A novel approach to exploring the dark genome and its application to mapping of the vertebrate virus 'fossil record'.

Daniel Blanco-Melo[1]*, Matthew A. Campbell[2]*, Henan Zhu[3], Tristan P.W. Dennis[3], Sejal Modha[3], Spyros Lytras[3], Joseph Hughes[3], Anna Gatseva[3], Robert J. Gifford[3,4]**

[1] Vaccine and Infectious Disease Division, *Fred Hutchinson Cancer Center, 1100 Fairview Ave N, Seattle, WA 98109, United States*

[2] *University of California, Davis, 1 Shields Ave, Davis, CA 95616, USA*

[3] *MRC-University of Glasgow Centre for Virus Research, 464 Bearsden Rd, Bearsden, Glasgow, UK, G61 1QH*

[4] *Centre for Epidemic Response and Innovation (CERI), School of Data Science and Computational Thinking, Stellenbosch University, Stellenbosch, South Africa*

*** Equal contributions.**

**** Correspondence:** Robert J Gifford (robert.gifford@glasgow.ac.uk)

ABSTRACT

Background:

Genomic regions that remain poorly understood, often referred to as the "dark genome," contain a variety of functionally relevant and biologically informative genome features. These include endogenous viral elements (EVEs) - virus-derived sequences that can dramatically impact host biology and serve as a virus “fossil record”. In this study, we introduce a database-integrated genome screening (DIGS) approach to investigating the dark genome *in silico*, focusing on EVEs found within vertebrate genomes.

Results:

Using DIGS on 874 vertebrate species genomes, we uncovered approximately 1.1 million EVE sequences, with over 99% originating from endogenous retroviruses or transposable elements that contain EVE DNA. We show that the remaining 6038 sequences represent over a thousand distinct horizontal gene transfer events across ten virus families, including some that have not previously been reported as EVEs. We explore the genomic and phylogenetic characteristics of non-retroviral EVEs and determine their rates of acquisition during vertebrate evolution. Our study uncovers novel virus diversity and broadens our knowledge of virus distribution among vertebrate hosts. It also provides new insights into the long-term evolution of highly pathogenic filoviruses.

Conclusions:

We comprehensively catalogue and analyse EVEs within 874 vertebrate genomes, shedding light on the distribution, diversity and long-term evolution of viruses, and revealing their extensive impact on vertebrate genome evolution. Our results demonstrate the power of linking a relational database management system to a similarity search-based screening pipeline for *in silico* exploration of the dark genome.

INTRODUCTION

The availability of whole genome sequence (WGS) data from a broad range of species provides unprecedented scope for comparative genomic investigations [1-3]. However, these investigations rely to a large extent on *annotation* - the process of identifying and labelling genome features - which usually lags far behind the generation of sequence data. Consequently, most whole genome sequences are comprised of DNA that is incompletely understood in terms of its evolutionary origins and functional significance. The portion of sequenced genome space that lacks annotations is sometimes referred to as the 'dark genome' [4], and contains a wide variety of yet-to-be-characterized genome features. Some of these may have functional roles, such as encoding proteins [5] or regulating gene expression [6]. Others, such as non-expressed pseudogenes, may not, but can nonetheless provide valuable insights into genome biology and evolution.

Within the dark genome, endogenous viral elements (EVEs) constitute a particularly intriguing group of genome features. EVEs are virus-derived DNA sequences that become integrated into the germline genome of host species and are stably inherited as host alleles – a form of horizontal gene transfer [7-14]. While once considered genetic 'junk,' it has become evident over recent years that EVEs can profoundly impact host biology and genome evolution, with many now known to have physiologically relevant roles [15-19]. In addition, EVE sequences (whether functional or not) provide a rare source of retrospective information about ancient viruses, akin to a viral 'fossil record' [7, 20-22].

Identifying genome features contained within the dark genome, such as EVEs, often relies on the use of sequence similarity searches, such as those implemented in the Basic Local Alignment Search Tool (BLAST) [23, 24], to search WGS databases. Because sequence similarity reflects homology (evolutionary relatedness), novel genome features can often be identified based on their resemblance to ones that have been described previously. One example of this approach is implemented in the PSI-BLAST [5] and HMMR [8] programs, in

which iterated search strategies are used to progressively increase sensitivity so that novel homologs of previously characterised genes may be detected. A related approach is 'systematic *in silico* genome screening' which extends the basic concept of a similarity search in two ways: (i) inclusion of multiple query sequences and/or target databases; (ii) similarity-based classification of matching sequences ('hits') via comparison to reference sequence library (**Fig. 1a**). Hits may also be further investigated using additional comparative or experimental approaches (**Fig. 1b, Table 1**). Thus, screening can provide one component of a broader analytical pipeline.

While straightforward in principle*, in silico* genome screening is computationally expensive and can be difficult to implement efficiently. Moreover, large-scale screens can produce copious output data that are difficult to manage and interpret without an appropriate analytical framework. To address these issues, we developed a database-oriented approach to *in silico* screening, called *database-integrated genome screening* (DIGS). To demonstrate the use of this approach, we first created an open software framework for performing it, then used this framework to search published vertebrate genomes for EVE loci. Besides demonstrating that DIGS provides a powerful, flexible approach for exploring the dark genome, our analysis provides a comprehensive and detailed overview of EVE diversity in vertebrate genomes and reveals new information about the long-term evolutionary relationships between viruses and vertebrate hosts.

RESULTS

1. A database-integrated approach to exploring the dark genome

We developed a robust, database-integrated approach to systematic *in silico* genome screening, referred to as database-integrated genome screening (DIGS). This approach integrates a similarity search-based screening pipeline with a relational database management system (RDBMS) to enable efficient exploration of the dark genome. The rationale for this integration is twofold: it not only provides a solid foundation for conducting

large-scale, automated screens in an efficient and non-redundant manner but also allows for the structured querying of screening output using SQL, a powerful and well-established tool for database interrogation [25]. Additionally, an RDBMS offers advantages such as data recoverability, multi-user support, and networked data access.

The DIGS process comprises three key input data components:

Target Database (TDb): A collection of whole genome sequence assemblies (or other large sequence datasets such as transcriptomes) that will serve as the target for sequence similarity searches.

Query Sequences (Probes): A set of sequences to be used as input for similarity searches of the TDb.

Reference Sequence Library (RSL): The RSL represents the broad range of genetic diversity associated with the genome feature(s) under investigation. Its composition varies according to the analysis context (see **Table 1**). It should always include sequences representing diversity within the genome feature under investigation. It may also include genetic marker sequences and potentially cross-matching genome features. Probes are typically a subset of sequences contained in the RSL.

As illustrated in **Fig. 2**, the DIGS process involves systematic searching of a user-defined TDb with user-defined probes, merging fragmented hits, and classifying merged sequences through BLAST-based comparison to the RSL. The output - a set of non-redundant, defragmented 'hits' – is captured in a project-specific relational database. Importantly, this integration allows database queries to be employed in real time, with SQL queries referencing any information captured by the database schema. SQL-based querying of screening databases facilitates the identification of loci of interest, which can then be explored further using comparative approaches (see **Fig. 1b**).

It is important to note that screening is usually an iterative discovery process, wherein initial results inform the development of subsequent screens. For instance, novel diversity detected by an initial screen can subsequently be incorporated into the RSL, and hits within the screening database can be reclassified using the updated library (**Fig. 2**). Additionally, probe sets used in initial searches can be expanded to incorporate sequences identified during screening, broadening the range of sequences detected in subsequent screens [26]. However, care must be taken when using this approach, since it can potentially produce misleading results, or generate excessive hits (e.g. if highly repetitive sequences are contained within the new probes). Importantly, database integration allows results to be observed and interrogated in real time - as they are being generated. This means that configuration issues (e.g. badly composed RSL, inappropriate choice of probes) can be detected early on – potentially saving a significant amount of time and effort. Furthermore, it facilitates the implementation of agile, heuristic screening strategies, in which approaches are adjusted in line with results.

2. An open software framework for implementing DIGS

We constructed a software framework for implementing DIGS, called 'the DIGS tool'. The DIGS tool is implemented using the PERL scripting language. It uses the BLAST+ program suite [24] to perform similarity searches, and the MySQL RDBMS (to capture their output. Accessible through a text-based console interface, it simplifies the complex process of large-scale genome screening, and provides a versatile basis for implementing screens.

To initiate screening using the DIGS tool, researchers provide a project-specific command file (**Fig. S1**) that serves as the blueprint for the screening process. This command file specifies parameters, including the user-defined name of the screening database, and file paths to the TDb, RSL, and probe sequences. When a screen is initiated a project-specific database is created with the schema shown in **Fig. S2**. This core schema can be extended to include any relevant "side data" – e.g., taxonomic information related to the species and

sequences included in the screen - increasing the power of SQL queries to reveal informative patterns (**Fig. S2**, **Fig. S3**).

Systematic screening proceeds automatically until all searches have been completed. If the process is interrupted at any point, or if novel probe/target sequences are incorporated into the project, screening will proceed in a non-redundant way on restarting. Thus, screening projects can readily be expanded to incorporate new TDb files (e.g. recently published WGS assemblies), or novel probe/reference sequences, as they become available. The DIGS tool console allows reclassification of sequences held in the results table (e.g. following an RSL update). To increase efficiency, this process can be tailored to specific subsets of database sequences by supplying SQL constraints via the DIGS tool console.

BLAST algorithms emphasise local similarity and consequently tend to fragment contiguous matches into several separate hits if similarity across some internal regions of the match is low. The DIGS tool allows screening pipelines to be configured with respect to how overlapping/adjacent hits are handled, so that the screening process can be tailored to the specific needs of diverse projects. The DIGS tool also provides a ‘consolidation’ function that concatenates, rather than merges, adjacent hits and stores concatenated results, along with information about their structure, in a new screening database table.

For program validation, we mined mammalian genomes for sequences disclosing similarity to the antiviral restriction factor tetherin [27, 28]. Tetherin provides a useful test case as it is a relatively unique gene and its evolution, distribution and diversity have previously been examined [27, 28]. Results were compared with those provided by two alternative, widely used genome mining pipelines: OrthoDB [29] and Ensembl [30] and found to overlap by >99% (**Fig. S4**).

The DIGS tool provides functionality for exporting FASTA-formatted sequences and managing screening database tables (e.g., add/drop tables, import table data). Further information regarding program installation and usage is provided online, in a repository associated website [31]. In the sections below we illustrate the application of the DIGS tool to cataloguing of EVEs in vertebrate genomes, focussing on both high and low copy number elements.

3. Use of DIGS to catalogue RT-encoding endogenous retroviruses

Unlike other vertebrate viruses, retroviruses (family *Retroviridae*) integrate their genome into the nuclear genome of infected cells as an obligate part of their life cycle. As a result, retroviruses gain more opportunities to become a permanent part of the host germline. Furthermore, the integrated form of a retrovirus genome, called a provirus, is typically replication competent and increases in germline copy number can thus occur through reinfection of germ line cells [32]. Accordingly, 'endogenous retroviruses' (ERVs) are by far the most common type of EVE found in vertebrate genomes [7, 33].

Retrovirus genomes contain a *pol* coding domain that encodes a reverse transcriptase (RT) gene. The RT gene can be used to reconstruct phylogenetic relationships across the entire *Retroviridae* and hence provides the lynchpin for unravelling the evolutionary history and origins of ERV loci [34, 35]. We therefore implemented a screening procedure to detect RT-encoding ERV loci, based on an RSL comprised of previously classified exogenous retrovirus and ERV RT sequences (**see Methods**). Screening involved more than 1.5 million discrete tBLASTn searches and resulted in the identification of 1,107,9451 ERV RT hits. This set was filtered based on a BLAST bitscore cut-off to obtain a high confidence set of 702,232 loci.

High confidence ERV RT hits were identified in all vertebrate classes. However, the frequency among classes was found to vary dramatically (**Fig. 3**). ERVs occur most

frequently in mammals (class Mammalia) and amphibians (class Amphibia), and at relatively similar, intermediate frequencies in the genomes of reptiles (class Squamata) and birds (class Aves). By contrast, RT-encoding ERVs are relatively rare in the genomes of fish, including ray-finned fish (class Actinopterygii) lobe-finned fish (class Sarcopterygii), and jawless vertebrates (class Agnatha). Interestingly, however, cartilaginous fish (class Chondrichthyes) represent an exception, exhibiting a median ERV copy number per species genome approaching that observed in amphibians and mammals (**Fig. 3**). These findings are broadly consistent with previous studies, conducted using a smaller amount of species genomes [33, 36].

ERVs have been taxonomically grouped into three clades (I, II and III) based on their phylogenetic relatedness in the reverse transcriptase (RT) gene to the exogenous *Gammaretrovirus*, *Betaretrovirus* and *Spumavirus* genera respectively [1,2]. We incorporated into our RT screening database taxonomic information for (i) host species examined in our screen and (ii) RSL RT sequences. We then used SQL statements referencing these tables to summarise the frequency of clade I, II and III ERVs in distinct vertebrate classes (**Fig. 3b**). Whereas clade I and III ERVs are found in all vertebrate groups, clade II ERVs appear to have a more restricted distribution, occurring only at low frequency in amphibians, and being completely absent from agnathans, cartilaginous fish and ray-finned fish. A few clade II ERVs were identified in these groups, but they are very closely related to mammalian ERVs and likely represent contamination of WGS builds with mammalian genomic DNA. While RT-encoding ERV copy number is relatively high in cartilaginous fish, RT diversity is relatively low, with the majority of ERV RT sequences belonging to clade III.

4. Use of DIGS to catalogue non-retroviral EVEs vertebrate genomes.

To identify non-retroviral EVEs, we first obtained an RSL representing all known viruses [37]. From this library, a set of representative probes was selected. Probes include representative

proteomes of all known vertebrate viruses except retroviruses. Screening entailed >1.5 million discrete tBLASTn searches, and initial results comprised 33,654 hits. However, many of these represented matches to host genes and TEs. We identified these spurious matches by interrogating screening results with a combination of SQL queries and *ad hoc* phylogenetic analysis. We also identified and excluded hits that appeared likely to derived from exogenous viruses (see **Table S2**). For example, SQL-generated summaries of our initial screen results revealed several WGS sequences disclosing unexpected similarity to plant virus genomes (**Fig. S3a**). Among these, matches to geminiviruses (family *Geminiviridae*) and potyviruses (family *Potyviridae*) lack evidence for germline integration and likely to represent diet-related contamination. Other unusual matches were contained within large contigs and thus could not be dismissed as contaminating DNA but were revealed to be spurious by genomic and phylogenetic analysis. For example, a sequence identified in the genome of the pig-nosed turtle (*Carettochelys insculpta*) disclosed similarity to caulimoviruses (family Caulimoviruses) - but was revealed by closer analysis to represent an unusual ERV (**Fig. S3a**, **Fig. S5**).

Next we removed matches to transposons partly comprised of virus-derived DNA such as the adintovirus-derived mavericks [38] and alloherpesvirus-derived teratorns [39] (**Fig. S3a**). Once TEs had been removed, results comprised 6038 putative non-retroviral EVE sequences, representing 10 virus families (**Table 2, Table S3-S11**). We did not identify any EVEs derived from vertebrate viruses with genomes comprised of double-stranded RNA (e.g., order Reovirales) or circular single-stranded RNA (e.g., genus *Deltavirus*). However, all other virus genome ‘classes’ were represented including reverse-transcribing DNA (DNArt) viruses, double-stranded DNA (DNAds) viruses, single-stranded DNA (DNAss) viruses, single-stranded negative sense RNA (RNAss-ve) viruses, and single-stranded positive sense RNA (RNAss+ve) viruses. Plotting the distribution of EVEs and exogenous viruses from distinct virus families and genera across vertebrate phyla, revealed that many

virus groups have had a broader distribution across vertebrate hosts than recognised on the basis of exogenous isolates (**Fig. 4**).

We examined all EVE loci identified in our study to determine their coding potential. We identified numerous EVE loci encoding open reading frames (ORFs) >300 amino acids (aa) in length (**Fig. S6**). Among these, four encoded ORFs longer than 1000 aa. One of these – a 1718aa ORF encoded by an endogenous borna-like L-protein (EBLL) element in bats (EBLL-Cultervirus.29-EptFus) – has been reported previously [40]. However, we also identified an endogenous chuvirus-like L-protein (ECLL) element encoding an ~1400 aa ORF in livebearers (subfamily Poeciliinae). This element encodes long ORFs in two distinct livebearer species (*P. formosa and P. latapina*), indicating it's coding capacity has been conserved for >10 million years [41]. We also detected herpesvirus and alloherpesvirus EVEs encoding ORFs >1000 aa, but as discussed below, the integration status of these sequences remains unclear.

5. Diversity of non-retroviral EVEs in vertebrate genomes

5.1 EVEs derived from viruses with double-stranded DNA genomes

We detected DNA derived from herpesviruses (family *Herpesviridae*) in mammalian and reptilian genomes (**Fig. 4, Table 2, Table S3**). DNA sequences derived from betaherpesviruses (subfamily *Betaherpesvirinae*) and gammaherpesviruses (subfamily *Gammaherpesvirinae*) have previously been reported in WGS assemblies of the tarsier (*Carlito syrichta*) and aye-aye (*Daubentonia madagascensis*), respectively [42]. In addition to these sequences, we detected gammaherpesvirus DNA in WGS data of red squirrels (*Sciurus vulgaris*) and the Amazon river dolphin (*Inia geoffrensis*), while betaherpesvirus DNA was detected in the stoat (*Mustela ermina*) WGS assembly, and DNA derived from an alphaherpesvirus (subfamily *Alphaherpesvirinae*) in WGS data of the Oriximina lizard (*Tretioscincus oriximinensis*) (**Fig. S7a-b**). Germline integration of human betaherpesviruses has been demonstrated [43, 44], and the presence of a betaherpesvirus-derived EVE in the

tarsier genome EVE has been established [42]. However, herpesviruses can also establish latent infections, and might be considered likely to occur as contaminants of DNA samples used to generate whole genome sequence assemblies. Due to the limitations of the WGS assemblies in which they were identified, it was not possible to confirm that the novel herpesvirus DNA sequences detected here represent EVEs rather than DNA derived from contaminating exogenous viruses.

DNA derived from alloherpesviruses (family *Alloherpesviridae*) was detected in fish and amphibians. In ray-finned fish, most of these sequences belonged to the 'teratorn' lineage of transposable elements, which have arisen via fusion of alloherpesvirus genomes and piggyBac transposons, and have been intragenomically amplified in the genomes of teleost fish (Infraclass Teleostei) [39]. Additional alloherpesvirus-related elements were identified in three amphibian species and five ray-finned fish species (**Table S3**). One of these elements, identified in the Asiatic toad (*Bufo gargarizans*) occurred within a contig that was significantly larger than a herpesvirus genome, demonstrating that it represents an EVE rather than an exogenous virus. Phylogenetic analysis revealed that alloherpesvirus-like sequences identified in amphibian genomes clustered robustly with amphibian alloherpesviruses, while those identified in fish genomes clustered with fish alloherpesviruses (**Fig S7c**).

5.2 EVEs derived from viruses with single-stranded DNA genomes

EVEs derived from parvoviruses (family *Parvoviridae*) and circoviruses (family *Circoviridae*) are widespread in vertebrate genomes, being found in the majority of vertebrate classes (**Fig. 4**). Both endogenous circoviral elements (ECVs) and endogenous parvoviral elements (EPVs) are only absent in major vertebrate groups represented by a relatively small number of sequenced species genomes (i.e. between 1 and 6). No ECVs or EPVs were identified in the tuatara (order Rhynchocephalia) or in crocodiles (order Crocodilia). EPVs were not identified in agnathans, while ECVs were not identified in cartilaginous fish.

We identified a total of 1192 ECVs, most of which derived from an element in carnivore (Class Mammalia: order Carnivora) genomes that is embedded within a non-LTR retrotransposon and has undergone intragenomic amplification (**Fig. S8**). While many of the ECVs identified in our have been reported in previous publications [7, 26, 45-47], we also identified novel loci in mammals, reptiles, amphibians, and ray-finned fish (**Table S4**). Phylogenetic analysis (see **Fig. S7d**) revealed that a novel ECV locus in turtles was found to group with avian circoviruses, while amphibian ECV elements grouped with fish circoviruses, though bootstrap support for this relationship was lacking. A circovirus-like sequence detected in the WGS data of Allen's wood mouse (*Hylomyscus alleni*) grouped robustly with the exogenous 'rodent circovirus', but integration of this sequence into the *H. alleni* genome could not be confirmed.

We identified 627 EPVs, representing two distinct subfamilies within the *Parvoviridae* and five distinct genera (see **Fig. 4**). The majority of these loci have been reported in a previous study of vertebrate genomes [47] or were related to these loci. However, we also identified novel EPVs in reptiles, amphibians and mammals (**Table 2, Table S5**). In reptiles the novel elements derived from genus *Dependoparvovirus* while the amphibian elements were more closely related to viruses in genus *Protoparvovirus*. Interestingly, these EPVs clustered basally within a clade of protoparvovirus-related viruses in phylogenetic reconstructions (**Fig. S7e**), consistent with previous analyses indicating that this genus may have broadly co-diverged with vertebrate phyla [47].

5.3 EVEs derived from reverse-transcribing DNA viruses

EVEs derived from hepadnaviruses (family *Hepadnaviridae*), which are reverse-transcribing DNA viruses, were identified in reptiles, birds and amphibians (**Table 2, Table S6**). Most of these EVEs, which are commonly referred to as 'endogenous hepatitis B viruses' (eHBVs), have been reported previously [48, 49]. However, we identified novel elements in the plateau fence lizard (*Sceloporus tristichus*), and also in vertebrate classes where they have not been

reported previously. These include one eHBV identified in a cartilaginous fish, the Australian ghostshark (*Callorhinchus milii*), and another identified in an amphibian, the common coquí (*Eleutherodactylus coqui*).

Phylogenetic analysis (see **Fig. S7f**) revealed that novel eHBV elements identified in lizards (suborder Lacertilia) group robustly with the exogenous skink hepadnavirus (SkHBV), while the amphibian element groups with robustly within a clade comprised of the exogenous spiny lizard hepadnavirus (SlHBV), Tibetan frog hepadnavirus (TfHBV) and eHBV elements identified in crocodile genomes. Unfortunately, the shark eHBV was relatively short and not amenable to phylogenetic analysis.

5.4 EVEs derived from viruses with single-stranded, negative sense RNA genomes

Screening revealed that vertebrate genomes contain numerous EVEs derived from mononegaviruses (order *Mononegavirales*), which are characterised by non-segmented ssRNA-ve genomes. These EVEs derive from four mononegavirus families: bornaviruses (family *Bornaviridae*), filoviruses (family *Filoviridae*), paramxyoviruses (family *Paramyxoviridae*) and chuviruses (family *Chuviridae*) (**Fig. 4, Table 2, Table S7-9**). We did not detect any EVEs derived from other mononegavirus families that infect vertebrates (*Pneumoviridae*, *Rhabdoviridae*, *Nyamiviridae*, *Sunviridae*), nor any EVEs derived from virus families with segmented, negative sense RNA genomes (e.g., *Peribunyaviridae*, *Orthomyxoviridae*).

The majority of mononegavirus EVEs identified in our screen were derived from bornaviruses and filoviruses and have been described in previous reports [7, 45, 47, 48, 50]. However, we also identified novel EVEs derived from these groups, as well as previously unreported EVEs derived from paramyxoviruses and chuviruses (**Table 2**). Germline integration of DNA derived from mononegaviruses can occur if, in an infected germline cell, viral mRNA sequences are reverse transcribed and integrated into the nuclear genome by

cellular retroelements [51]. EVE loci generated in this way preserve the sequences of individual genes of ancient mononegaviruses, but not entire viral genomes. Among mononegavirus-derived EVEs, regardless of which family, EVEs derived from the nucleoprotein (NP) and large polymerase (L) genes predominate, but other genes are also represented, including the glycoprotein (GP) genes of filoviruses, bornaviruses, and chuviruses, the VP30 and VP35 genes of filoviruses, and the hemagglutinin-neuraminidase (HA-NM) gene of paramyxoviruses.

Paramyxovirus-like EVEs were identified in ray-finned fish, amphibians and sharks (**Fig. 4, Table 2**, **Table S7**). Many of these EVEs were highly divergent and/or degenerated and consequently their evolutionary relationships to contemporary paramyxoviruses were poorly resolved in phylogenetic analysis. However, an L polymerase-derived sequence identified in the pobblebonk frog (*Limnodynastes dumerilii*) genome was found to group robustly with sunshine virus, a contemporary paramyxovirus of Australian pythons [52], in phylogenetic trees (**Fig. S7g**).

Chuvirus-like sequences were identified in agnathans, ray-finned fish, reptiles, and mammals (**Fig. 4**, **Table 2**, **Table S8**). The majority of the mammalian elements were identified in marsupials, but we also identified a single chuvirus-like EVE in the genome of a laurasiatherian mammal – the bottlenose dolphin (*Tursiops truncatus*). Phylogenetic trees reconstructed using alignments of NP-derived chuvirus EVEs and NP genes of contemporary chuviruses revealed evidence for the existence of distinct clades specific to particular vertebrate classes (**Fig. S7h**). These included a clade including both a snake EVE and an exogenous chuviruses of snakes, two clades comprised of EVEs and viruses of teleost fish. In addition, these phylogenies revealed a robustly supported relationship between chuvirus EVEs in the Tibetan frog (*Nanorana parkeri*) and zebrafish (*Danio rerio*) genomes. Taken together, these results provide evidence for the existence of numerous diverse lineages of vertebrate chuviruses.

Filovirus-derived EVEs were mainly identified in mammals (**Fig. 4**, **Table 2**, **Table S9**). However, we also identified one filovirus-derived EVE in an amphibian – the mimic poison frog (*Ranitomeya imitator*) - providing the first evidence that filoviruses infect this vertebrate group (**Table 1**). We identified novel, ancient filovirus EVEs in anteaters (family Myrmecophagidae) and spiny mice (genus *Acomys*).

Strikingly, the inclusion of Tapajos virus (TAPV), a snake filovirus, in phylogenetic reconstructions revealed evidence for the existence of two highly distinct filovirus lineages in mammals (**Fig. 5**). These two lineages, which are robustly separated from one another by TAPV, are evident in phylogenies constructed for both the NP and VP35 genes. One lineage (here labelled 'Mammal-1') is comprised of EVEs and all contemporary mammalian filoviruses, whereas the other ('Mammal-2') is comprised exclusively of EVEs. Notably, within the Mammal-1 group, EVEs identified in mammals indigenous to Southern Hemisphere continents (e.g. marsupials, xenarthrans) group basally, whereas EVEs and viruses isolated from 'Old World'-associated placental mammals occupy a more derived position.

The 'Mammal-2' clade contains filovirus EVEs from rodents, primates and bats. Because EVEs belonging to this clade were obtained from several distinct lineages, and show conservation across these groups, we can be reasonably confident they represent a *bona fide* lineage within the *Filoviridae*, rather than just a set of highly degraded filo-like EVEs that group together due to long branch attraction [53]. One member of this group (eflp-filo.1-Myotis) encodes an intact VP35 protein, the properties of which have been experimentally investigated in recent studies [54, 55]. Interestingly, we found that spiny mice also harbour a filovirus EVE encoding an intact VP35 protein (eflp-filo.3-Acomys), however, this insertion belongs to the 'Mammal 1' clade and is relatively closely related to the VP35 proteins found in contemporary mammalian filoviruses (**Fig. 5b**).

Bornavirus-like EVEs were identified in all vertebrate classes except Chondrichthyes (**Fig. 4**, **Table 2**, **Table S10**). The majority have been reported previously or are orthologs of previously reported EVEs. However, we identified novel bornavirus-like EVEs in the genomes of ray-finned fish, amphibians and reptiles. In fish and amphibians, these EVEs disclosed similarity to culterviruses (genus *Cultervirus*) whereas in reptiles they disclosed similarity to orthobornaviruses (genus *Orthobornavirus*).

5.5 EVEs derived from viruses with single-stranded, positive sense RNA genomes

EVEs derived from positive sense RNA viruses are rare in vertebrate genomes (**Fig. 4, Table 2**, **Table S11**). The only examples we identified were a small number of sequences derived from flavivirids (family *Flaviviridae*). These include an EVE derived from the *Pestivirus* genus of flavivirids, the reference genome of the Indochinese shrew (*Crocidura indochinensis*), as reported previously [56], and EVEs identified in ray-finned fish, also reported previously [57]. In fish genomes, flavivirid EVEs derive from the proposed 'Tamanavirus' genus, and a lineage labelled 'X2' that groups as a sister taxon to the proposed 'Jingmenvirus' genus. However, jingmenviruses are actually segmented, RNAss-ve viruses whose genomes include flavivirid-derived segments [58]. Since is possible that the X2 lineage shares a common RNAss-ve ancestor with jingmenviruses, EVEs belonging to this lineage may in fact be derived from viruses with ssRNA-ve genomes.

6. Frequency of germline incorporation events across distinct vertebrate phyla

We used the DIGS framework to dissect the history of horizontal gene transfer events involving germline incorporation of DNA derived from non-retroviral viruses. We excluded EVEs derived from DNAds viruses from this analysis because most of these are mavericks/polintons or teratorn elements that have undergone intragenomic amplification. For these groups, the large number of insertions, and the fact that amplified lineages appear to have been independently established on multiple occasions, meant that such an analysis

would be beyond the scope of this study. Furthermore, for most of the few dsDNA-derived EVEs that did not belong to these groups, it was not possible to determine if they represented germline-integrated elements, exogenous viruses, or integrations occurring in somatic cells.

To examine the rate of germline incorporation in the remaining groups of non-retroviral EVEs representing DNAss, DNArt, RNAss-ve and RNAss+ve viruses, we compiled an expanded RSL containing a single reference sequence for each putative (or previously confirmed) ortholog. By classifying our hits against this expanded RSL, we could discriminate novel EVE loci (paralogs) from orthologs of previously described EVE loci. Where novel paralogs were identified, we incorporated these into our RSL and then reclassified related sequences in our screening database against this updated library. By investigating loci in this way, and iteratively reclassifying database sequences, we progressively resolved the various non-retroviral EVEs identified in our screen into sets of putatively orthologous insertions. Via this analysis we estimated that the non-retroviral EVEs identified in our study (excluding those derived from dsDNA viruses) represent ~1137 distinct germline incorporation events (**Table 2**). Using information about orthology across vertebrate hosts, we calculated minimum age estimates for all non-retroviral EVEs identified in two or more species (**Table S12**). We applied standardised nomenclature to EVE loci (see **Methods**), capturing information about EVE orthology, taxonomy, and distribution that was obtained in our study (**Table S3-11**).

Next, we estimated the rate of germline incorporation for each endogenized virus family, in all vertebrate classes represented by at least ten species (**Fig. 6**). Rates were found vary dramatically across each of the vertebrate groups examined. Overall, rates were highest in mammals, and lowest in reptiles. Fish and amphibians disclosed similar rates with ssDNA and ssRNA-ve viruses being incorporated at similar, intermediate rates. Birds were generally similar to reptiles, but show a a higher rate of ssDNA virus incorporations and a markedly elevated rate of hepadnavirus incorporation. Rates of parvovirus, filovirus, and bornavirus

infiltration were very high in mammals compared to other vertebrate classes. with bornaviruses being incorporated at a particularly high rate (>0.03 per million years of species evolution)elevated rate of incorporation for the DNArt hepadnaviruses. A relatively high rate of incorporation of RNAss+ve viruses was observed in ray-finned fish, but since the elements in question are closely related to jingmenviruses, as described above, they may in fact reflect incorporation of DNA derived from an RNAss-ve virus group [58].

In addition to estimating the frequency of germline incorporation of non-retroviral viruses, we used our screening data to reconstruct a time-calibrated overview of virus integration throughout vertebrate evolutionary history (**Fig. 7, Table S12**). Among putatively orthologous groups of EVEs for which we were able to estimate minimum dates of integration, the majority were found to have been incorporated in the Cenozoic Era (1-66 Mya). So far, the oldest integration event identified involves a metahepadnavirus (genus *Metahepadnavirus*)-derived EVE that appears to be orthologous in tuataras and birds, indicating it was incorporated into the saurian germline >280-300 Mya (see [48]). Other ancient EVEs include circovirus and herpetohepadnavirus (genus *Herpetohepadnavirus*)-derived EVEs in turtles (order Testudines) (see [49]), a circovirus-derived EVE in frogs (order Anura), and bornavirus integrations in placental mammals (see [50]). Besides revealing the landscape of non-retroviral EVE integration throughout vertebrate history, plotting EVE distribution in this way clearly reveals the main differences in EVE distribution across host groups (**Fig. 7**).

DISCUSSION

Sequencing of genomes is advancing rapidly but deciphering the complex layers of information they contain is a challenging, long-term endeavour [58, 59]. Genomes are not only inherently complex, they also exhibit remarkable dynamism, with phenomena such as recombination, transposition and horizontal gene transfer contributing to the creation of genomic ‘churn’ that makes feature distribution difficult to map [60]. These issues, combined with rapid data accumulation, coverage limitations, and assembly errors – make generation

of complete and accurate annotations difficult [62, 63]. Consequently, labour intensive manual genome annotation remains important [58, 61], and most published whole genome sequences are comprised of genomic 'dark matter'.

An exciting aspect of these circumstances is that they provide immense scope to make interesting biological discoveries using low cost, *in silico* approaches. While experimental studies are generally required to characterise genome features at a functional level, approaches based solely on comparative sequence analysis (see **Fig 1b**) can often reveal useful insights into their biology and evolution [1, 59]. Furthermore, comparative investigations *in silico* can often be productively combined with functional genomics or experimental approaches (**Fig 1b, Table 1**).

Systematic *in silico* genome screening is computational approach that facilitates investigation of the dark genome (**Fig. 1**). However, it can be challenging to implement efficiently. Automated pipelines are generally required to implement large-scale screens [60], and these can produce copious output data that are difficult to manage and interpret without an appropriate analytical framework. In this report, we introduce DIGS – a robust analytical platform for conducting large-scale *in silico* screens - and describe an open software framework (the DIGS tool) for implementing it.

EVEs constitute one interesting and informative group of genome features that can be found within the dark genome [22]. They are poorly annotated for several reasons. Firstly, they arise sporadically via horizontal gene transfer, and consequently their distribution is unpredictable [7, 22]. Additionally, uncharacterised EVE loci may be hard to recognise due to their being highly degraded or fragmented, or because their exogenous virus counterparts are either unknown or extinct [61, 62]. Finally, there are numerous potential sources of confounding or artefactual results that can arise during EVE screening, including host genes

that exhibit similarity to virus genes, and contamination of WGS assemblies with DNA derived from exogenous viruses.

To illustrate how DIGS facilitates identification and characterisation of features hidden within the dark genome, we use the DIGS tool to perform a broad-based investigation of EVE diversity in vertebrates. We first focussed on high-copy number EVEs - which in vertebrate genomes mainly comprise ERVs. We screened 874 vertebrate genomes for RT-encoding ERVs and identified 700,000 high confidence matches. This screen revealed marked differences in ERV RT copy number between vertebrate classes. An in-depth investigation of ERV diversity in vertebrates – for example, examining their composition in finer detail, or incorporating insertions that lack RT sequences, was considered beyond the scope of this study. However, the RT dataset generated here provides a robust foundation for further ERV studies that are underpinned by phylogenetic analysis. For example, we have previously used RT data in combination with other *in silico* approaches for in-depth, phylogenetical characterisation of ERVs within discrete mammalian subgroups (e.g. see [63]).

ERVs constitute a unique type of EVE, in that they can remain replication-competent following integration and may increase their germline copy number through continued virus replication. However, the germline copy number of any EVE can potentially increase through interactions with TEs - this has been described for ERVs [64-66], as well as for EVEs derived from dsDNA viruses [38, 39, 67]. In addition, data obtained here and in our previous investigations show that EVEs derived from hepadnaviruses have been amplified in cormorants [48], while circovirus-derived sequences have been amplified in carnivore genomes [45], apparently in association with LINE1 activity (**Table S4**). These findings underline the impact of fusion between EVEs and vertebrate transposons on vertebrate genome evolution. This phenomenon has occurred on multiple independent occasions and involved a diverse range of vertebrate viruses. Interestingly, we found that circovirus EVEs in carnivore genomes are associated with a retroelement lineage (LINE1) that has also

inserted into gammaherpesvirus and Chikungunya virus genomes (**Fig. S8**). These findings indicate that retroelement-mediated transposition has established a complex network of horizontal transfer events linking diverse host and virus genomes.

DIGS is not only well-suited to exploring the distribution and diversity of high copy number genome features such as ERVs and TEs, it can also be used in 'beach combing' searches that aim to identify rare and unusual genome features. These kinds of screens typically require a rigorous filtering process to distinguish genuine from spurious matches, and as shown here, this is facilitated by database integration. DIGS enabled the efficient identification of EVEs derived from non-retroviral viruses (which are relatively rare and diverse) and provided a powerful framework for filtering spurious results (**Fig. S3**).

Via DIGS, we established a broad overview of non-retroviral EVE diversity in vertebrate genomes (**Fig. 4**, **Fig. 6**), shedding new light on virus distribution and diversity in vertebrates. Notably, our findings extend the known host range of important virus families. For example, we identify a filovirus-derived EVE in a frog (order Anura), providing the first evidence for the existence of amphibian filoviruses. In addition, we provide the first evidence for the presence (at least historically) of hepadnaviruses in sharks, and chuviruses in mammals (**Fig. 4**). In addition, we reveal novel virus diversity. For example, we identify novel lineages of parvoviruses and circoviruses in amphibians (**Fig S7d-e**), as well as a novel circovirus lineage in turtles (**Fig S7d**) and a novel hepadnavirus lineage in frogs (**Fig S7f**). We also identify novel paramyxovirus and chuvirus lineages in fish and amphibians (**Fig. S7g-h**).

Mammalian filoviruses include some of the most lethal viruses in the world [68], and while the natural reservoirs of some are known, they remain unclear for the highly pathogenic ebolavirus (EBOV) and its closest relatives (**Fig. 5**). EBOV is assumed to have a zoonotic origin, but it has rarely been possible to formally link outbreaks to a given animal reservoir,

limiting understanding of its emergence. So far, efforts to identify the true reservoirs of ebolaviruses have tended to focus on bats [69]. However, the widespread presence of filovirus EVEs in rodents (**Table S9**), including some groups that have not been examined as potential EBOV reservoirs, such as spiny mice, suggests that the potential of this group to serve as a reservoir should not be overlooked.

Previous studies have noted that filovirus EVE sequences in the genomes of cricetid rodents (family Cricetidae) robustly split the *Ebolavirus* and *Cuevavirus* genera from the *Marburgvirus* and *Dianlovirus* genera, demonstrating that these groups diverged >20 million years ago (Mya) [70], rather than within the past 10,000 years as suggested by molecular clock-based analysis of contemporary filovirus genomes [71]. Here, we found that TAPV, an exogenous virus of snakes, robustly separates two clades of mammalian filoviruses in phylogenetic reconstructions. Since transmission of filoviruses between reptiles and mammals is likely quite rare, and both lineages contain ancient EVEs (**Fig. 5, Table S12**), these findings support the long-term existence of two highly distinct filovirus lineages ('mammal 1' and 'mammal 2') within mammals. Notably, basal taxa within the 'mammal 1' lineage – which also includes all known contemporary filoviruses of mammals – disclose associations with Southern Hemisphere continents (Australia, South America) that were largely isolated throughout extensive periods of the Cenozoic Era. These data suggest that filoviruses were present in ancestral mammals inhabiting Gondwanaland (an ancient supercontinent comprised of South America, Africa, India, and Australia) and diversified into at least two major lineages as mammalian populations became compartmentalised in distinct continental regions during the early to mid-Cenozoic. An interesting question is whether the 'mammal 2' group represents filoviruses that evolved in Northern hemisphere-associated, boreoeutherian mammals (magnorder Boreoeutheria), while 'mammal 1' represents filoviruses that initially evolved in Southern hemisphere-associated marsupials (infraclass Marsupialia) and xenarthrans (magnorder Xenarthra) before disseminating throughout the globe (possibly in association with volant mammals – i.e., bats).

While several previous studies have described EVE diversity in vertebrates [33, 36, 72], our investigation is significantly larger in scale and breadth. Furthermore, for non-retroviral viruses, we introduced a higher level of order to EVE data, making use of the DIGS framework to discriminate orthologous versus paralogous EVE loci, and to identify intra-genomically amplified EVE lineages. This allowed us to establish a panoramic view of germline incorporation by non-retroviral viruses during vertebrate evolution (**Fig. 6**). Furthermore, discriminating orthologous and paralogous EVEs enabled us to infer the rates of germline infiltration by non-retroviral virus families with greater accuracy than in previous studies (**Fig. 5**, **Fig. 6**). Notably, we did not find strong evidence for a reduced rate of germline infiltration in avian genomes, as suggested by a previous study [73]. Incorporation of DNArt viruses is higher in birds than in any other vertebrate class (**Fig. 5**), while acquisition of EVES derived from ssDNA-ve viruses does appear to be limited in this group, they closely resemble reptiles in this respect. Furthermore, avian hosts appear similar to reptiles with regard to ERV RT copy number (**Fig. 3**).

The absence, or near absence, of many virus groups from our catalogue of vertebrate EVEs is noteworthy. For example, vertebrates are infected with many ssRNA+ve viruses, but EVEs derived from these viruses are extremely rare, while EVEs derived from viruses with circular RNA genomes, or double-stranded RNA genomes, were not detected at all (**Table 2**). All other virus genome types were represented by EVEs in the vertebrate germline, but their distribution is patchy and limited to a relatively small number of virus families (**Fig. 4**, **Fig 7**). For example, among ssRNA-ve viruses, only mononegaviruses were found to be present – we found no evidence for germline integration of segmented ssRNA-ve viruses such as orthomyxoviruses and bunyaviruses. The scarcity of EVEs derived from these virus groups suggests that aspects of their biology strictly limit their capacity to for germline incorporation. These likely include cell tropism (whether germline cells are typically infected) and the site of cellular replication (with viruses replicating in the nucleus more likely to be

incorporated). Additionally, vertebrate germline cells may present strong intrinsic barriers to the replication of certain virus groups.

The catalogue of EVE loci generated here provides a foundation for further investigations in virology, genomics, and human health. From the virology perspective, EVEs provide information about the long-term evolutionary history viruses, which greatly influences how we understand their biology. As well as enabling future studies of vertebrate ‘paleoviruses’, the EVE catalogue can inform efforts to identify and characterise new viruses (both by providing ecological and evolutionary insights [46], and by helping identify 'false positive' hits arising from genomic DNA). From the genomics side, EVEs are of interest due to their important roles in physiology and genome evolution [74]. These include roles antiviral immunity [11, 75, 76], and a diverse range of other physiological processes [18, 54, 55, 77-79]. Notably, we identified numerous non-retroviral EVEs encoding ORFs longer than 300aa (**Fig. S6**), indicating that their coding capacity has been conserved during vertebrate evolution. One of these - a chuvirus-derived L-protein identified in livebearers – adds to previous evidence that viral RdRp sequences have been co-opted by vertebrate genomes [40]. Mapping of EVE loci can also inform efforts to develop new medical treatments - in a recent study, EVE loci identified using DIGS were used to identify potential genomic safe harbours for human transgene therapy applications [80].

The EVE screen performed here has several important limitations. Firstly, it relied on published WGS data generated for extant species. Secondly, our results have likely been influenced by aspects of our screening configuration, such as the composition of the probe set with respect to viral taxa and polypeptide probe length [81, 82]. This might mean that we failed to detect some of the potentially recognisable EVE loci present in our TDb. For example, counts of RT-encoding ERV loci were found to be generally lower in ray-finned fish and jawless fish (**Fig. 3**), but previous studies have shown that RT loci related to other families of reverse-transcribing virus, such as metaviruses (family *Metaviridae*) [83] and

‘lokiretroviruses’ [84] are relatively common in these hosts. These would likely have been missed in our search because they were not included in our RT RSL. Finally, previous studies have indicated that vertebrate genomes contain EVEs that lack any clear homology to extant viruses [85], and these would not be detected using a sequence similarity-based approach.

As vertebrate genome sequencing progresses, further opportunities to identify novel EVEs will arise, since: (i) any novel genome could in theory contain a lineage-specific EVE, and; (ii) ongoing characterization of exogenous virus diversity may allow for detection of previously undetectable EVEs. The DIGS project created here, which is openly available online, can be reused to accommodate newly sequenced vertebrate genomes (expanded TDb) and newly discovered vertebrate virus diversity (expanded RSL/probe set). In addition, similar projects can readily be created to screen for EVEs in other host groups.

The use of DIGS is not limited to EVEs – it can be used to investigate any sufficiently conserved genome feature lurking within the dark genome, including both coding and non-coding elements (**Table 1**). Importantly, some of the most important and interesting genes have evolved relatively rapidly and are highly duplicated [86]. Such gene families may be unreliably annotated by automated pipelines, because increases in copy number have been followed by a relaxation of purifying selection, leading to mutational decay [87]. DIGS is particularly well-suited to investigations of genome features that comprise a single contiguous unit >100 nucleotides in length, such as EVEs, TEs and processed pseudogenes, since these tend to produce unambiguous matches in sequence similarity searches. However, DIGS can also be used to investigate genome features that are shorter, or spliced genes that are split into multiple, short exons, providing a careful approach is used. In these kinds of screens, inclusion of conserved flanking features can be used help discriminate false positive matches.

The framework described here or implementing DIGS could be further developed and improved. For example, by including the option to use of other search tools, such as Diamond [88] and ElasticBLAST [89], or by integrating with functional genomics resources [90].

CONCLUSIONS

We demonstrate how a relational database management system can be linked to a similarity search-based screening pipeline to investigate the dark genome *in silico*. Using this approach, we catalogue and analyse EVEs throughout vertebrate genomes, providing a broad range of novel insights into the evolution of ancient viruses and their interactions with host species.

MATERIALS & METHODS

Whole genome sequence and taxonomic data

Whole genome shotgun (WGS) sequence assemblies of 874 vertebrate species were obtained from the NCBI genomes resource [91]. Taxonomic data for the vertebrate species included in our screen and the viruses in our reference sequence library were obtained from the NCBI taxonomy database [92], using PERL scripts included with the DIGS tool package.

Database-integrated screening for RT-encoding ERVs

An RT RSL was colllated represent diversity within the *Retroviridae*. We included representatives of previously identified ERV lineages and exogenous retrovirus species. A subset of these sequences was used as probes in similarity search-based screens [93]. For initial screening we used a bitscore cut-off of 60. For comparisons of ERV RT copy number across species we filtered initial results using a more conservative bitscore cut-off of 90. Our previous, DIGS-based studies of ERVs have shown that spurious matches (i.e. to sequences other than retroviral RTs) do not arise when this cut-off is applied, although some genuine ERV RT hits may be excluded [63].

Database-integrated screening for non-retroviral EVEs

We obtained an RSL representing the proteome of eukaryotic viruses from the NCBI virus genomes database [37]. We supplemented this with sequence likely to cross-match to virus probes during screening. These included the teratorn transposon found in fish, which is known to contain multiple alloherpesvirus-derived genes [94]. We included the polypeptide sequences of these genes, obtained from the subtype 1 Teratorn reference (Accession #: LC199500) in our RSL. We also included representatives of the maverick/polinton lineage of transposons, derived from sequences defined in a prevous study [95]. Since these element derive from a group of midsize eukaryotic linear dsDNA (MELD) viruses provisionally named ‘adintoviruses’ [67]. Probes constituted a subset of 685 sequences contained within our RSL, and incorporated polypeptide sequences representing all major protein-coding genes

of representative species of all recognised vertebrate virus families. We also included representative sequences of maverick/polinton elements in our probe set. We used a bit score cut-off of 60 as a threshold for counting non-retroviral EVE loci. This threshold was established through previous experience searching for non-retroviral EVEs using DIGS [45, 47, 48, 57]. Experience from previous studies had shown that nearly 100% of matches with bit scores >= 60 were either virus-derived or represented genuine similarity between virus genes and their cellular orthologs. By contrast, investigation of a subset of 100 hits with bit scores of b 40-59 showed that ~50% could not be confidently confirmed as having a viral origin (data not shown).

Artefactual hits to host DNA can occur since some virus genomes contain genes that have cellular homologs [96], some virus genomes contain captured host DNA [97]. To distinguish host from virus-derived DNA in these cases, we exported such hits from the screening database and virtually translated them to obtain a polypeptide sequence. We then used the translated sequences as query input to online BLAST searches of GenBank's non-redundant (nr) database. If searches revealed closer matching to host genes than to known viral genes, the input sequences were assumed to be host derived. Wherever this occurred we incorporated representatives of the matching host sequences into the RSL, so that they would be recognised as host hits on reclassification. By updating hit classifications in this way, we could progressively filter out host-derived hits from our final screening output.

Filtering sequences-derived from exogenous virus

Sequences derived from exogenous viruses are occasionally incorporated into WGS assemblies. We used SQL queries to identify and exclude these sequences based on hit characteristics. Where hits derived from virus species or species groups that have been sequenced previously, they could be discriminated on the basis of sequence identity (i.e., 98-99% nucleotide-level identity known viruses. The 'extract start' field could be used to identify sequences that lacked flanking genomic sequences, indicating a potential

exogenous origin. We also examined the virtually translated sequences to look for evidence of long-term presence in the host germline (e.g., stop codons, frameshifting mutations).

Filtering of cross-matching retrovirus-derived sequences

Hits that match more closely to virus genomes than to host DNA, and are clearly inserted into host DNA, are most likely *bona fide* EVE sequences. However, they may not necessarily be non-retroviral EVEs, because some filoviruses and arenaviruses (family *Arenaviridae*) contain glycoprotein genes that are distantly related to those found in certain retroviruses [98, 99]. When such hits were investigated and found to correspond to ERVs (established through the presence of proviral genome features adjacent to the hit) we included the putative sequences of glycoproteins encoded by these ERVs into our RSL and reclassified hits, so that spurious matches could be recognized as ERV-derived.

Genomic analysis

Previous studies of presence/absence patterns have shown that non-retroviral EVEs are present in many genomes due to orthology (ancient insertions) rather than paralogy (recent independent insertion) [45, 47-49]. To differentiate orthologs of previously described EVEs from newly identified paralogs, we substituted expanded our RSL to include consensus/reference sequences representing unique EVE loci. This set of EVE loci was comprised of insertions identified in previous studies [45, 47, 48, 50, 100], as well as a set of clearly novel EVEs identified in the present screen. For high-copy number, amplified lineages within this set, we only included a single reference sequence, rather than attempting to represent each individual ortholog, since it was clear that these elements derive from a single germline incorporation event (see **Fig. S8**). EVEs were considered novel if: (i) they derived from a virus group not previously reported in the host group in which they were identified, or (ii) occurred in species only distantly related to species in which similar EVEs had been identified previously (e.g. an entirely distinct host class). Whenever novel EVEs were defined, results were reclassified using the updated RSL (see **Fig. 2**).

Orthologs of previously identified EVEs could be inferred by using SQL queries to summarise screening results, as they disclosed high similarity to these EVE sequences and occurred in host species relatively closely related to the species in which the putatively orthologous EVEs had previously been identified. By contrast, novel paralogs either disclosed only limited similarity to previously identified EVE sequences or occurred in distantly related host species. Minimum age estimates were obtained for orthologous EVEs by using host species divergence time estimates collated in TimeTree [101]. This approach to discriminating between paralogs and orthologs has limitations, but can guide further investigations using more reliable approaches (e.g. via investigation of flanking sequences, or phylogeny) [48].

Phylogenetic analysis

Phylogenies were reconstructed using the maximum likelihood approach implemented in RAxML (version 8.2.12) [102]. Protein substitution models were selected via hierarchical maximum likelihood ratio test using the PROTAUTOGAMMA option in RAxML. Se-Al (version 2.0a11) was used to inspect MSAs.

Application of standardised nomenclature to EVE loci

We assigned all non-retroviral EVEs identified in our study unique identifiers (IDs), following a convention developed for ERVs [103], Each was assigned a unique identifier (ID) constructed from three components. The first component is a classifier denoting the type of EVE. The second component comprises: (i) the name of the taxonomic group of viruses the element derived from and; (ii) a numeric ID that uniquely identifies a specific integration locus, or for multicopy lineages, a unique founding event. The final component denotes the taxonomic distribution of the element. This approach has been applied in several previous studies of vertebrate EVEs [47, 48, 50, 57]. Where appropriate, we maintained consistency with previous studies.

FIGURE LEGENDS

Figure 1. Exploring the dark genome using in silico screening.

Sequence similarity search-based screening aims to identify and classify the sequences similar to a set of query sequences, so that these can be examined using comparative approaches. It entails systematically screening a set of target databases using probes selected from a user-composed 'reference sequence library' (RSL). The schematic in panel **(a)** shows one round of screening. (i) A 'reference sequence library' (RSL) curated to represent the known diversity of the genome feature(s) under investigation; (ii) A subset of sequences from the RSL are selected for use as probes; (iii) Each probe sequence is used to screen each target file in turn; (iv) Matching sequences identified in this screen are compared to sequences in the RSL, again using BLAST; The output of this process is a comprehensive set of matching sequences, each associated with a classification or 'genotype' and associated metrics based on pairwise comparison to the RSL. As shown in panel **(b)**, these data can be investigated using a wide range of in silico, comparative approaches: (i) analysis of feature distribution – e.g. annotating host phylogeny to show frequency (coloured circles). (ii) phylogenetic screening, in which sequences obtained via similarity search-based screening are investigated phylogenetically (iii) Pairwise sequence comparisons. (iv) Comparative phylogenetic analysis. The genetic properties of novel homologs can be inferred via comparative analysis (e.g. pairwise comparisons), while their phenotypic properties can potentially be investigated experimentally (e.g., via transcriptome sequencing).

Figure 2. The database-integrated genome screening (DIGS) process.

On initiation of screening a list of searches (i.e. each query sequence versus each TDb file) is composed based on the probe and TDb paths supplied. Subsequently, screening proceeds systematically as shown, via the following steps: (i) the status table of the project-associated screening database is queried to determine which searches have yet to be

performed. if there are no outstanding searches then screening ends, otherwise it proceeds to the next outstanding search as defined by a probe-TDb file pair. (iii) the search is performed using the appropriate BLAST+ program – i.e. BLASTn for nucleotide sequence probes, tBLASTn for polypeptide probes – and results are recorded in the data processing table ('active set'); (iv) Results in the processing table are compared to those (if any) obtained previously to derive a non-redundant set of non-overlapping loci, and an updated set of non-redundant hits is created, with the results from this round included, and each hit being represented by a single results table row. To create this non-redundant set, hits that overlap, or occur within a given range of one another, are merged to create a single entry; (v) The nucleotide sequences associated with results table rows are extracted from TDb files, classified via BLAST-based comparison to the RSL. For non-coding features classification is based on BLASTn whereas for coding features it is based on BLASTx. (vi) The header-encoded details of the best-matching sequence (species name, gene name) are recorded in the results table along with the sequence itself, and the associated data from BLAST based comparison; (vii) finally, the status table is updated to create a record of the search having been performed, and the next round of screening is initiated.

Figure 3. Counts of ERV RT loci identified by identified via database integrated genome screening of 874 vertebrate species.

Box plots showing the distribution of endogenous retrovirus (ERV) reverse transcriptase (RT) counts in distinct vertebrate classes. Median and range of values are indicated. Circles indicate counts for individual species. Counts are shown against a log scale. Figure created in R using ggplot2 and geom_boxplot.

Figure 4. Distribution of virus families across vertebrate hosts. Circles indicate the presence of exogenous viruses. Shaded boxes indicate the presence of confirmed endogenous viral elements. Abbreviations: DNArt = reverse transcribing DNA viruses; DNAss = single-stranded DNA viruses; DNAds double-stranded DNA viruses; RNAds =

double-stranded RNA viruses; RNAss-ve = single-stranded negative sense RNA viruses; RNAss+ve = RNAss-ve = single-stranded positive sense RNA viruses. * Only one family of RNAss+ve is shown, since all other families with single-stranded, positive sense RNA genomes infecting vertebrates do not occur as EVEs. Information on the distribution of exogenous viruses was obtained from the NCBI virus genomes resource [37], supplemented with information from published papers [104-109].

Figure 5. Evolutionary relationships of filoviruses and filovirus-derived EVEs

Bootstrapped maximum likelihood phylogenies showing the evolutionary relationships between filoviruses and filovirus EVEs in the nucleoprotein (NP) and viral protein 35 (VP35) genes. Phylogenies were constructed using maximum likelihood as implemented in RAxML, and codon-aligned nucleotides for each gene. Numbers adjacent internal nodes indicate bootstrap support (1000 bootstrap replicates). The scale bar indicates evolutionary distance in substitutions per site. Virus taxon names are shown in regular font, EVE names are shown bold. EVE names follow standardised nomenclature (see Methods). Brackets to the right of each tree indicate virus genera (italics) and major lineages (bold). Silhouettes indicate host groups following the key. For Ebola virus, Bundibugyo virus, and Tai Forest virus, the main reservoir hosts are unknown. The inset box adjacent these taxa shows host species in which one or more of these viruses has been isolated [69, 110], following the key.

Figure 6. Comparison of germline infiltration rates in five vertebrate classes.

Infiltration rates represent the rate of incorporation and fixation per million years (MY) of species branch length sampled. Rates are shown each non-retroviral family represented by vertebrate EVEs. Colours indicate reverse transcribing DNA (DNArt) viruses, single stranded DNA (DNAss) viruses, single stranded negative sense RNA (ssRNA-ve) virueses and single stranded positive sense RNA (ssRNA-ve) virueses, following the key. Total branch length sampled in each vertebrate class was derived from time-calibrated phylogenies obtained from TimeTree [101]).

Figure 7. Overview of germline incorporation in vertebrates.

A time-calibrated phylogeny of vertebrates with genome sequences examined for this study is presented with endogenization events indicated at nodes (diamonds) or at tips (circles). Nodal events are present in descendants of that node, providing minimum ages for the age of endogenization. The number of loci within each vertebrate genome is indicated at the tips including both unique and shared endogenization events. The points at nodes and tips are scaled by the number of endogenized loci, "Count" in legend, and color-coded to virus family as indicated in legend. (a) The distribution of 10 families of viruses is shown across vertebrates separately, (b) all virus families are shown on the same tree.

Table 1. Examples of published studies utilising database-integrated screening.

Genome feature	Target database	Reference sequence library & probes*	Reference	Year
Non-coding DNA				
ZP3AR (and SFP819)	Rodents	ZP3AR*. ZFP819* & related genes	[111] [b]	2022
SHIN (and IAP elements)	Rodents	SHIN*, IAP subgroups*, *Retroviridae*	[112] [b]	2023
Genes				
OAS1 gene	Mammals	OAS1* & related genes	[113] [b]	2021
APOBEC3 (A3) genes	Mammals	APOBEC3* & related genes	[114] [a]	2020
Interferon stimulated genes (ISGs)	Vertebrates	ISGs* & related genes	[115]	2017
Interferon lambda (IFNL) genes	Vertebrates	IFNLs* & locus marker genes*	In prep [a, b]	2023
Endogenous viral elements				
Family *Flaviviridae*	Metazoa	AVP, *Flaviviridae** & EFVs	[57] [a]	2022
Family *Parvoviridae*	Vertebrates	AVP, *Parvoviridae** & EPVs	[47] [a]	2022
Family *Parvoviridae*	Vertebrates	AVP, *Parvoviridae** & EPVs	[80] [b]	2023
Genus *Protoparvovirus*	Mammals	AVP, protoparvoviruses* & EPVs	[116] [a, b]	2019
Family *Hepadnaviridae*	Metazoa	AVP, *Hepadnaviridae** & eHBVs	[48] [a]	2021
Family *Circoviridae*	Metazoa	AVP, *Circoviridae** & ECVs	[45] [a]	2019
Endogenous retroviruses				
Genus *Lentiretrovirus*	Rodents	Lentiviruses*, other XRVs, & ERVs	[117] [a]	2022
Family *Retroviridae*	Perissodactyls	*Retroviridae*,* Retroelements	[63] [a]	2018
HERV-T	Hominids	Class I HERVs*, *Retroviridae*	[118] [a, b]	2017
MuERV-L	Mice	Class III ERVs*, *Retroviridae*	[119] [b]	2018

Footnote: [a)] DIGS was used as part of 'phylogenetic screening' pipeline (see **Fig 1c**) [b)] DIGS-based investigations were allied to experimental or functional genomics investigations. * Indicates subset of the RSL from which probes were derived (note that *Retroviridae* here denotes both endogenous and exogenous retroviruses). **Abbreviations:** zinc finger protein (ZFP); 2'-5'-oligoadenylate synthetase 1 (OAS1); intracisternal A-type particle (IAP); endogenous flaviviral element (EFV); endogenous parvoviral element (EPV); endogenous hepadnavirus (eHBV); endogenous circoviral element (ECV). Human endogenous retrovirus (HERV). Murine endogenous retrovirus (muERV); NCBI all virus proteins set (AVP).

Table 2. Number of non-retroviral EVE sequence identified and estimated number of germline incorporation events in distinct vertebrate classes.

Virus Group	# EVEs identified (Estimated # germline incorporation events)															
	Total		Mammalia		Aves		Reptiles		Amphibia		Actinopterygii		Chondrichthyes		Agnatha	
ssRNA-ve																
Bornaviridae	2566	(383)	2434	(292)	30	(11)	27	(14)	52	(44)	22	(21)	-	-	1	(1)
Chuviridae	182	(164)	24	(24)	-	-	23	(23)	9	(9)	119	(108)	-	-	7	-
Filoviridae	390	(69)	389	(68)	-	-	-	-	1	(1)	-	-	-	-	-	-
Paramyxoviridae	19	(17)	-	-	-	-	-	-	4	(3)	14	(3)	1	(1)	-	-
ssRNA+ve																
Flaviviridae	8	(11)	1	(1)	-	-	-	-	-	-	7	(10)	-	-	-	-
DNArt																
Hepadnaviridae	993	(108)	-	-	897	(89)	93	(17)	2	(1)	-	-	1	(1)	-	-
DNAss																
Circoviridae	1198	(131)	918	(29)	32	(15)	91	(19)	82	(24)	68	(38)	-	-	7	(6)
Parvoviridae	689	(238)	534	(199)	34	(10)	34	(13)	12	(7)	12	(6)	3	(3)	-	-
DNAds																
Herpesviridae	13	(8)	11	(6)	1	(1)	1	(1)	-	-	-	-	-	-	-	-
Alloherpesviridae	28	(8)	-	-	-	-	-	-	15	(3)	13	(5)	-	-	-	-
Total	6087	(1137)	4311	(619)	994	(126)	269	(87)	177	(92)	255	(191)	5	(5)	15	(7)

Legend: Germline incorporation here implies both integration into the germline and fixation.

ACKNOWLEDGEMENTS

We thank Connor Bamford, Paul Bieniasz, and Jamie Henzy for helpful discussions. Additional thanks to Ade Tukuru (Aaron Diamond AIDS Research Centre) and Scott Arkinson (MRC-University of Glasgow Centre for Virus Research) for bioinformatics support. DB-M is supported by the V Foundation for Cancer Research and the Searle Scholars Program. TD, SL, JH, and RJG, were funded by the Medical Research Council of the United Kingdom (MC_UU_12014/12).

REFERENCES

1. Margulies, E.H. and E. Birney, *Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes.* Nat Rev Genet, 2008. **9**(4): p. 303-13.
2. Cheng, J.F., J.R. Priest, and L.A. Pennacchio, *Comparative genomics: a tool to functionally annotate human DNA.* Methods Mol Biol, 2007. **366**: p. 229-51.
3. Nobrega, M.A. and L.A. Pennacchio, *Comparative genomic analysis as a tool for biological discovery.* J Physiol, 2004. **554**(Pt 1): p. 31-9.
4. Guan, D. and M.A. Lazar, *Shining light on dark matter in the genome.* Proc Natl Acad Sci U S A, 2019. **116**(50): p. 24919-24921.
5. Wright, B.W., et al., *The dark proteome: translation from noncanonical open reading frames.* Trends Cell Biol, 2022. **32**(3): p. 243-258.
6. Eisenstein, M., *Drug hunters uncloak the non-coding 'hidden' genome.* Nat Biotechnol, 2021. **39**(10): p. 1169-1171.
7. Katzourakis, A. and R.J. Gifford, *Endogenous viral elements in animal genomes.* PLoS Genet, 2010. **6**(11): p. e1001191.
8. Chiba, S., et al., *Widespread endogenization of genome sequences of non-retroviral RNA viruses into plant genomes.* PLoS Pathog, 2011. **7**(7): p. e1002146.
9. Diop, S.I., et al., *Tracheophyte genomes keep track of the deep evolution of the Caulimoviridae.* Sci Rep, 2018. **8**(1): p. 572.
10. Soucy, S.M., J. Huang, and J.P. Gogarten, *Horizontal gene transfer: building the web of life.* Nat Rev Genet, 2015. **16**(8): p. 472-82.
11. Parrish, N.F. and K. Tomonaga, *Endogenized viral sequences in mammals.* Curr Opin Microbiol, 2016. **31**: p. 176-183.
12. de Tomás, C. and C.M. Vicient, *Genome-wide identification of Reverse Transcriptase domains of recently inserted endogenous plant pararetrovirus (Caulimoviridae).* Front Plant Sci, 2022. **13**: p. 1011565.
13. Gong, Z., Y. Zhang, and G.Z. Han, *Molecular fossils reveal ancient associations of dsDNA viruses with several phyla of fungi.* Virus Evol, 2020. **6**(1): p. veaa008.
14. Bellas, C., et al., *Large-scale invasion of unicellular eukaryotic genomes by integrating DNA viruses.* Proc Natl Acad Sci U S A, 2023. **120**(16): p. e2300465120.
15. Dewannieux, M. and T. Heidmann, *Endogenous retroviruses: acquisition, amplification and taming of genome invaders.* Curr Opin Virol, 2013. **3**(6): p. 646-56.
16. Geis, F.K. and S.P. Goff, *Silencing and Transcriptional Regulation of Endogenous Retroviruses: An Overview.* Viruses, 2020. **12**(8).
17. Srinivasachar Badarinarayan, S. and D. Sauter, *Switching Sides: How Endogenous Retroviruses Protect Us from Viral Infections.* J Virol, 2021. **95**(12).
18. Fujino, K., et al., *A Human Endogenous Bornavirus-Like Nucleoprotein Encodes a Mitochondrial Protein Associated with Cell Viability.* J Virol, 2021. **95**(14): p. e0203020.

19. Ophinni, Y., et al., *piRNA-Guided CRISPR-like Immunity in Eukaryotes.* Trends Immunol, 2019. **40**(11): p. 998-1010.
20. Patel, M.R., M. Emerman, and H.S. Malik, *Paleovirology - ghosts and gifts of viruses past.* Curr Opin Virol, 2011. **1**(4): p. 304-9.
21. Holmes, E.C., *The evolution of endogenous viral elements.* Cell Host Microbe, 2011. **10**(4): p. 368-77.
22. 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.
23. 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.
24. Camacho, C., et al., *BLAST+: architecture and applications.* BMC Bioinformatics, 2009. **10**: p. 421.
25. Pearson, W.R. and A.J. Mackey, *Using SQL Databases for Sequence Similarity Searching and Analysis.* Curr Protoc Bioinformatics, 2017. **59**: p. 9.4.1-9.4.22.
26. 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.
27. Heusinger, E., et al., *Early Vertebrate Evolution of the Host Restriction Factor Tetherin.* J Virol, 2015. **89**(23): p. 12154-65.
28. Blanco-Melo, D., S. Venkatesh, and P.D. Bieniasz, *Origins and Evolution of tetherin, an Orphan Antiviral Gene.* Cell Host Microbe, 2016. **20**(2): p. 189-201.
29. Waterhouse, R.M., et al., *OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs.* Nucleic Acids Res, 2013. **41**(Database issue): p. D358-65.
30. Cunningham, F., et al., *Ensembl 2015.* Nucleic Acids Res, 2015. **43**(Database issue): p. D662-9.
31. Gifford, R.J. *Database-integrated genome screening (DIGS) tool*. 2022; Available from: https://giffordlabcvr.github.io/DIGS-tool/.
32. Johnson, W.E., *Origins and evolutionary consequences of ancient endogenous retroviruses.* Nat Rev Microbiol, 2019. **17**(6): p. 355-370.
33. Hayward, A., M. Grabherr, and P. Jern, *Broad-scale phylogenomics provides insights into retrovirus-host evolution.* Proc Natl Acad Sci U S A, 2013. **110**(50): p. 20146-51.
34. Xiong, Y. and T.H. Eickbush, *Origin and evolution of retroelements based upon their reverse transcriptase sequences.* Embo j, 1990. **9**(10): p. 3353-62.
35. Tristem, M., *Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database.* J Virol, 2000. **74**(8): p. 3715-30.
36. Hayward, A., C.K. Cornwallis, and P. Jern, *Pan-vertebrate comparative genomics unmasks retrovirus macroevolution.* Proc Natl Acad Sci U S A, 2015. **112**(2): p. 464-9.
37. Brister, J.R., et al., *NCBI viral genomes resource.* Nucleic Acids Res, 2015. **43**(Database issue): p. D571-7.
38. Koonin, E.V., M. Krupovic, and N. Yutin, *Evolution of double-stranded DNA viruses of eukaryotes: from bacteriophages to transposons to giant viruses.* Ann N Y Acad Sci, 2015. **1341**(1): p. 10-24.
39. Inoue, Y. and H. Takeda, *Teratorn and its relatives - a cross-point of distinct mobile elements, transposons and viruses.* Front Vet Sci, 2023. **10**: p. 1158023.
40. Horie, M., et al., *An RNA-dependent RNA polymerase gene in bat genomes derived from an ancient negative-strand RNA virus.* Scientific Reports, 2016. **6**(1): p. 25873.
41. Ho, A.L.F.C., C.L. Pruett, and J. Lin, *Phylogeny and biogeography of Poecilia (Cyprinodontiformes: Poeciliinae) across Central and South America based on mitochondrial and nuclear DNA markers.* Molecular Phylogenetics and Evolution, 2016. **101**: p. 32-45.
42. Aswad, A. and A. Katzourakis, *The first endogenous herpesvirus, identified in the tarsier genome, and novel sequences from primate rhadinoviruses and lymphocryptoviruses.* PLoS Genet, 2014. **10**(6): p. e1004332.

43. Aswad, A., et al., *Evolutionary History of Endogenous Human Herpesvirus 6 Reflects Human Migration out of Africa.* Mol Biol Evol, 2021. **38**(1): p. 96-107.
44. Liu, X., et al., *Endogenization and excision of human herpesvirus 6 in human genomes.* PLoS Genet, 2020. **16**(8): p. e1008915.
45. Dennis, T.P.W., et al., *The evolution, distribution and diversity of endogenous circoviral elements in vertebrate genomes.* Virus Res, 2019. **262**: p. 15-23.
46. Dennis, T.P.W., et al., *Insights into Circovirus Host Range from the Genomic Fossil Record.* J Virol, 2018. **92**(16).
47. Campbell, M.A., et al., *Comparative analysis reveals the long-term co-evolutionary history of parvoviruses and vertebrates.* bioRxiv, 2022: p. 2021.10.25.465781.
48. 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.
49. Suh, A., et al., *Early mesozoic coexistence of amniotes and hepadnaviridae.* PLoS Genet, 2014. **10**(12): p. e1004559.
50. Kawasaki, J., et al., *100-My history of bornavirus infections hidden in vertebrate genomes.* Proc Natl Acad Sci U S A, 2021. **118**(20).
51. Horie, M., et al., *Endogenous non-retroviral RNA virus elements in mammalian genomes.* Nature, 2010. **463**(7277): p. 84-7.
52. Hyndman, T.H., et al., *Isolation and molecular identification of Sunshine virus, a novel paramyxovirus found in Australian snakes.* Infect Genet Evol, 2012. **12**(7): p. 1436-46.
53. Gorbalenya, A.E. and C. Lauber, *Phylogeny of Viruses.* Reference Module in Biomedical Sciences, 2017.
54. Edwards, M.R., et al., *Conservation of Structure and Immune Antagonist Functions of Filoviral VP35 Homologs Present in Microbat Genomes.* Cell Rep, 2018. **24**(4): p. 861-872.e6.
55. Kondoh, T., et al., *Putative endogenous filovirus VP35-like protein potentially functions as an IFN antagonist but not a polymerase cofactor.* PLoS One, 2017. **12**(10): p. e0186450.
56. Li, Y., et al., *Endogenous Viral Elements in Shrew Genomes Provide Insights into Pestivirus Ancient History.* Mol Biol Evol, 2022. **39**(10).
57. Bamford, C.G.G., et al., *Comparative analysis of genome-encoded viral sequences reveals the evolutionary history of flavivirids (family Flaviviridae).* Virus Evolution, 2022. **8**(2).
58. Qin, X.C., et al., *A tick-borne segmented RNA virus contains genome segments derived from unsegmented viral ancestors.* Proc Natl Acad Sci U S A, 2014. **111**(18): p. 6744-9.
59. Birney, E., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature, 2007. **447**(7146): p. 799-816.
60. Schattner, P., *Automated querying of genome databases.* PLoS Comput Biol, 2007. **3**(1): p. e1.
61. Obbard, D.J., *Expansion of the metazoan virosphere: progress, pitfalls, and prospects.* Curr Opin Virol, 2018. **31**: p. 17-23.
62. Zhang, Y.Z., M. Shi, and E.C. Holmes, *Using Metagenomics to Characterize an Expanding Virosphere.* Cell, 2018. **172**(6): p. 1168-1172.
63. Zhu, H., R.J. Gifford, and P.R. Murcia, *Distribution, Diversity, and Evolution of Endogenous Retroviruses in Perissodactyl Genomes.* J Virol, 2018. **92**(23).
64. Reus, K., et al., *HERV-K(OLD): ancestor sequences of the human endogenous retrovirus family HERV-K(HML-2).* J Virol, 2001. **75**(19): p. 8917-26.
65. Pavlícek, A., et al., *Processed pseudogenes of human endogenous retroviruses generated by LINEs: their integration, stability, and distribution.* Genome Res, 2002. **12**(3): p. 391-9.
66. Belshaw, R., et al., *High copy number in human endogenous retrovirus families is associated with copying mechanisms in addition to reinfection.* Mol Biol Evol, 2005. **22**(4): p. 814-7.

67. Starrett, G.J., et al., *Adintoviruses: a proposed animal-tropic family of midsize eukaryotic linear dsDNA (MELD) viruses.* Virus Evolution, 2021. **7**(1).
68. Mahanty, S. and M. Bray, *Pathogenesis of filoviral haemorrhagic fevers.* Lancet Infect Dis, 2004. **4**(8): p. 487-98.
69. Mari Saez, A., et al., *Investigating the zoonotic origin of the West African Ebola epidemic.* EMBO Mol Med, 2015. **7**(1): p. 17-23.
70. Taylor, D.J., et al., *Evidence that ebolaviruses and cuevaviruses have been diverging from marburgviruses since the Miocene.* PeerJ, 2014. **2**: p. e556.
71. Carroll, S.A., et al., *Molecular evolution of viruses of the family Filoviridae based on 97 whole-genome sequences.* J Virol, 2013. **87**(5): p. 2608-16.
72. Kryukov, K., et al., *Systematic survey of non-retroviral virus-like elements in eukaryotic genomes.* Virus Res, 2019. **262**: p. 30-36.
73. Cui, J., et al., *Low frequency of paleoviral infiltration across the avian phylogeny.* Genome Biol, 2014. **15**(12): p. 539.
74. Frank, J.A. and C. Feschotte, *Co-option of endogenous viral sequences for host cell function.* Curr Opin Virol, 2017. **25**: p. 81-89.
75. Aswad, A. and A. Katzourakis, *Paleovirology and virally derived immunity.* Trends Ecol Evol, 2012. **27**(11): p. 627-36.
76. Bravo, A., et al., *Antiviral Activity of an Endogenous Parvoviral Element.* Viruses, 2023. **15**(7).
77. Lavialle, C., et al., *Paleovirology of 'syncytins', retroviral env genes exapted for a role in placentation.* Philos Trans R Soc Lond B Biol Sci, 2013. **368**(1626): p. 20120507.
78. 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).
79. Pastuzyn, E.D., et al., *The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer.* Cell, 2018. **172**(1-2): p. 275-288.e18.
80. Quezada-Ramírez, M.A., et al., *Identification of genome safe harbor loci for human gene therapy based on evolutionary biology and comparative genomics.* bioRxiv, 2023: p. 2023.09.08.556857.
81. Hu, G. and L. Kurgan, *Sequence Similarity Searching.* Curr Protoc Protein Sci, 2019. **95**(1): p. e71.
82. Pearson, W.R., *An introduction to sequence similarity ("homology") searching.* Curr Protoc Bioinformatics, 2013. **Chapter 3**: p. Unit3.1.
83. Miller, K., et al., *Identification of multiple Gypsy LTR-retrotransposon lineages in vertebrate genomes.* J Mol Evol, 1999. **49**(3): p. 358-66.
84. Wang, J. and G.Z. Han, *A Sister Lineage of Sampled Retroviruses Corroborates the Complex Evolution of Retroviruses.* Mol Biol Evol, 2021. **38**(3): p. 1031-1039.
85. Kojima, S., et al., *Virus-like insertions with sequence signatures similar to those of endogenous nonretroviral RNA viruses in the human genome.* Proc Natl Acad Sci U S A, 2021. **118**(5).
86. Bruno, M., M. Mahgoub, and T.S. Macfarlan, *The Arms Race Between KRAB-Zinc Finger Proteins and Endogenous Retroelements and Its Impact on Mammals.* Annu Rev Genet, 2019. **53**: p. 393-416.
87. Shakhnovich, B.E. and E.V. Koonin, *Origins and impact of constraints in evolution of gene families.* Genome Res, 2006. **16**(12): p. 1529-36.
88. Buchfink, B., K. Reuter, and H.G. Drost, *Sensitive protein alignments at tree-of-life scale using DIAMOND.* Nat Methods, 2021. **18**(4): p. 366-368.
89. Camacho, C., et al., *ElasticBLAST: accelerating sequence search via cloud computing.* BMC Bioinformatics, 2023. **24**(1): p. 117.
90. Grabowski, P. and J. Rappsilber, *A Primer on Data Analytics in Functional Genomics: How to Move from Data to Insight?* Trends Biochem Sci, 2019. **44**(1): p. 21-32.
91. Kitts, P.A., et al., *Assembly: a resource for assembled genomes at NCBI.* Nucleic Acids Res, 2016. **44**(D1): p. D73-80.

92. Schoch, C.L., et al., *NCBI Taxonomy: a comprehensive update on curation, resources and tools.* Database (Oxford), 2020. **2020**.
93. Gifford, R.J. *DIGS-for-EVEs*. 2023; Available from: https://github.com/giffordlabcvr/DIGS-for-EVEs.
94. Inoue, Y., et al., *Fusion of piggyBac-like transposons and herpesviruses occurs frequently in teleosts.* Zoological Lett, 2018. **4**: p. 6.
95. Barreat, J.G.N. and A. Katzourakis, *Phylogenomics of the Maverick Virus-Like Mobile Genetic Elements of Vertebrates.* Mol Biol Evol, 2021. **38**(5): p. 1731-1743.
96. Koonin, E.V., *On the origin of cells and viruses: primordial virus world scenario.* Ann N Y Acad Sci, 2009. **1178**(1): p. 47-64.
97. Becher, P. and N. Tautz, *RNA recombination in pestiviruses: cellular RNA sequences in viral genomes highlight the role of host factors for viral persistence and lethal disease.* RNA Biol, 2011. **8**(2): p. 216-24.
98. Benit, L., P. Dessen, and T. Heidmann, *Identification, phylogeny, and evolution of retroviral elements based on their envelope genes.* Journal of Virology, 2001. **75**(23): p. 11709-11719.
99. Gallaher, W.R., C. DiSimone, and M.J. Buchmeier, *The viral transmembrane superfamily: possible divergence of Arenavirus and Filovirus glycoproteins from a common RNA virus ancestor.* BMC Microbiol, 2001. **1**: p. 1.
100. Hildebrandt, E., et al., *Evolution of dependoparvoviruses across geological timescales – implications for design of AAV-based gene therapy vectors.* Virus Evolution, 2020.
101. Kumar, S., et al., *TimeTree 5: An Expanded Resource for Species Divergence Times.* Mol Biol Evol, 2022. **39**(8).
102. Stamatakis, A., *RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.* Bioinformatics, 2006. **22**(21): p. 2688-90.
103. Gifford, R.J., et al., *Nomenclature for endogenous retrovirus (ERV) loci.* Retrovirology, 2018. **15**(1): p. 59.
104. Ariel, E., *Viruses in reptiles.* Veterinary Research, 2011. **42**(1): p. 100.
105. Waller, S.J., et al., *Cloacal virome of an ancient host lineage - The tuatara (Sphenodon punctatus) - Reveals abundant and diverse diet-related viruses.* Virology, 2022. **575**: p. 43-53.
106. Soto, E., et al., *First Isolation of a Novel Aquatic Flavivirus from Chinook Salmon (Oncorhynchus tshawytscha) and Its In Vivo Replication in a Piscine Animal Model.* J Virol, 2020. **94**(15).
107. Koda, S.A., et al., *Complete genome sequences of infectious spleen and kidney necrosis virus isolated from farmed albino rainbow sharks Epalzeorhynchos frenatum in the United States.* Virus Genes, 2021. **57**(5): p. 448-452.
108. Harding, E.F., et al., *Revealing the uncharacterised diversity of amphibian and reptile viruses.* ISME Communications, 2022. **2**(1): p. 95.
109. Harvey, E. and E.C. Holmes, *Diversity and evolution of the animal virome.* Nat Rev Microbiol, 2022. **20**(6): p. 321-334.
110. Leroy, E.M., et al., *Multiple Ebola virus transmission events and rapid decline of central African wildlife.* Science, 2004. **303**(5656): p. 387-90.
111. Fernandes, L.P., et al., *A satellite DNA array barcodes chromosome 7 and regulates totipotency via ZFP819.* Science Advances, 2022. **8**(43): p. eabp8085.
112. Enriquez-Gasca, R., et al., *Co-option of endogenous retroviruses through genetic escape from TRIM28 repression.* Cell Rep, 2023. **42**(6): p. 112625.
113. Wickenhagen, A., et al., *A prenylated dsRNA sensor protects against severe COVID-19.* Science, 2021. **374**(6567): p. eabj3624.
114. Ito, J., R.J. Gifford, and K. Sato, *Retroviruses drive the rapid evolution of mammalian APOBEC3 genes.* Proc Natl Acad Sci U S A, 2020. **117**(1): p. 610-618.
115. Shaw, A.E., et al., *Fundamental properties of the mammalian innate immune system revealed by multispecies comparison of type I interferon responses.* PLoS Biol, 2017. **15**(12): p. e2004086.

116. 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).
117. Kambol, R., A. Gatseva, and R.J. Gifford, *An endogenous lentivirus in the germline of a rodent.* Retrovirology, 2022. **19**(1): p. 30.
118. Blanco-Melo, D., R.J. Gifford, and P.D. Bieniasz, *Co-option of an endogenous retrovirus envelope for host defense in hominid ancestors.* Elife, 2017. **6**.
119. Blanco-Melo, D., R.J. Gifford, and P.D. Bieniasz, *Reconstruction of a replication-competent ancestral murine endogenous retrovirus-L.* Retrovirology, 2018. **15**(1): p. 34.

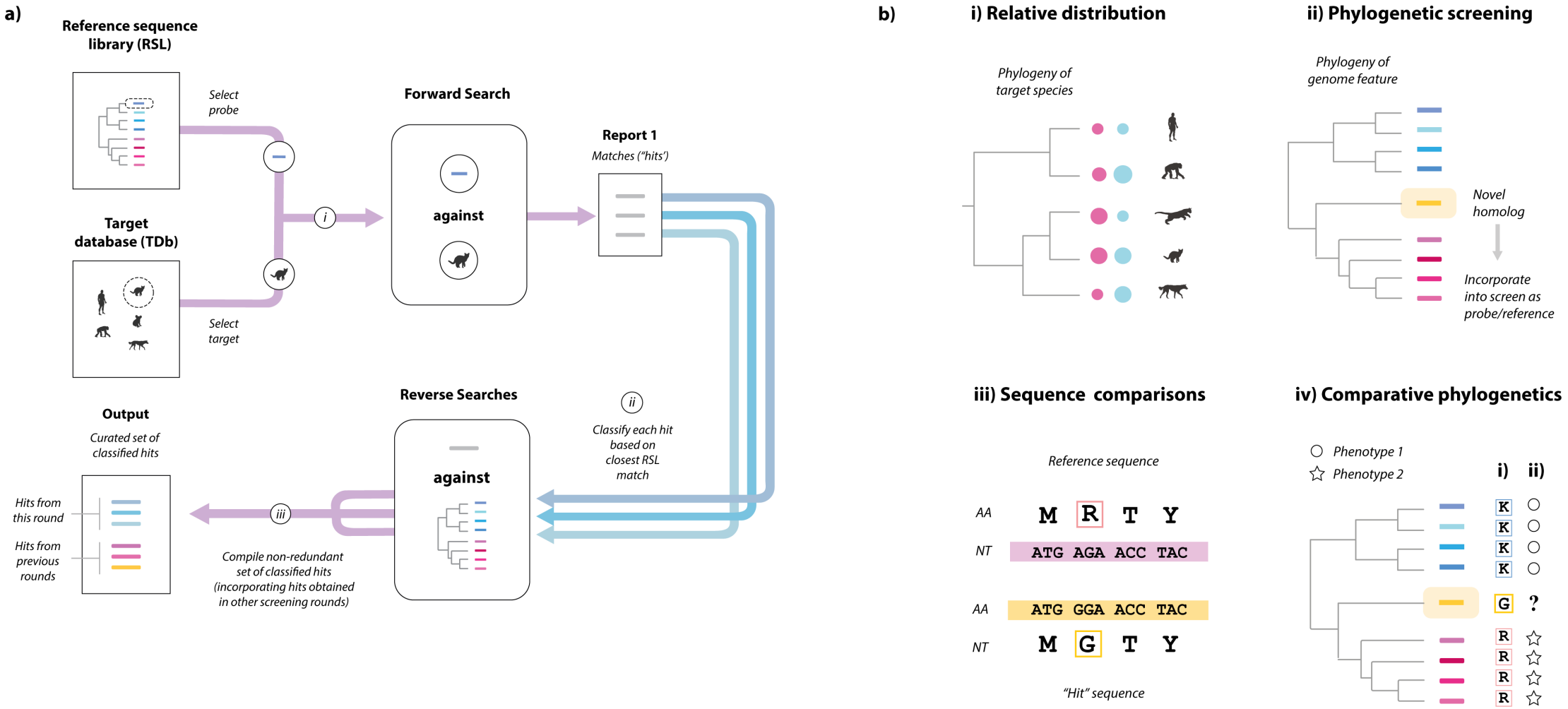
a)
Reference sequence library (RSL)
Select probe
Target database (TDb)
Select target
i
Forward Search
against
Report 1
Matches ("hits')
ii
Classify each hit based on closest RSL match
Reverse Searches
against
iii
Compile non-redundant set of classified hits (incorporating hits obtained in other screening rounds)
Output
Curated set of classified hits
Hits from this round
Hits from previous rounds
b)
i) Relative distribution
Phylogeny of target species
ii) Phylogenetic screening
Phylogeny of genome feature
Novel homolog
Incorporate into screen as probe/reference
iii) Sequence comparisons
Reference sequence
AA M R T Y
NT ATG AGA ACC TAC
AA ATG GGA ACC TAC
NT M G T Y
"Hit" sequence
iv) Comparative phylogenetics
Phenotype 1
Phenotype 2
i) ii)
K
G ?
R

Figure 1.

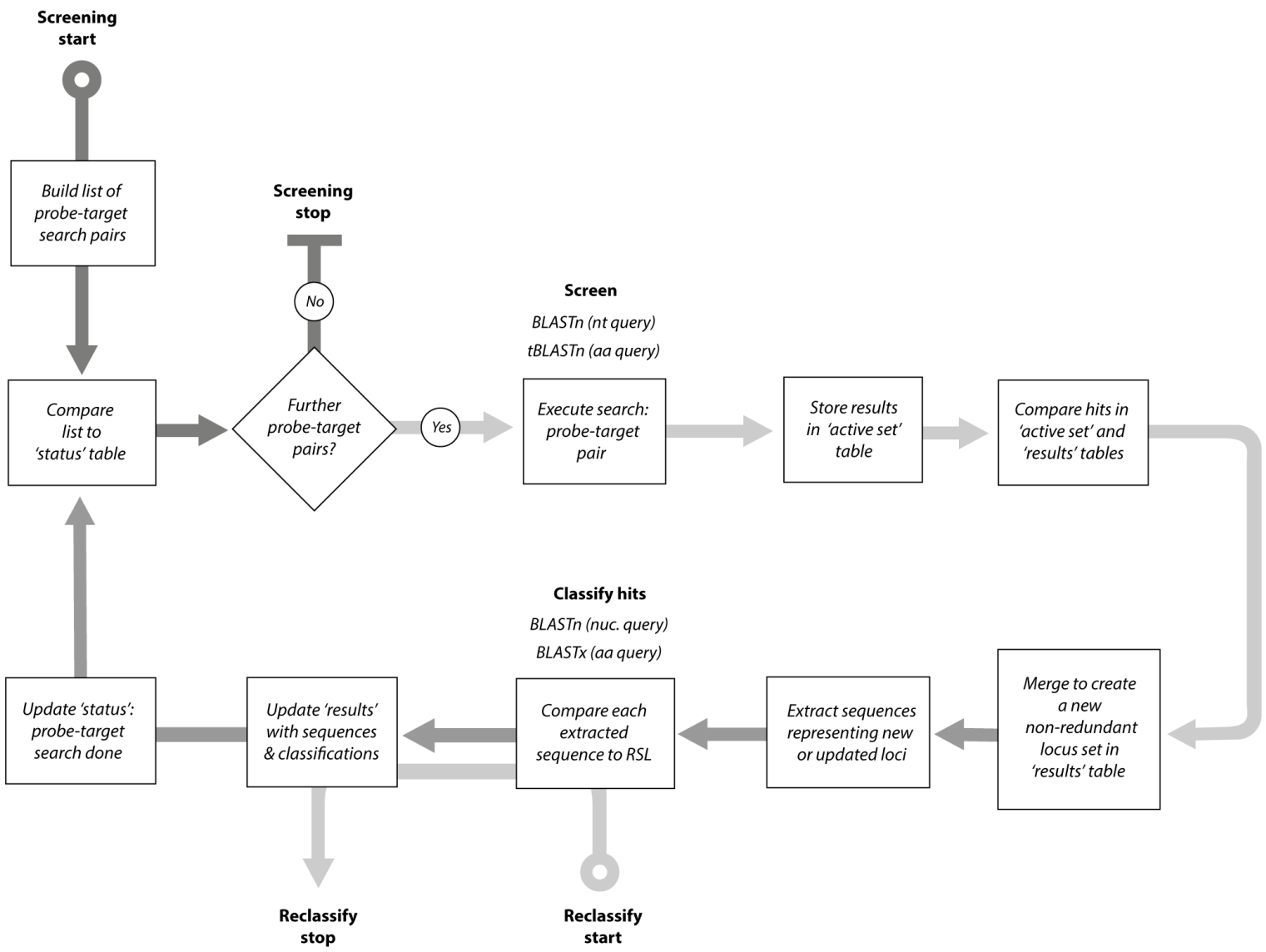
Screening start
Build list of probe-target search pairs
Screening stop
No
Screen
BLASTn (nt query)
tBLASTn (aa query)
Compare list to 'status' table
Further probe-target pairs?
Yes
Execute search: probe-target pair
Store results in 'active set' table
Compare hits in 'active set' and 'results' tables
Classify hits
BLASTn (nuc. query)
BLASTx (aa query)
Update 'status': probe-target search done
Update 'results' with sequences & classifications
Compare each extracted sequence to RSL
Extract sequences representing new or updated loci
Merge to create a new non-redundant locus set in 'results' table
Reclassify stop
Reclassify start

Figure 2.

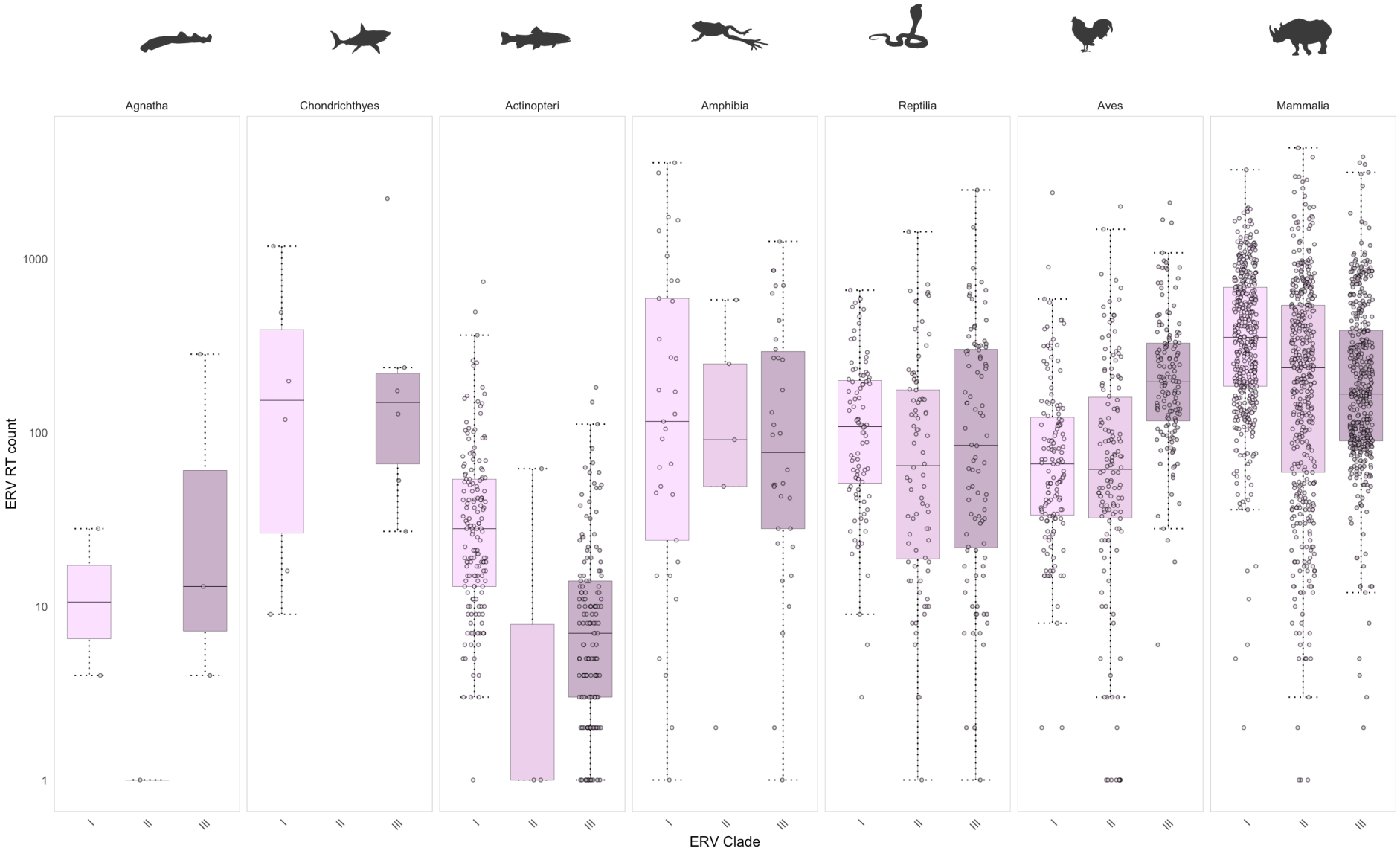
Agnatha
Chondrichthyes
Actinopteri
Amphibia
Reptilia
Aves
Mammalia
ERV RT count
1000
100
10
1
I
II
III
ERV Clade

Figure 3.

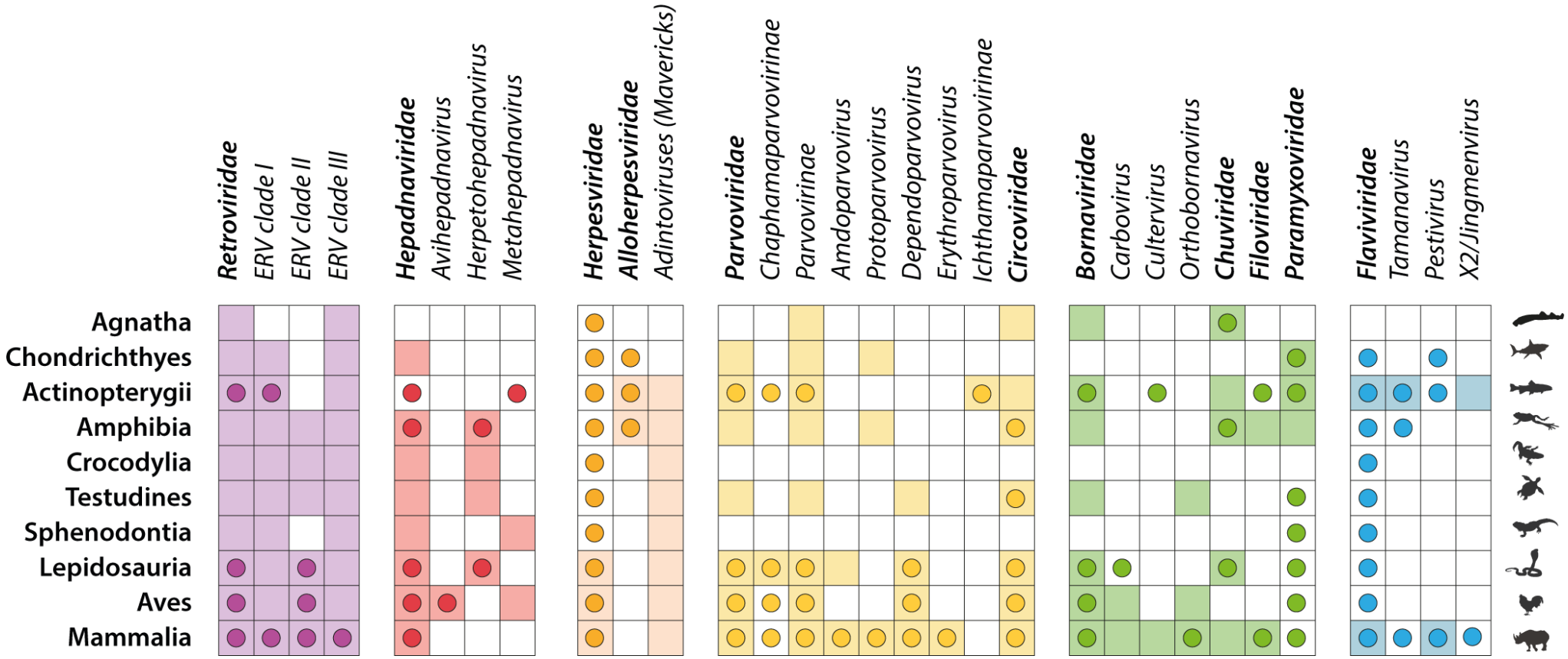
Retroviridae
ERV clade I
ERV clade II
ERV clade III
Hepadnaviridae
Avihepadnavirus
Herpetohepadnavirus
Metahepadnavirus
Herpesviridae
Alloherpesviridae
Adintoviruses (Mavericks)
Parvoviridae
Chaphamaparvovirinae
Parvovirinae
Amdoparvovirus
Protoparvovirus
Dependoparvovirus
Erythroparvovirus
Ichthamaparvovirinae
Circoviridae
Bornaviridae
Carbovirus
Cultervirus
Orthobornavirus
Chuviridae
Filoviridae
Paramyxoviridae
Flaviviridae
Tamanavirus
Pestivirus
X2/Jingmenvirus
Agnatha
Chondrichthyes
Actinopterygii
Amphibia
Crocodylia
Testudines
Sphenodontia
Lepidosauria
Aves
Mammalia

Figure 4.

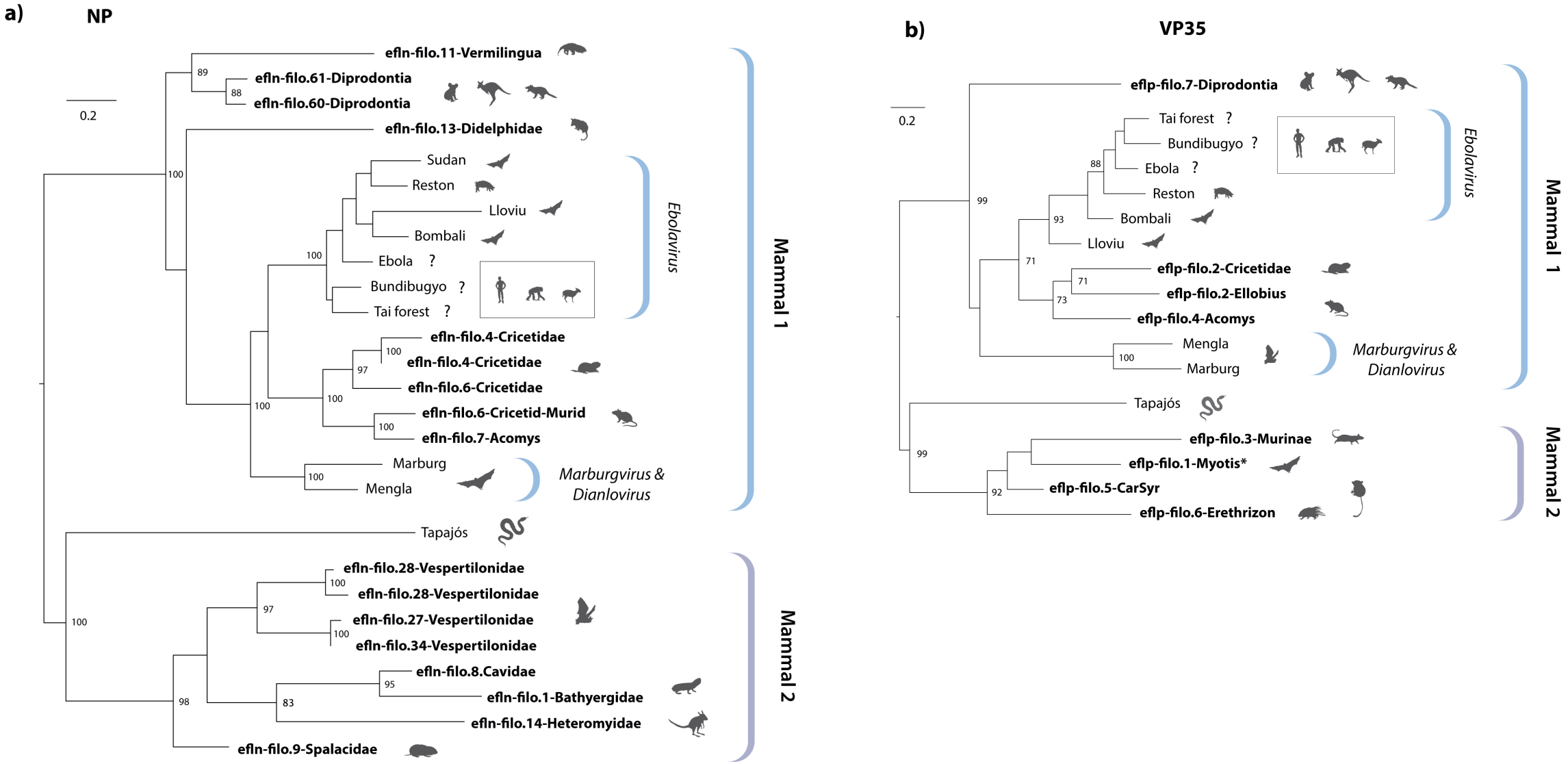
a)
NP
0.2
efln-filo.11-Vermilingua
efln-filo.61-Diprodontia
efln-filo.60-Diprodontia
efln-filo.13-Didelphidae
Sudan
Reston
Lloviu
Bombali
Ebola ?
Bundibugyo ?
Tai forest ?
efln-filo.4-Cricetidae
efln-filo.4-Cricetidae
efln-filo.6-Cricetidae
efln-filo.6-Cricetid-Murid
efln-filo.7-Acomys
Marburg
Mengla
Tapajós
efln-filo.28-Vespertilonidae
efln-filo.28-Vespertilonidae
efln-filo.27-Vespertilonidae
efln-filo.34-Vespertilonidae
efln-filo.8.Cavidae
efln-filo.1-Bathyergidae
efln-filo.14-Heteromyidae
efln-filo.9-Spalacidae
Ebolavirus
Marburgvirus & Dianlovirus
Mammal 1
Mammal 2
b)
VP35
0.2
eflp-filo.7-Diprodontia
Tai forest ?
Bundibugyo ?
Ebola ?
Reston
Bombali
Lloviu
eflp-filo.2-Cricetidae
eflp-filo.2-Ellobius
eflp-filo.4-Acomys
Mengla
Marburg
Tapajós
eflp-filo.3-Murinae
eflp-filo.1-Myotis*
eflp-filo.5-CarSyr
eflp-filo.6-Erethrizon
Ebolavirus
Marburgvirus & Dianlovirus
Mammal 1
Mammal 2

Figure 5.

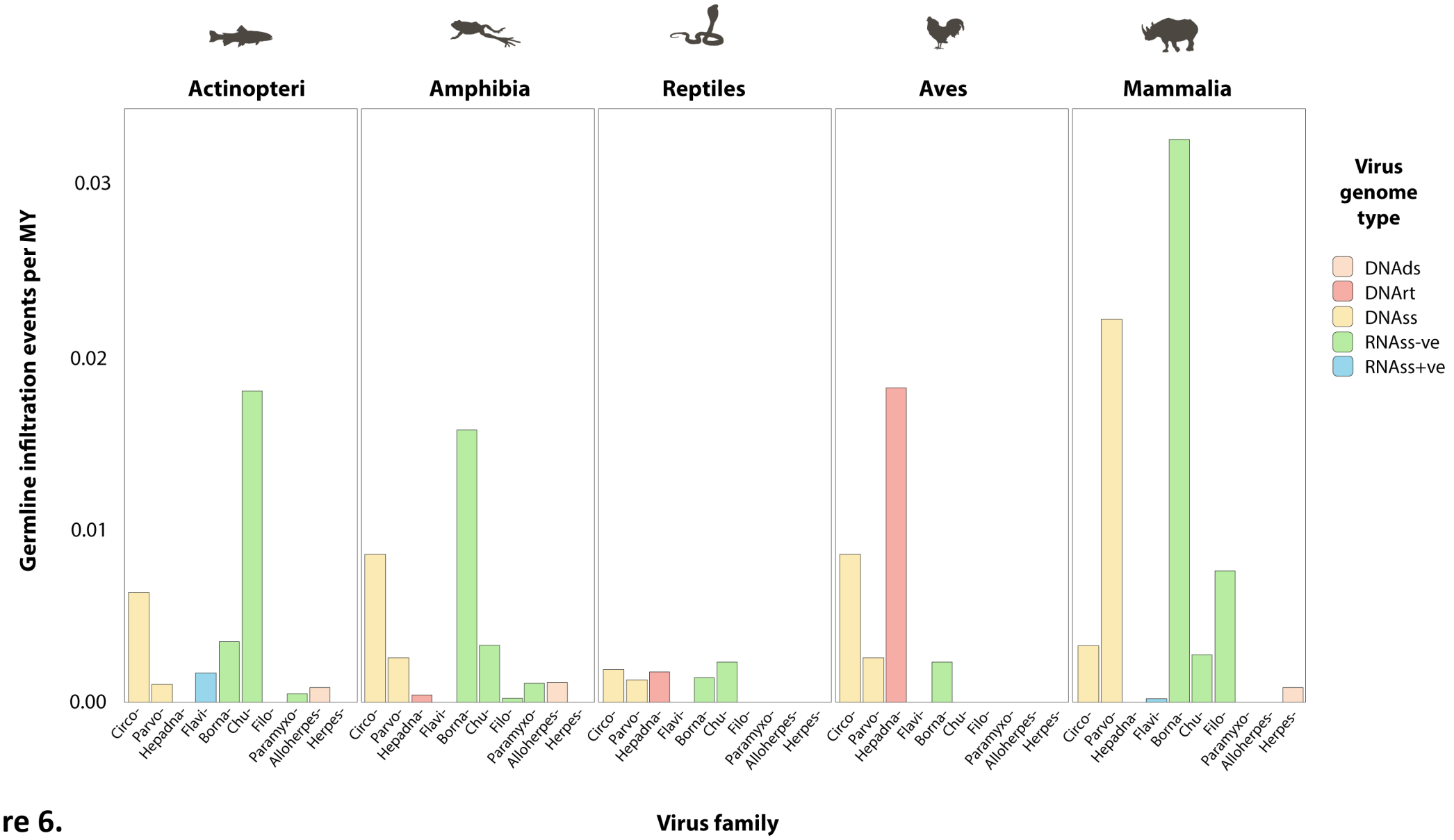
Actinopteri
Amphibia
Reptiles
Aves
Mammalia
Germline infiltration events per MY
0.03
0.02
0.01
0.00
Circo-
Parvo-
Hepadna-
Flavi-
Borna-
Chu-
Filo-
Paramyxo-
Alloherpes-
Herpes-
Virus genome type
DNAds
DNArt
DNAss
RNAss-ve
RNAss+ve
Virus family

Figure 6.

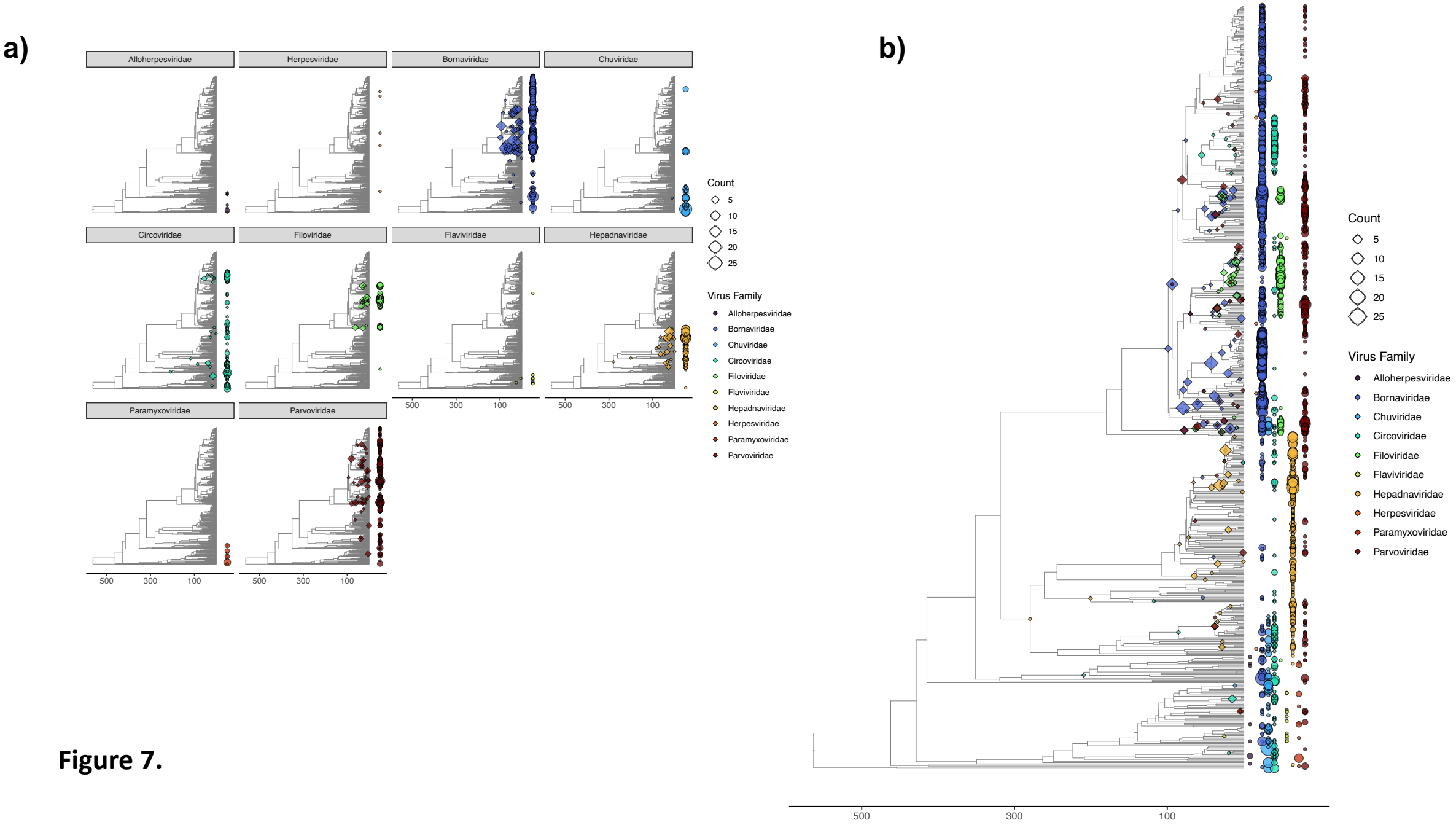
a)
Alloherpesviridae
Herpesviridae
Bornaviridae
Chuviridae
Circoviridae
Filoviridae
Flaviviridae
Hepadnaviridae
Paramyxoviridae
Parvoviridae
500
300
100
Count
5
10
15
20
25
Virus Family
Alloherpesviridae
Bornaviridae
Chuviridae
Circoviridae
Filoviridae
Flaviviridae
Hepadnaviridae
Herpesviridae
Paramyxoviridae
Parvoviridae
b)
500
300
100
Count
5
10
15
20
25
Virus Family
Alloherpesviridae
Bornaviridae
Chuviridae
Circoviridae
Filoviridae
Flaviviridae
Hepadnaviridae
Herpesviridae
Paramyxoviridae
Parvoviridae

Figure 7.