagascar. The resulting vicariance is thought have contributed to the diversification of mammals, reflected in the mammalian phylogeny as shown in in panel (c). The majority of placental mammals (including rodents, primates, ungulates and bats) evolved in Laurasia. However, these groups later expanded into other continents, and fossil evidence indicates that the ancestors of today's "New World rodents" had arrived on the South American continent by~35 million years ago (Mya), if not earlier **(C)** A time-calibrated phylogeny of mammals with

annotations indicating the biogeographic associations of the major taxonomic groups of contemporary mammals and ancestral mammalian groups, following panel (b) and key 1. **(D)** A time-calibrated phylogeny of mammals annotated to indicate the distribution of protoparvovirus subgroups among mammalian groups, following key 2. Question marks indicate where it is unknown whether viral counterparts of the lineages represented by EPVs still circulate among contemporary members of the host species groups in which they are found. **Abbreviations**: Mya = millions of years ago; NW=New World; NW=New World; (OW); CPV=carnivore parvovirus type 1; PPV=porcine parvovirus; HV=Hamster parvovirus; TuV=Tusavirus.

Figure 5. Dependoparvovirus evolution and the influence of inter-class transmission.
(A) A maximum likelihood phylogeny showing the reconstructed evolutionary relationships between contemporary dependoparvovirus species and the ancient dependoparvovirus species represented by EPVs. Virus taxa names are shown in bold, EPVs are shown in regular text. The phylogeny was constructed from a multiple sequence alignment spanning (MSA) 330 amino acid residues of the Rep protein and the LG likelihood substitution model and is rooted on the reptilian lineage as proposed by Penzes *et al* [45]. Brackets to the right indicate proposed taxonomic groupings. Shapes on leaf nodes indicate full-length EPVs and EPVs containing intact/expressed genes. Numbers next to leaf nodes indicate minimum age calibrations for EPV orthologs. Shapes on branches and internal nodes indicate different kinds of minimum age estimates for parvovirus lineages, as shown in the key. Numbers adjacent to node shapes show minimum age estimates in millions of years before present. For taxa that are not associated with mammals, organism silhouettes indicate species associations, following the key. The scale bar shows evolutionary distance in substitutions per site. Asterisks in circles indicate nodes with bootstrap support >70% (1000 replicates) in the tree shown. Plain asterisks next to internal nodes indicate nodes that are not supported in the tree shown here but do have bootstrap support >70% (1000 replicates) in phylogenies based on longer MSA partitions within Rep (but including less taxa). *Age calibrations based on data obtained in separated publications – see references [13] and [25]. **A contemporary virus derived from the marsupial clade has been reported in marsupials, but only transcriptome-based evidence is available [12]. **(B)** A time-calibrated phylogeny of vertebrate lineages showing proposed patterns of inter-class transmission in the Shirdaldependoparvovirus lineage. **Abbreviations**: PV=Parvovirus; AAV=Adeno-associated virus; BrdPV=Bearded dragon parvovirus; MdPV=Muscovy duck parvovovirus.

ACKNOWLEDGEMENTS

This work was supported by funding from the Association Monégasque Contre les Myopathies, and the Bill & Melinda Gates Foundation (OPP1202116).

COMPETING INTERESTS

R.M.K. is a co-founder of Synteny Therapeutics, Inc., which is a co-assignee of a patent application filed on behalf of University of Massachusetts Medical School and Synteny Therapeutics, Inc.

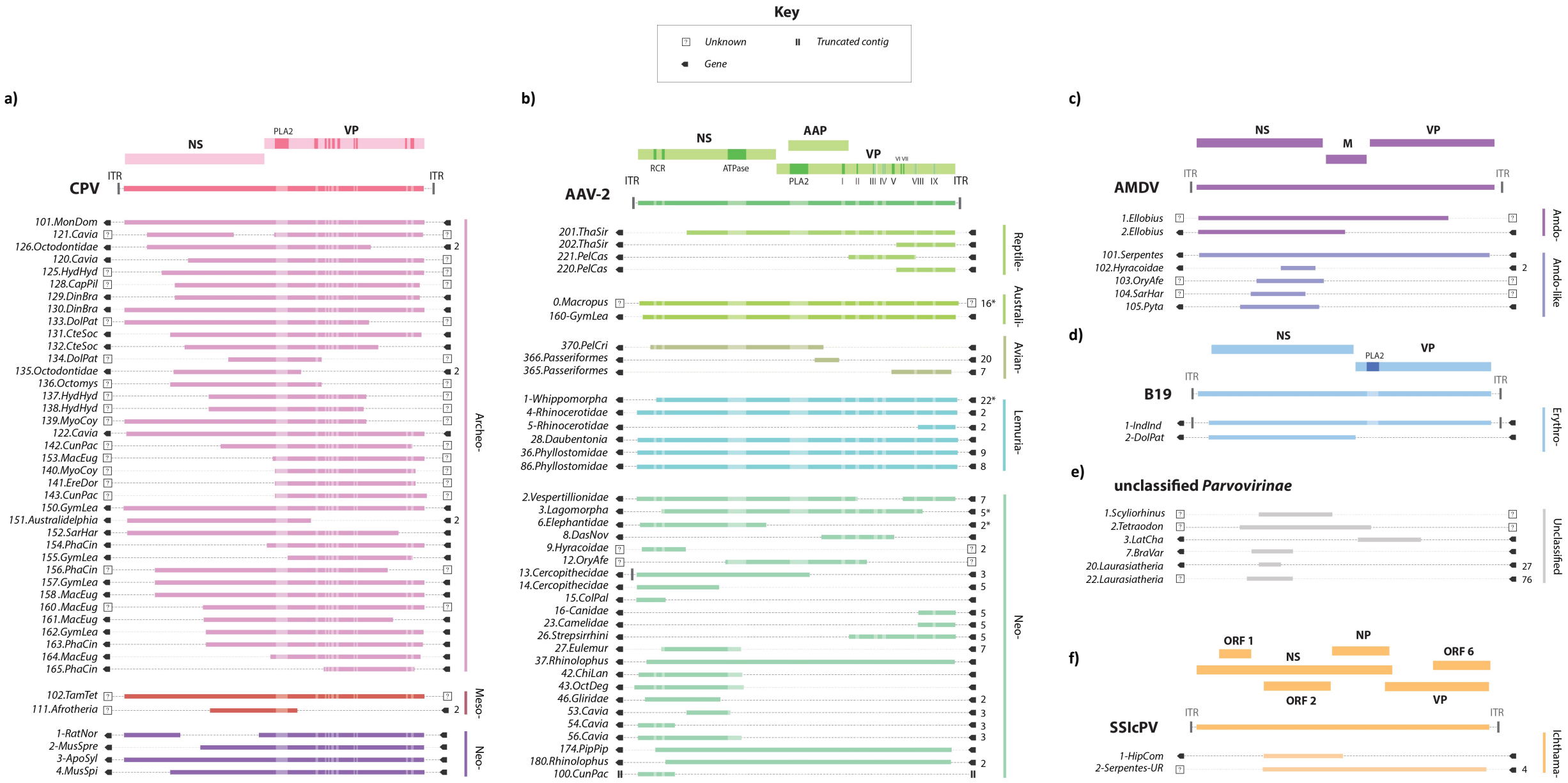
References

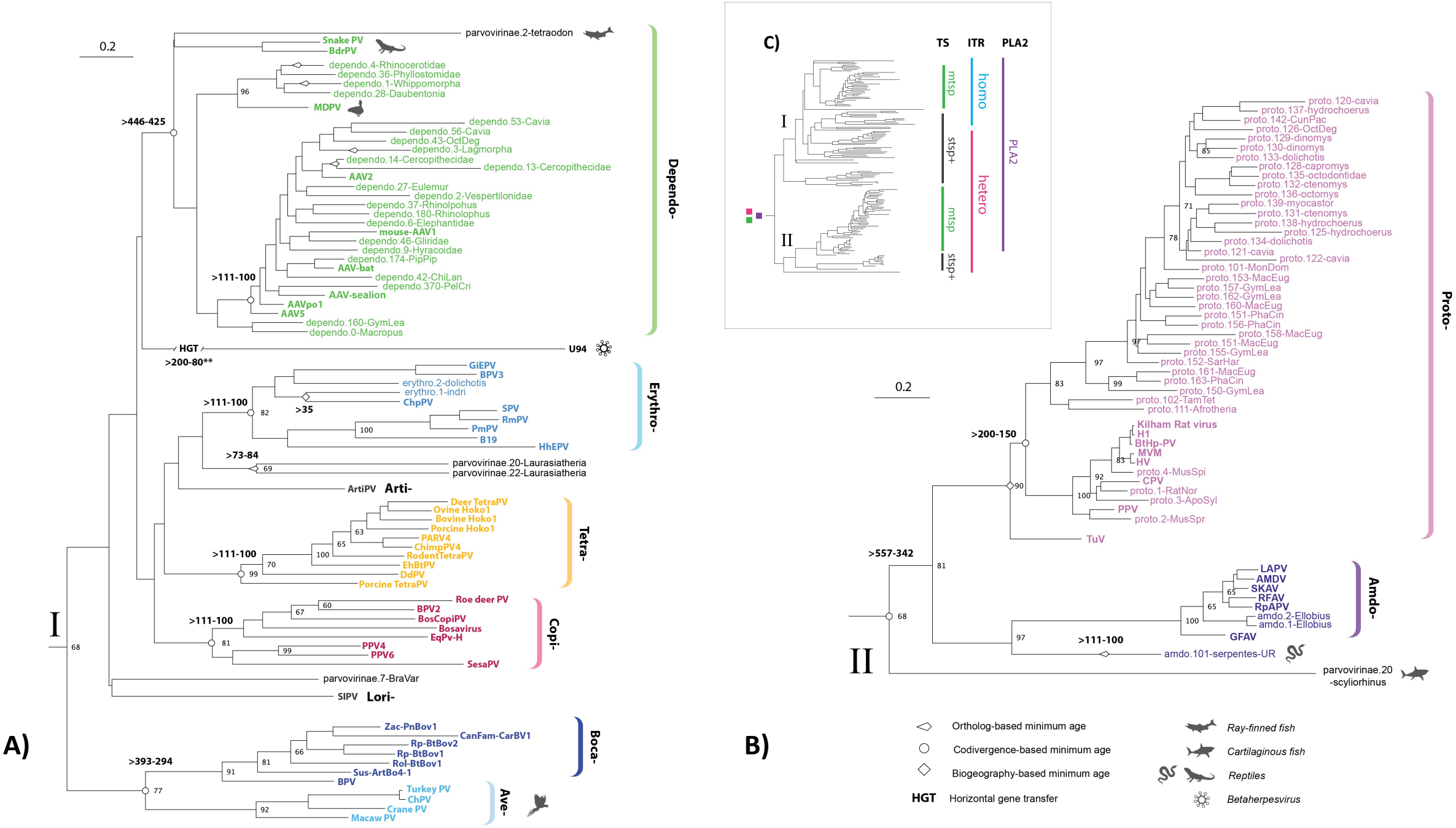
1. Cotmore, S.F., et al., *ICTV Virus Taxonomy Profile: Parvoviridae.* J Gen Virol, 2019. **100**(3): p. 367-368.
2. Cotmore, S.F. and P. Tattersall, *Parvoviruses: Small Does Not Mean Simple.* Annu Rev Virol, 2014. **1**(1): p. 517-37.
3. Nüesch, J.P., et al., *Molecular pathways: rodent parvoviruses--mechanisms of oncolysis and prospects for clinical cancer treatment.* Clin Cancer Res, 2012. **18**(13): p. 3516-23.
4. Hartley, A., et al., *A Roadmap for the Success of Oncolytic Parvovirus-Based Anticancer Therapies.* Annu Rev Virol, 2020. **7**(1): p. 537-557.
5. Fakhiri, J. and D. Grimm, *Best of most possible worlds: Hybrid gene therapy vectors based on parvoviruses and heterologous viruses.* Mol Ther, 2021.
6. Naso, M.F., et al., *Adeno-Associated Virus (AAV) as a Vector for Gene Therapy.* BioDrugs, 2017. **31**(4): p. 317-334.
7. Katzourakis, A. and R.J. Gifford, *Endogenous viral elements in animal genomes.* PLoS Genet, 2010. **6**(11): p. e1001191.
8. Belyi, V.A., A.J. Levine, and A.M. Skalka, *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**(23): p. 12458-62.
9. Kapoor, A., P. Simmonds, and W.I. Lipkin, *Discovery and characterization of mammalian endogenous parvoviruses.* J Virol, 2010. **84**(24): p. 12628-35.
10. Liu, H., et al., *Widespread endogenization of densoviruses and parvoviruses in animal and human genomes.* J Virol, 2011. **85**(19): p. 9863-76.
11. Francois, S., et al., *Discovery of parvovirus-related sequences in an unexpected broad range of animals.* Sci Rep, 2016. **6**: p. 30880.
12. Smith, R.H., et al., *Germline viral "fossils" guide in silico reconstruction of a mid-Cenozoic era marsupial adeno-associated virus.* Sci Rep, 2016. **6**: p. 28965.
13. Hildebrandt, E., et al., *Evolution of dependoparvoviruses across geological timescales – implications for design of AAV-based gene therapy vectors.* Virus Evolution, 2020.
14. Kotin, R.M., R.M. Linden, and K.I. Berns, *Characterization of a preferred site on human chromosome 19q for integration of adeno-associated virus DNA by non-homologous recombination.* Embo j, 1992. **11**(13): p. 5071-8.
15. Weitzman, M.D., et al., *Adeno-associated virus (AAV) Rep proteins mediate complex formation between AAV DNA and its integration site in human DNA.* Proc Natl Acad Sci U S A, 1994. **91**(13): p. 5808-12.

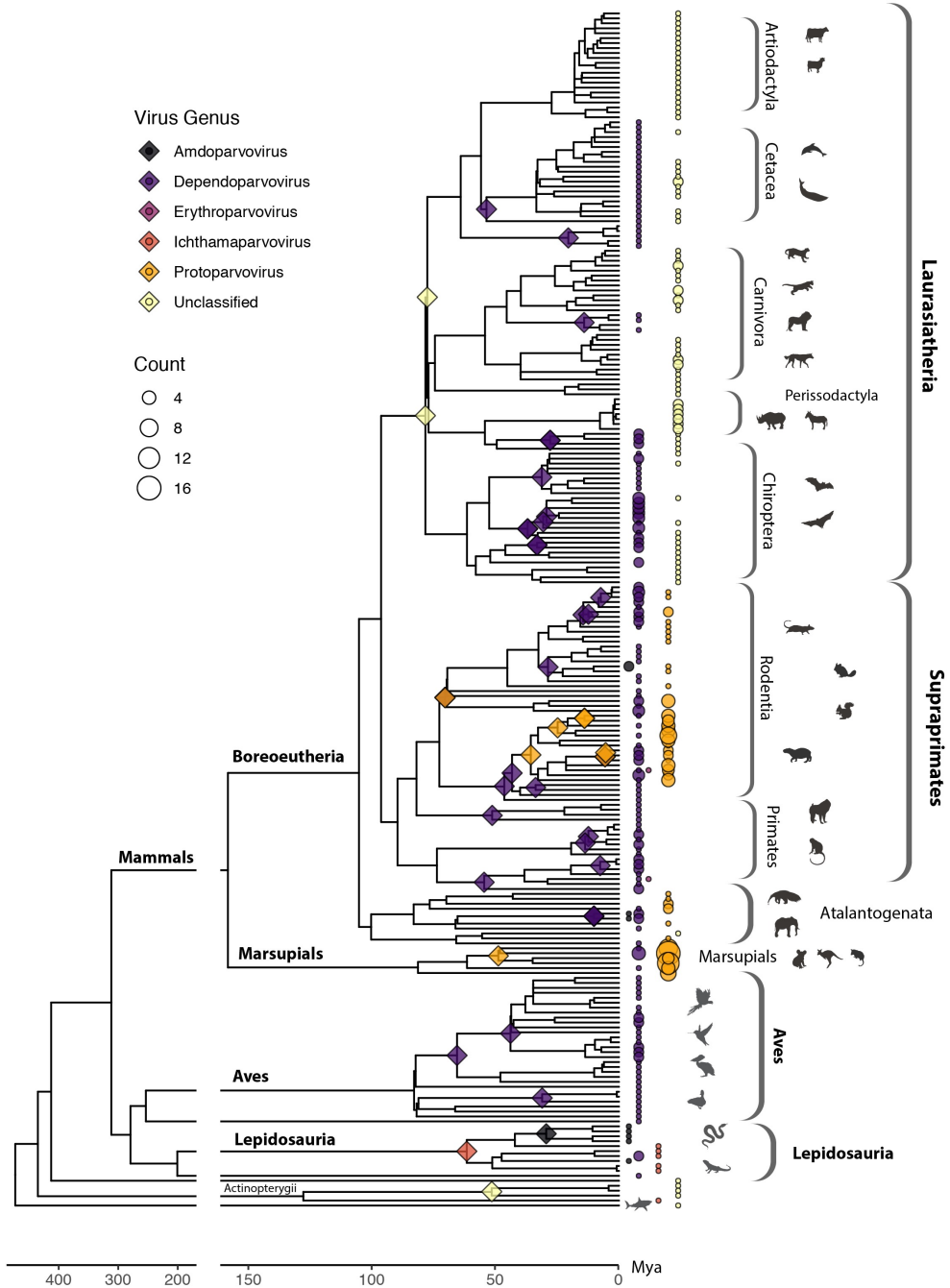
16. Kawasaki, J., et al., *One hundred million years history of bornavirus infections hidden in vertebrate genomes.* Proc Natl Acad Sci U S A, 2021: p. 2020.12.02.408005.
17. Katzourakis, A., et al., *Discovery and analysis of the first endogenous lentivirus.* Proc Natl Acad Sci U S A, 2007. **104**(15): p. 6261-5.
18. Feschotte, C. and C. Gilbert, *Endogenous viruses: insights into viral evolution and impact on host biology.* Nat Rev Genet, 2012. **13**(4): p. 283-96.
19. Lytras, S., G. Arriagada, and R.J. Gifford, *Ancient evolution of hepadnaviral paleoviruses and their impact on host genomes.* Virus Evol, 2021. **7**(1): p. veab012.
20. Kawasaki, J., et al., *100-My history of bornavirus infections hidden in vertebrate genomes.* Proc Natl Acad Sci U S A, 2021. **118**(20).
21. Cornelis, G., et al., *An endogenous retroviral envelope syncytin and its cognate receptor identified in the viviparous placental Mabuya lizard.* Proc Natl Acad Sci U S A, 2017. **114**(51): p. E10991-e11000.
22. Pastuzyn, E.D., et al., *The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer.* Cell, 2018. **173**(1): p. 275.
23. Herniou, E.A., et al., *When parasitic wasps hijacked viruses: genomic and functional evolution of polydnaviruses.* Philos Trans R Soc Lond B Biol Sci, 2013. **368**(1626): p. 20130051.
24. Arriagada, G. and R.J. Gifford, *Parvovirus-derived endogenous viral elements in two South American rodent genomes.* J Virol, 2014. **88**(20): p. 12158-62.
25. Kobayashi, Y., et al., *An endogenous adeno-associated virus element in elephants.* Virus Res, 2018.
26. Valencia-Herrera, I., et al., *Molecular Properties and Evolutionary Origins of a Parvovirus-Derived Myosin Fusion Gene in Guinea Pigs.* J Virol, 2019. **93**(17).
27. Grüning, B., et al., *Practical Computational Reproducibility in the Life Sciences.* Cell Syst, 2018. **6**(6): p. 631-635.
28. Ali, R.H., M. Bogusz, and S. Whelan, *Identifying Clusters of High Confidence Homologies in Multiple Sequence Alignments.* Mol Biol Evol, 2019. **36**(10): p. 2340-2351.
29. Holmes, E.C. and S. Duchêne, *Can Sequence Phylogenies Safely Infer the Origin of the Global Virome?* mBio, 2019. **10**(2).
30. Singer, J.B., et al., *GLUE: a flexible software system for virus sequence data.* BMC Bioinformatics, 2018. **19**(1): p. 532.
31. Gifford, R.J. *Parvovirus-GLUE*. 2021; Available from: https://giffordlabcvr.github.io/Parvovirus-GLUE/.
32. Merkel, D., *Docker: lightweight linux containers for consistent development and deployment.* Linux Journal, 2014. **239**(2).
33. Campbell, K., et al., *Making Genomic Surveillance Deliver: A Lineage Classification and Nomenclature System to Inform Rabies Elimination.* bioRxiv, 2021: p. 2021.10.13.464180.
34. Pénzes, J.J., et al., *Endogenous amdoparvovirus-related elements reveal insights into the biology and evolution of vertebrate parvoviruses.* Virus evolution, 2018. **4**(2): p. vey026-vey026.
35. Cui, J., et al., *Low frequency of paleoviral infiltration across the avian phylogeny.* Genome Biol, 2014. **15**(12): p. 539.
36. Gompels, U.A., et al., *The DNA sequence of human herpesvirus-6: structure, coding content, and genome evolution.* Virology, 1995. **209**(1): p. 29-51.
37. Aswad, A. and A. Katzourakis, *The first endogenous herpesvirus, identified in the tarsier genome, and novel sequences from primate rhadinoviruses and lymphocryptoviruses.* PLoS genetics, 2014. **10**(6): p. e1004332-e1004332.

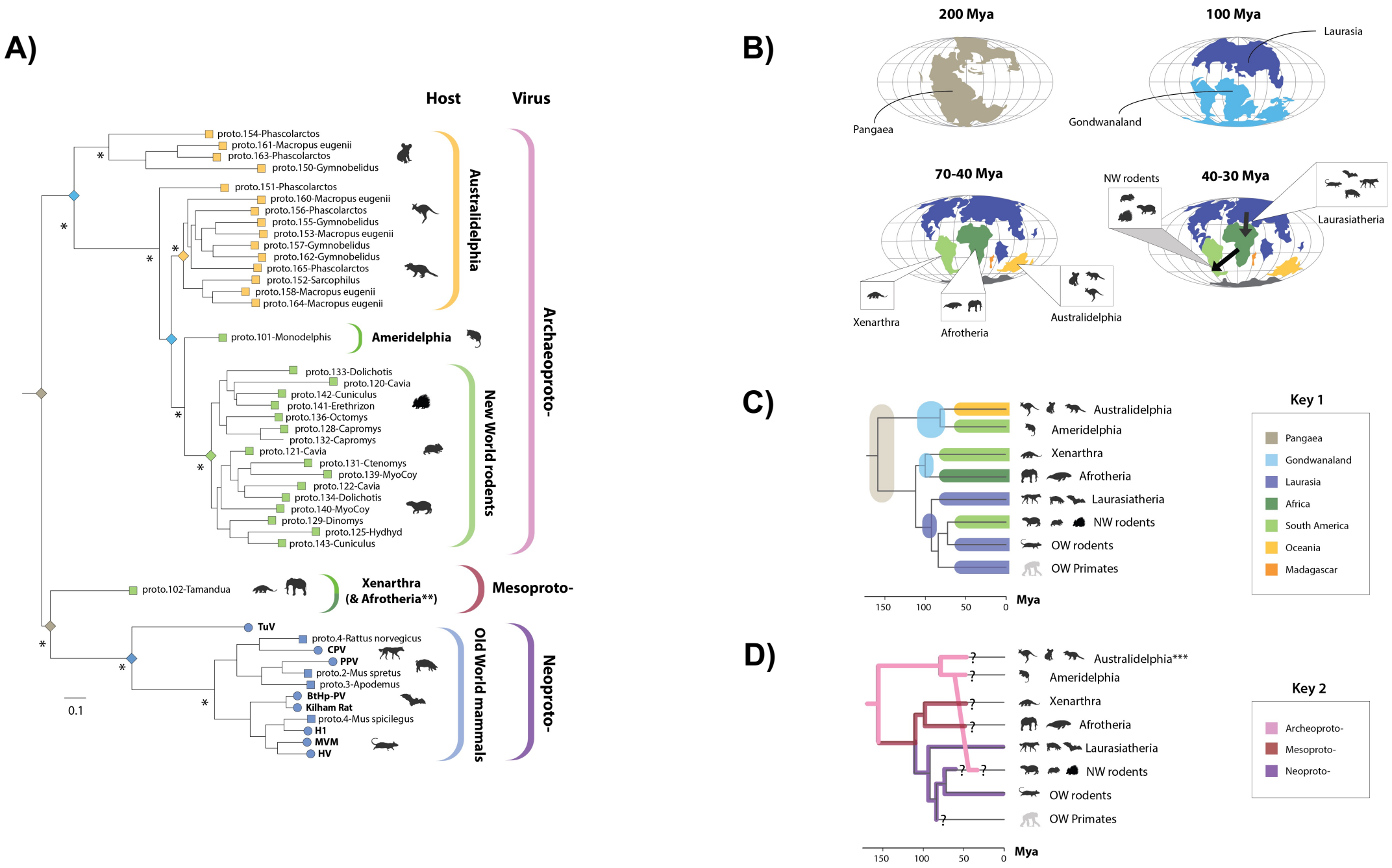
38. McGeoch, D.J., F.J. Rixon, and A.J. Davison, *Topics in herpesvirus genomics and evolution.* Virus Res, 2006. **117**(1): p. 90-104.
39. Cotmore, S.F. and P. Tattersall, *Parvovirus diversity and DNA damage responses.* Cold Spring Harb Perspect Biol, 2013. **5**(2).
40. Callaway, H.M., et al., *Examination and Reconstruction of Three Ancient Endogenous Parvovirus Capsid Protein Gene Remnants Found in Rodent Genomes.* J Virol, 2019. **93**(6).
41. Wu, Z., et al., *Deciphering the bat virome catalog to better understand the ecological diversity of bat viruses and the bat origin of emerging infectious diseases.* Isme j, 2016. **10**(3): p. 609-20.
42. Väisänen, E., et al., *Human Protoparvoviruses.* Viruses, 2017. **9**(11).
43. Springer, M.S., et al., *The historical biogeography of Mammalia.* Philos Trans R Soc Lond B Biol Sci, 2011. **366**(1577): p. 2478-502.
44. Poux, C., et al., *Arrival and diversification of caviomorph rodents and platyrrhine primates in South America.* Syst Biol, 2006. **55**(2): p. 228-44.
45. Pénzes, J.J., et al., *Novel parvoviruses in reptiles and genome sequence of a lizard parvovirus shed light on Dependoparvovirus genus evolution.* J Gen Virol, 2015. **96**(9): p. 2769-2779.
46. Souza, W.M., et al., *Chapparvoviruses occur in at least three vertebrate classes and have a broad biogeographic distribution.* J Gen Virol, 2017. **98**(2): p. 225-229.
47. Pénzes, J.J., et al., *An Ancient Lineage of Highly Divergent Parvoviruses Infects both Vertebrate and Invertebrate Hosts.* Viruses, 2019. **11**(6).
48. Imperiale, M.J. and M. Jiang, *What DNA viral genomic rearrangements tell us about persistence.* J Virol, 2015. **89**(4): p. 1948-50.
49. Hickman, A.B., et al., *Structural unity among viral origin binding proteins: crystal structure of the nuclease domain of adeno-associated virus Rep.* Mol Cell, 2002. **10**(2): p. 327-37.
50. Koonin, E.V., *Temporal order of evolution of DNA replication systems inferred by comparison of cellular and viral DNA polymerases.* Biol Direct, 2006. **1**: p. 39.
51. Schmidt, M., S. Afione, and R.M. Kotin, *Adeno-associated virus type 2 Rep78 induces apoptosis through caspase activation independently of p53.* J Virol, 2000. **74**(20): p. 9441-50.
52. Zimmermann, B., et al., *PrKX is a novel catalytic subunit of the cAMP-dependent protein kinase regulated by the regulatory subunit type I.* J Biol Chem, 1999. **274**(9): p. 5370-8.
53. Di Pasquale, G. and J.A. Chiorini, *PKA/PrKX activity is a modulator of AAV/adenovirus interaction.* Embo j, 2003. **22**(7): p. 1716-24.
54. Hermonat, P.L., *The adeno-associated virus Rep78 gene inhibits cellular transformation induced by bovine papillomavirus.* Virology, 1989. **172**(1): p. 253-61.
55. Brister, J.R., et al., *NCBI viral genomes resource.* Nucleic Acids Res, 2015. **43**(Database issue): p. D571-7.
56. Zhu, H., et al., *Database-integrated genome screening (DIGS): exploring genomes heuristically using sequence similarity search tools and a relational database.* bioRxiv, 2018.
57. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nuc. Acids Res., 1997. **25**: p. 3389-3402.
58. Kitts, P.A., et al., *Assembly: a resource for assembled genomes at NCBI.* Nucleic Acids Res, 2016. **44**(D1): p. D73-80.
59. Stamatakis, A., *RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.* Bioinformatics, 2014. **30**(9): p. 1312-3.
60. Gifford, R.J., et al., *Nomenclature for endogenous retrovirus (ERV) loci.* Retrovirology, 2018. **15**(1): p. 59.

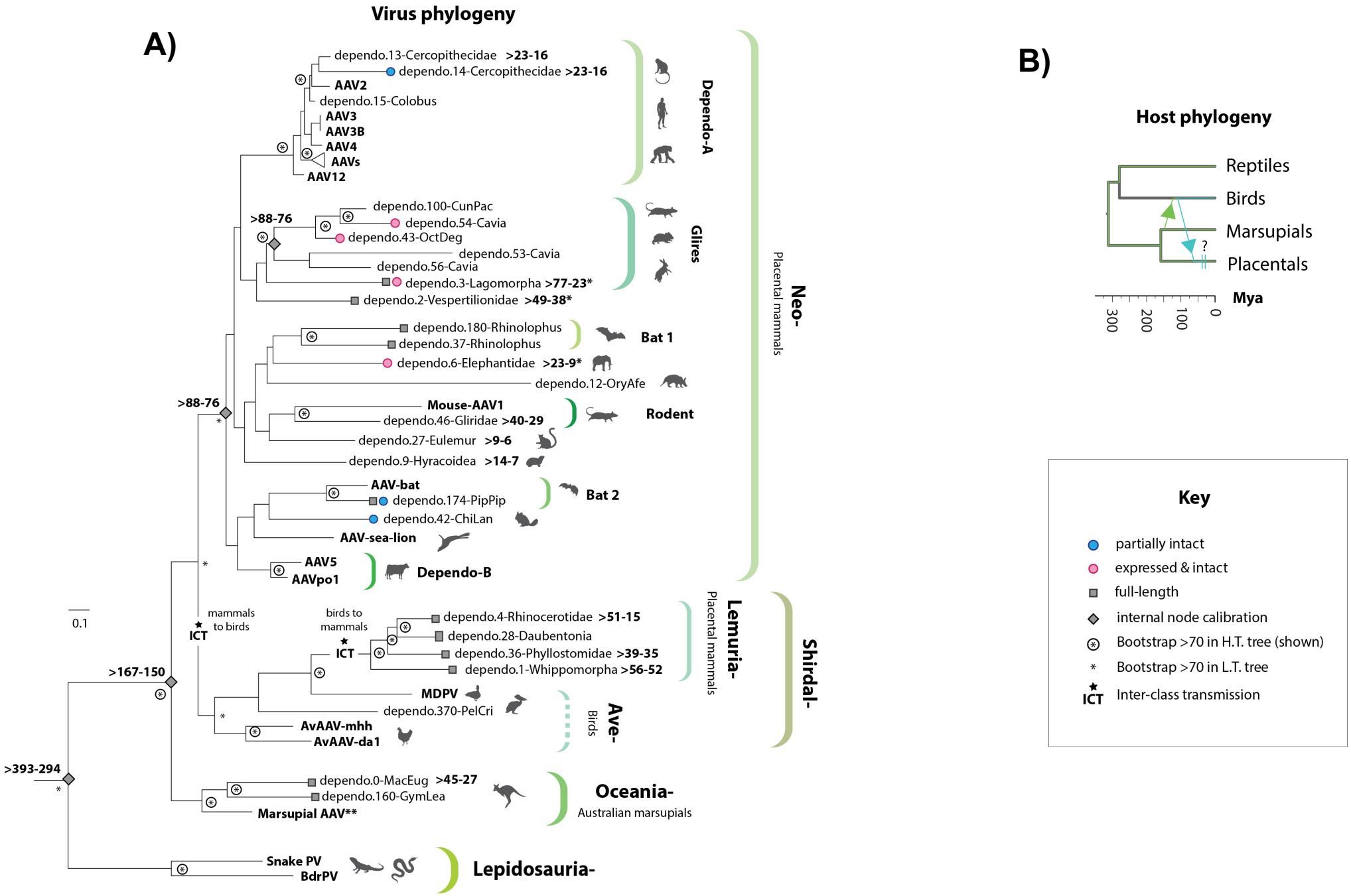
61. Kumar, S., et al., *TimeTree: A Resource for Timelines, Timetrees, and Divergence Times.* Mol Biol Evol, 2017. **34**(7): p. 1812-1819.
62. Clark, J.R., et al., *A comparative study in ancestral range reconstruction methods: retracing the uncertain histories of insular lineages.* Syst Biol, 2008. **57**(5): p. 693-707.
63. Chamberlain, S., et al., *rgbif: Interface to the Global Biodiversity Information Facility API*. 2021.
64. Ree, R.H. and S.A. Smith, *Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis.* Syst Biol, 2008. **57**(1): p. 4-14.
65. Wickham, H., *ggplot2: Elegant Graphics for Data Analysis*. 2016.
66. Yu, G., *Using ggtree to Visualize Data on Tree-Like Structures.* Curr Protoc Bioinformatics, 2020. **69**(1): p. e96.
67. Suyama, M., D. Torrents, and P. Bork, *PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W609-12.


Key
Unknown
Truncated contig
Gene
a)
CPV
NS
PLA2
VP
ITR
ITR
101.MonDom
121.Cavia
126.Octodontidae
120.Cavia
125.HydHyd
128.CapPil
129.DinBra
130.DinBra
133.DolPat
131.CteSoc
132.CteSoc
134.DolPat
135.Octodontidae
136.Octomys
137.HydHyd
138.HydHyd
139.MyoCoy
122.Cavia
142.CunPac
153.MacEug
140.MyoCoy
141.EreDor
143.CunPac
150.GymLea
151.Australidelphia
152.SarHar
154.PhaCin
155.GymLea
156.PhaCin
157.GymLea
158 .MacEug
160 .MacEug
161.MacEug
162.GymLea
163.PhaCin
164.MacEug
165.PhaCin
Archeo-
102.TamTet
111.Afrotheria
Meso-
1-RatNor
2-MusSpre
3-ApoSyl
4.MusSpi
Neo-
b)
AAV-2
NS
RCR
ATPase
AAP
VP
PLA2
I II III IV V VI VII VIII IX
ITR
ITR
201.ThaSir
202.ThaSir
221.PelCas
220.PelCas
Reptile-
0.Macropus
160-GymLea
Australi-
370.PelCri
366.Passeriformes
365.Passeriformes
Avian-
1-Whippomorpha
4-Rhinocerotidae
5-Rhinocerotidae
28.Daubentonia
36.Phyllostomidae
86.Phyllostomidae
Lemuria-
2.Vespertillionidae
3.Lagomorpha
6.Elephantidae
8.DasNov
9.Hyracoidae
12.OryAfe
13.Cercopithecidae
14.Cercopithecidae
15.ColPal
16-Canidae
23.Camelidae
26.Strepsirrhini
27.Eulemur
37.Rhinolophus
42.ChiLan
43.OctDeg
46.Gliridae
53.Cavia
54.Cavia
56.Cavia
174.PipPip
180.Rhinolophus
100.CunPac
Neo-
c)
AMDV
NS
M
VP
ITR
ITR
1.Ellobius
2.Ellobius
Amdo-
101.Serpentes
102.Hyracoidae
103.OryAfe
104.SarHar
105.Pyta
Amdo-like
d)
B19
NS
PLA2
VP
ITR
ITR
1-IndInd
2-DolPat
Erythro-
e)
unclassified Parvovirinae
1.Scyliorhinus
2.Tetraodon
3.LatCha
7.BraVar
20.Laurasiatheria
22.Laurasiatheria
Unclassified
f)
SSIcPV
ORF 1
NS
NP
ORF 6
ORF 2
VP
ITR
ITR
1-HipCom
2-Serpentes-UR
Ichthama-


A)
0.2
parvovirinae.2-tetraodon
Snake PV
BdrPV
dependo.4-Rhinocerotidae
dependo.36-Phyllostomidae
dependo.1-Whippomorpha
dependo.28-Daubentonia
MDPV
>446-425
dependo.53-Cavia
dependo.56-Cavia
dependo.43-OctDeg
dependo.3-Lagmorpha
dependo.14-Cercopithecidae
dependo.13-Cercopithecidae
AAV2
dependo.27-Eulemur
dependo.2-Vespertilonidae
dependo.37-Rhinolpohus
dependo.180-Rhinolophus
dependo.6-Elephantidae
mouse-AAV1
dependo.46-Gliridae
dependo.9-Hyracoidae
dependo.174-PipPip
AAV-bat
dependo.42-ChiLan
dependo.370-PelCri
AAV-sealion
AAVpo1
AAV5
dependo.160-GymLea
dependo.0-Macropus
>111-100
Dependo-
HGT
>200-80**
U94
GiEPV
BPV3
erythro.2-dolichotis
erythro.1-indri
ChpPV
>111-100
>35
SPV
RmPV
PmPV
B19
HhEPV
Erythro-
>73-84
parvovirinae.20-Laurasiatheria
parvovirinae.22-Laurasiatheria
ArtiPV
Arti-
Deer TetraPV
Ovine Hoko1
Bovine Hoko1
Porcine Hoko1
PARV4
ChimpPV4
RodentTetraPV
EhBtPV
DdPV
Porcine TetraPV
>111-100
Tetra-
Roe deer PV
BPV2
BosCopiPV
Bosavirus
EqPv-H
PPV4
PPV6
SesaPV
>111-100
Copi-
I
68
parvovirinae.7-BraVar
SIPV
Lori-
Zac-PnBov1
CanFam-CarBV1
Rp-BtBov2
Rp-BtBov1
Rol-BtBov1
Sus-ArtBo4-1
BPV
>393-294
Boca-
Turkey PV
ChPV
Crane PV
Macaw PV
Ave-
77
B)
C)
TS
ITR
PLA2
mtsp
stsp+
mtsp
stsp+
homo
hetero
PLA2
I
II
proto.120-cavia
proto.137-hydrochoerus
proto.142-CunPac
proto.126-OctDeg
proto.129-dinomys
proto.130-dinomys
proto.133-dolichotis
proto.128-capromys
proto.135-octodontidae
proto.132-ctenomys
proto.136-octomys
proto.139-myocastor
proto.131-ctenomys
proto.138-hydrochoerus
proto.125-hydrochoerus
proto.134-dolichotis
proto.121-cavia
proto.122-cavia
proto.101-MonDom
proto.153-MacEug
proto.157-GymLea
proto.162-GymLea
proto.160-MacEug
proto.151-PhaCin
proto.156-PhaCin
proto.158-MacEug
proto.151-MacEug
proto.155-GymLea
proto.152-SarHar
proto.161-MacEug
proto.163-PhaCin
proto.150-GymLea
proto.102-TamTet
proto.111-Afrotheria
Proto-
0.2
>200-150
Kilham Rat virus
H1
BtHp-PV
MVM
HV
proto.4-MusSpi
CPV
proto.1-RatNor
proto.3-ApoSyl
PPV
proto.2-MusSpr
TuV
>557-342
81
68
LAPV
AMDV
SKAV
RFAV
RpAPV
amdo.2-Ellobius
amdo.1-Ellobius
GFAV
Amdo-
>111-100
amdo.101-serpentes-UR
II
parvovirinae.20
-scyliorhinus
Ortholog-based minimum age
Codivergence-based minimum age
Biogeography-based minimum age
HGT Horizontal gene transfer
Ray-finned fish
Cartilaginous fish
Reptiles
Betaherpesvirus


Virus Genus
Amdoparvovirus
Dependoparvovirus
Erythroparvovirus
Ichthamaparvovirus
Protoparvovirus
Unclassified
Count
4
8
12
16
Mammals
Boreoeutheria
Marsupials
Aves
Lepidosauria
Actinopterygii
Artiodactyla
Cetacea
Carnivora
Perissodactyla
Chiroptera
Rodentia
Primates
Atalantogenata
Marsupials
Aves
Lepidosauria
Laurasiatheria
Supraprimates
400
300
200
150
100
50
0
Mya


A)
Host
Virus
proto.154-Phascolarctos
proto.161-Macropus eugenii
proto.163-Phascolarctos
proto.150-Gymnobelidus
proto.151-Phascolarctos
proto.160-Macropus eugenii
proto.156-Phascolarctos
proto.155-Gymnobelidus
proto.153-Macropus eugenii
proto.157-Gymnobelidus
proto.162-Gymnobelidus
proto.165-Phascolarctos
proto.152-Sarcophilus
proto.158-Macropus eugenii
proto.164-Macropus eugenii
Australidelphia
proto.101-Monodelphis
Ameridelphia
proto.133-Dolichotis
proto.120-Cavia
proto.142-Cuniculus
proto.141-Erethrizon
proto.136-Octomys
proto.128-Capromys
proto.132-Capromys
proto.121-Cavia
proto.131-Ctenomys
proto.139-MyoCoy
proto.122-Cavia
proto.134-Dolichotis
proto.140-MyoCoy
proto.129-Dinomys
proto.125-Hydhyd
proto.143-Cuniculus
New World rodents
Archaeoproto-
proto.102-Tamandua
Xenarthra
(& Afrotheria**)
Mesoproto-
TuV
proto.4-Rattus norvegicus
CPV
PPV
proto.2-Mus spretus
proto.3-Apodemus
BtHp-PV
Kilham Rat
proto.4-Mus spicilegus
H1
MVM
HV
Old World mammals
Neoproto-
0.1
B)
200 Mya
Pangaea
100 Mya
Laurasia
Gondwanaland
70-40 Mya
Xenarthra
Afrotheria
Australidelphia
NW rodents
40-30 Mya
Laurasiatheria
C)
Australidelphia
Ameridelphia
Xenarthra
Afrotheria
Laurasiatheria
NW rodents
OW rodents
OW Primates
150
100
50
0
Mya
Key 1
Pangaea
Gondwanaland
Laurasia
Africa
South America
Oceania
Madagascar
D)
Australidelphia***
Ameridelphia
Xenarthra
Afrotheria
Laurasiatheria
NW rodents
OW rodents
OW Primates
150
100
50
0
Mya
Key 2
Archeoproto-
Mesoproto-
Neoproto-


A)
Virus phylogeny
dependo.13-Cercopithecidae >23-16
dependo.14-Cercopithecidae >23-16
AAV2
dependo.15-Colobus
AAV3
AAV3B
AAV4
AAVs
AAV12
Dependo-A
dependo.100-CunPac
dependo.54-Cavia
dependo.43-OctDeg
dependo.53-Cavia
dependo.56-Cavia
dependo.3-Lagomorpha >77-23*
dependo.2-Vespertilionidae >49-38*
Glires
>88-76
dependo.180-Rhinolophus
dependo.37-Rhinolophus
Bat 1
dependo.6-Elephantidae >23-9*
dependo.12-OryAfe
Mouse-AAV1
dependo.46-Gliridae >40-29
Rodent
dependo.27-Eulemur >9-6
dependo.9-Hyracoidea >14-7
AAV-bat
dependo.174-PipPip
dependo.42-ChiLan
Bat 2
AAV-sea-lion
AAV5
AAVpo1
Dependo-B
Neo-
Placental mammals
0.1
mammals to birds
ICT
birds to mammals
ICT
dependo.4-Rhinocerotidae >51-15
dependo.28-Daubentonia
dependo.36-Phyllostomidae >39-35
dependo.1-Whippomorpha >56-52
Lemuria-
Placental mammals
Shirdal-
>167-150
MDPV
dependo.370-PelCri
AvAAV-mhh
AvAAV-da1
Ave-
Birds
>393-294
dependo.0-MacEug >45-27
dependo.160-GymLea
Marsupial AAV**
Oceania-
Australian marsupials
Snake PV
BdrPV
Lepidosauria-
B)
Host phylogeny
Reptiles
Birds
Marsupials
Placentals
?
Mya
300
200
100
0
Key
partially intact
expressed & intact
full-length
internal node calibration
Bootstrap >70 in H.T. tree (shown)
Bootstrap >70 in L.T. tree
ICT Inter-class transmission