Virus Research 262 (2019) 30–36

Contents lists available at ScienceDirect

Virus Research

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

Systematic survey of non-retroviral virus-like elements in eukaryotic genomes

Kirill Kryukov[a], Mahoko Takahashi Ueda[b], Tadashi Imanishi[a], So Nakagawa[a,b,*]

[a] Department of Molecular Life Science, Tokai University School of Medicine, 143 Shimokasuya, Isehara, Kanagawa 259-1193, Japan
[b] Micro/Nano Technology Center, Tokai University, 411 Kitakaname, Hiratsuka, Kanagawa 259-1292, Japan

ARTICLE INFO

Keywords:
Endogenous viral elements
Database
Evolution
Comparative genomics

ABSTRACT

Endogenous viral elements (EVEs) are viral sequences that are endogenized in the host cell. Recently, several eukaryotic genomes have been shown to contain EVEs. To improve the understanding of EVEs in eukaryotes, we have developed a system for detecting EVE-like sequences in eukaryotes and conducted a large-scale nucleotide sequence similarity search using all available eukaryotic and viral genome assembly sequences (excluding those from retroviruses) stored in the National Center for Biotechnology Information genome database (as of August 14, 2017). We found that 3856 of 7007 viral genomes were similar to 4098 of 4102 eukaryotic genomes. For those EVE-like sequences, we constructed a database, Predicted Endogenous Viral Elements (pEVE, http://peve.med.u-tokai.ac.jp) which provides comprehensive search results summarized from an evolutionary viewpoint. A comparison of EVE-like sequences among closely related species may be useful to avoid false-positive hits. We believe that our search system and database will facilitate studies on EVEs.

1. Introduction

Virus DNA can become integrated into a host genome, where, if infected into germline cells, it is inherited like host DNA. Several retroviruses endogenized in various eukaryotic genomes have been observed (reviewed in Feschotte and Gilbert, 2012; Mager and Stoye, 2015; Imakawa et al., 2015). It was formerly believed that only retroviruses could be endogenized in host cells via a process involving their reverse transcriptase and integrase. However, Horie et al. have identified bornaviruses, nonsegmented single strand RNA viruses, endogenized in several mammalian genomes multiple times via the activity of long interspersed nuclear element-1 (LINE-1) (Horie et al., 2010). Subsequently, several research groups have reported various non-retroviral RNA/DNA virus-like sequences in eukaryotic genomes (Katzourakis and Gifford, 2010; Belyi et al., 2010a,b; Feschotte and Gilbert, 2012; Aiewsakun and Katzourakis, 2015). Such sequences are referred to as endogenous viral elements (EVEs), and some have been observed to have acquired new functions in the host species (Fujino et al., 2014; Parrish et al., 2015; Kobayashi et al., 2016).

Because of recent advances in sequencing technologies, numerous genomes from a wide range of species are now publicly available (Fig. 1), providing an opportunity to conduct large-scale genome-wide study for EVEs in the genomes of various species. However, this rapid increase in genome availability may cause difficulties in the comprehensive detection of EVEs in eukaryotic genomes. Therefore, we have developed a system for detecting EVE-like sequences in eukaryotes.

The GenomeSync database (http://genomesync.org) includes all available genome sequences in the National Center for Biotechnology Information (NCBI) GenBank and RefSeq databases (Benson et al., 2017; O'Leary et al., 2016) for which genome data are accompanied by a matching subset in the NCBI Taxonomy Database (Federhen, 2015). The GenomeSync database is continuously updated via peer-to-peer networking. The Genome Search Toolkit (GSTK; http://kirill-kryukov.com/study/tools/gstk/) is a script kit for conducting comprehensive searches of the GenomeSync database. In this study, we used the database and pipeline to comprehensively search all available eukaryotic genomes for EVEs originating from non-retroviral DNA/RNA viruses and developed a new database – the Predicted Endogenous Viral Element database (pEVE; http://peve.med.u-tokai.ac.jp) – to include the search results and allow them to be presented from an evolutionary viewpoint. Importantly, the GenomeSync database automatically updates to include newly available genomes, which can then be included as GSTK query and/or target sequences to iteratively identify EVEs. Therefore, we believe that our search system and database will facilitate the study of EVEs.

Abbreviations: BoHV4, bovine herpesvirus 4; ds, double strand; EVE, endogenous viral element; GSTK, Genome Search Toolkit; Hsp70, heat shock protein 70; MpV, *Micromonas pusilla* virus; NCBI, National Center for Biotechnology Information; pEVE, Predicted Endogenous Viral Element database; ss, single strand

[*] Corresponding author at: Department of Molecular Life Science, Tokai University School of Medicine, 143 Shimokasuya, Isehara, Kanagawa 259-1193, Japan.
E-mail address: so@tokai.ac.jp (S. Nakagawa).

https://doi.org/10.1016/j.virusres.2018.02.002
Received 30 September 2017; Received in revised form 30 December 2017; Accepted 5 February 2018
Available online 06 February 2018

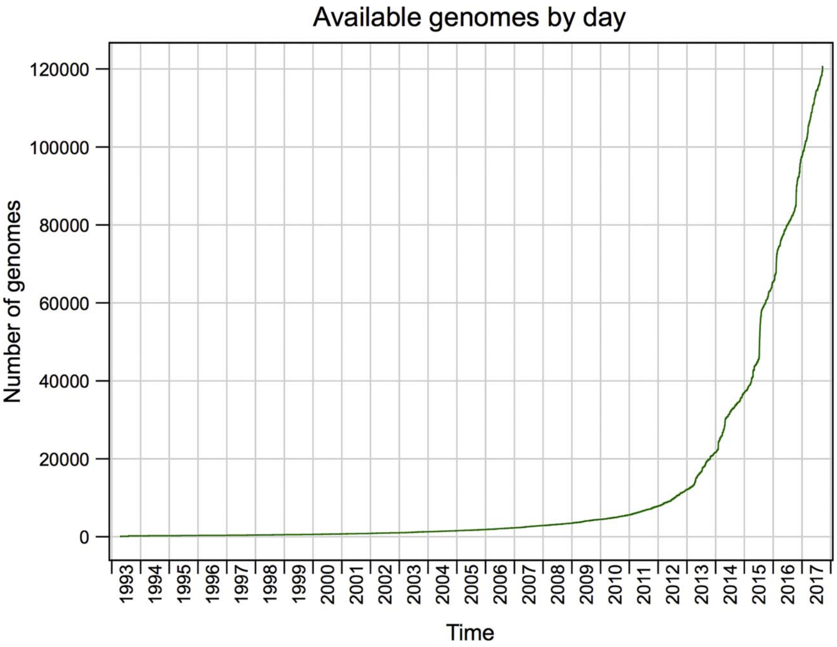


Fig. 1. The number of genomes available in the NCBI GenBank database. On September 25, 2017, a total of 120,679 genome sequences were available in the GenomeSync database, including 4158 eukaryotic and 7451 viral genomes. The total sequence length was 1,637,227,929 nt, of which 1,544,973,691,610 nt sequences (94.4%) corresponded to ATCG.

2. Materials and methods

2.1. GenomeSync database for eukaryotic and viral genomes

The GenomeSync database (http://genomesync.org) is organized as a central repository of genome sequences, distributed to users via peer-to-peer networking. The central node periodically downloads new or updated genome sequences from the NCBI GenBank (Benson et al., 2017) and RefSeq (O'Leary et al., 2016) databases. The genomes are sorted in the FASTA and BLAST formats and labeled according to the NCBI Taxonomy Database (Federhen, 2015). These data, including the taxonomy information, are then automatically distributed to users. As well as new files, information regarding files that have been deleted, changed, or moved is automatically propagated to each client node.

In this study, we used all available eukaryotic and viral genomes (as of August 14, 2017) as target and query sequences, respectively. Eukaryotic genomes comprised 4102 sequences from 335 vertebrates, 472 invertebrates, 2518 fungi, 224 land plants, and 553 other eukaryotic species (listed in Supplementary Data). For viral genomes, we focused on non-retroviral DNA/RNA virus-like sequences, excluding those that belonged to retro-transcribing viruses including *Retroviridae*, *Caulimoviridae*, and *Hepadnaviridae*. As a result, a total of 7007 viral genomes were included in the study: 2697 double-strand DNA (dsDNA) viruses, 284 double-strand RNA (dsRNA) viruses, 911 single-strand DNA (ssDNA) viruses, 1624 single-strand RNA (ssRNA) viruses, 223 Satellites, 4 environmental samples, and 1264 other viruses including phages and viroids (listed in Supplementary Data).

2.2. Sequence search

The GSTK (http://kirill-kryukov.com/study/tools/gstk/) is a set of scripts for managing massive local homology searches, including BLAST, which is particularly useful for searching the GenomeSync database. In this study, we conducted BLASTN searches (e-value < 1E-10) using GSTK scripts for the 4102 eukaryotic and 7007 viral genomes as target and query sequences, respectively, to comprehensively identify EVEs in eukaryotic genomes. These scripts can be downloaded from the pEVE database (http://peve.med.u-tokai.ac.jp). Before the search, we masked simple repeat regions in the query genome sequences using tantan (Frith, 2011) with the options "−x N−r 0.0005."

2.3. Phylogenetic analyses for bovine herpesvirus 4-like sequences

We obtained nucleotide sequences similar to those of bovine herpesvirus 4 (BoHV4) from 27 mammalian species of cetartiodactyla (a group that includes even-toed ungulates and cetaceans). We aligned the nucleotide sequences, including the matched region of BoHV4, using MAFFT version 7.307, with the default options, (Katoh and Standley, 2013) and constructed a maximum-likelihood tree using RAxML version 8.2.10 (Stamatakis, 2014) with a general time-reversible model that applied distributed rate variation among the sites.

2.4. Host preference analysis

We used the Virus-Host database (VHDB) (Mihara et al., 2016) to examine whether there is any correlation between virus sequence integrations and known hosts. In this analysis, we used viruses and eukaryote hosts that are found both in our study and in the VHDB. This set consists of 2807 virus taxa and 485 host taxa. For all EVEs in this set, we measured the topological taxonomic distance between the known host and EVE-harboring eukaryote. The distance is computed as the number of branches on taxonomic tree that separate the two organisms. We then plotted the distances against BLASTN bit-scores to see if there is correlation.

3. Results and discussion

3.1. Summary of the EVE-like sequence search

We used 7007 non-retro DNA/RNA viral genomes as query sequences. Of these, 2.2% were masked as simple repeat regions. We then conducted a BLASTN search in 4102 eukaryotic genomes, which revealed that 55.0% (3856/7007) of the viral genomes were included in at least one eukaryotic genome. Interestingly, 99.9% (4098/4102) of the eukaryotic genomes contained at least one EVE-like sequence. It was found that the most and the second-most abundant viruses (found in 4091 and 4089 eukaryotic genomes, respectively) were uncultured human fecal viruses found in metagenomic analyses; this may have been an artifact. Other than these viruses, many of abundant viruses in eukaryotic genomes were from dsDNA viruses. (See Supplementary

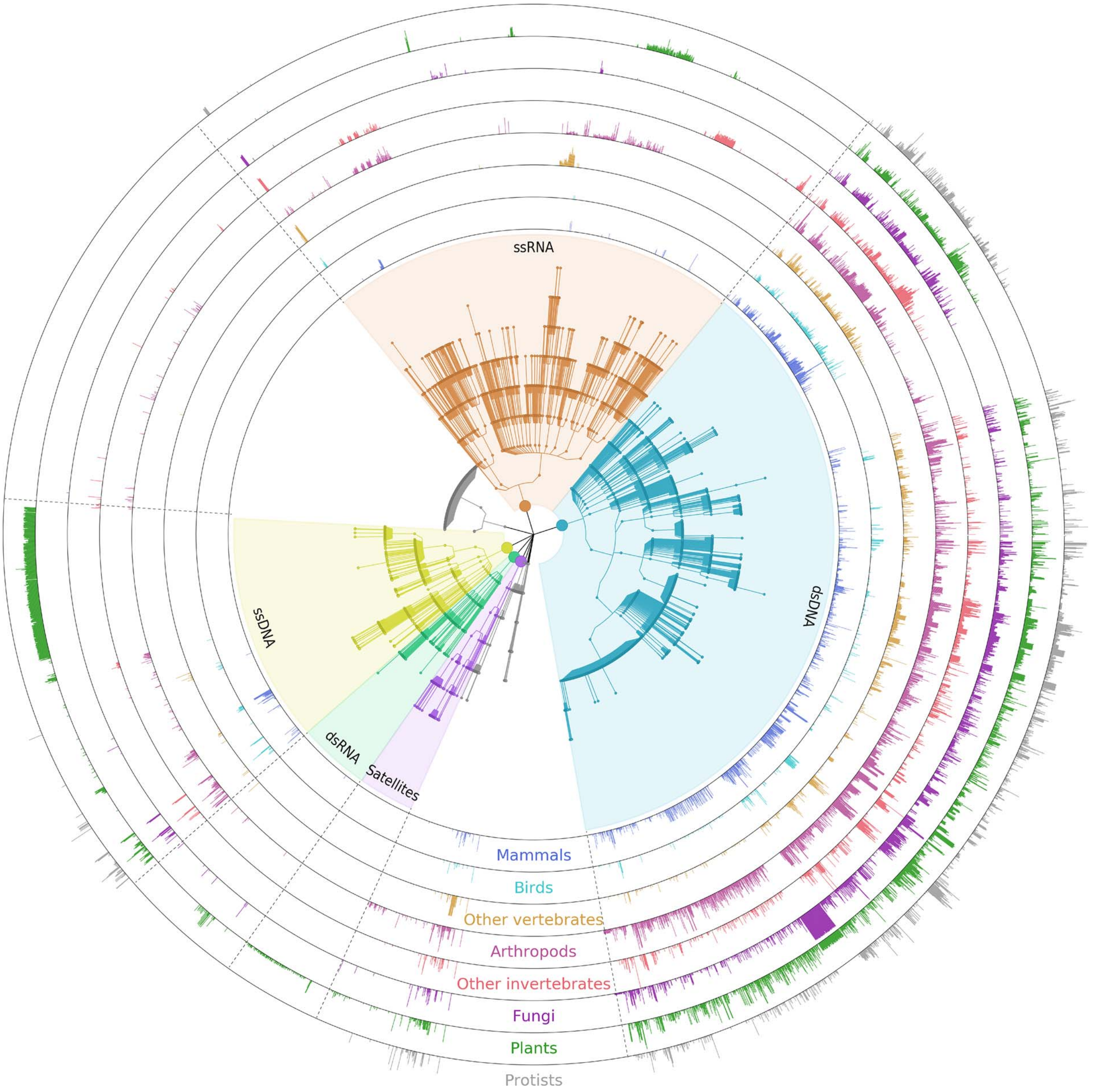


Fig. 2. Taxonomic tree of the 7007 viruses used in this study, labeled and colored according to the virus groups. Viruses not classified in any of the five virus groups are not highlighted. The large circles at the tree center indicate the root of each virus group. The bars in each circle indicate the maximum BLAST bit scores for hits of each virus to genomes in eight categories of eukaryotes; the height of the bars is log-scaled. The labels of eukaryote categories are shown at the bottom of each circle. Taxonomic trees for double-stranded DNA, single-stranded DNA, double-stranded RNA, and single-stranded RNA are shown in Supplementary Figs. S3–S6, respectively. All trees were generated using GraPhlAn (Asnicar et al., 2015).

Table S1 and Supplementary Fig. S1 for details). This was also suggested by taxonomic trees which summarized all the fractions of eukaryotic genomes that contained EVE-like sequences (Figs. 2 and 3, Supplementary Figs. S2–S6). The number of EVE-like sequences in eukaryotic genomes is summarized in Table 1.

The taxonomic trees of eukaryotes showed that sequences similar to dsDNA viruses were widely identified in all eukaryotic genomes (Figs. 2 and 3, and Supplementary Figs. S2 and S3). By contrast, EVE sequences similar to the viruses in other groups, especially dsRNA, ssRNA, and satellites, exhibited strong preference for genomes in particular eukaryotic categories (Figs. 2 and 3). However, in all virus groups, we detected sequence clusters that are specific for particular eukaryotic categories. We found several clusters of fungi-plant specific EVE-like sequences. A detailed taxonomic tree of dsDNA viruses showed that sequences similar to viruses in *Caudovirales* and in *Alloherpeviridae, Mimiviridae*, and *Poxviridae* were abundant in the fungi and plant genomes, respectively (Supplementary Figs. S2 and S3). We also identified that sequences similar to viruses in *Beta-, Gamma-herpeviridae* were highly abundant in the genomes of vertebrates, especially of mammals. Moreover, we found that sequences similar to viruses in *Polydnaviridae* showed high scores in Arthropod genomes. While all groups of satellite, dsRNA, and ssDNA viruses contain large clusters of plant-specific EVE-like sequences, ssDNA group also contain clusters shared among all eukaryotes (Figs. 2 and 3, and Supplementary Figs. S4 and S5). In a group of ssRNA viruses, eukaryotic category-specific EVE-like sequences were identified; many of such EVE-like sequences were invertebrate-specific clusters (Fig. 2, and Supplementary Figs. S3 and S6). One interesting observation in this virus group is the presence of the largest cluster of reptile-fish (other vertebrates) specific sequences among all virus groups.

This study identified both known and novel non-retroviral EVE-like sequences as well as sequences that were probably horizontally transferred from host species to viruses. Some of the well-known examples of non-retrovirus EVEs include sequences derived from nonsegmented and

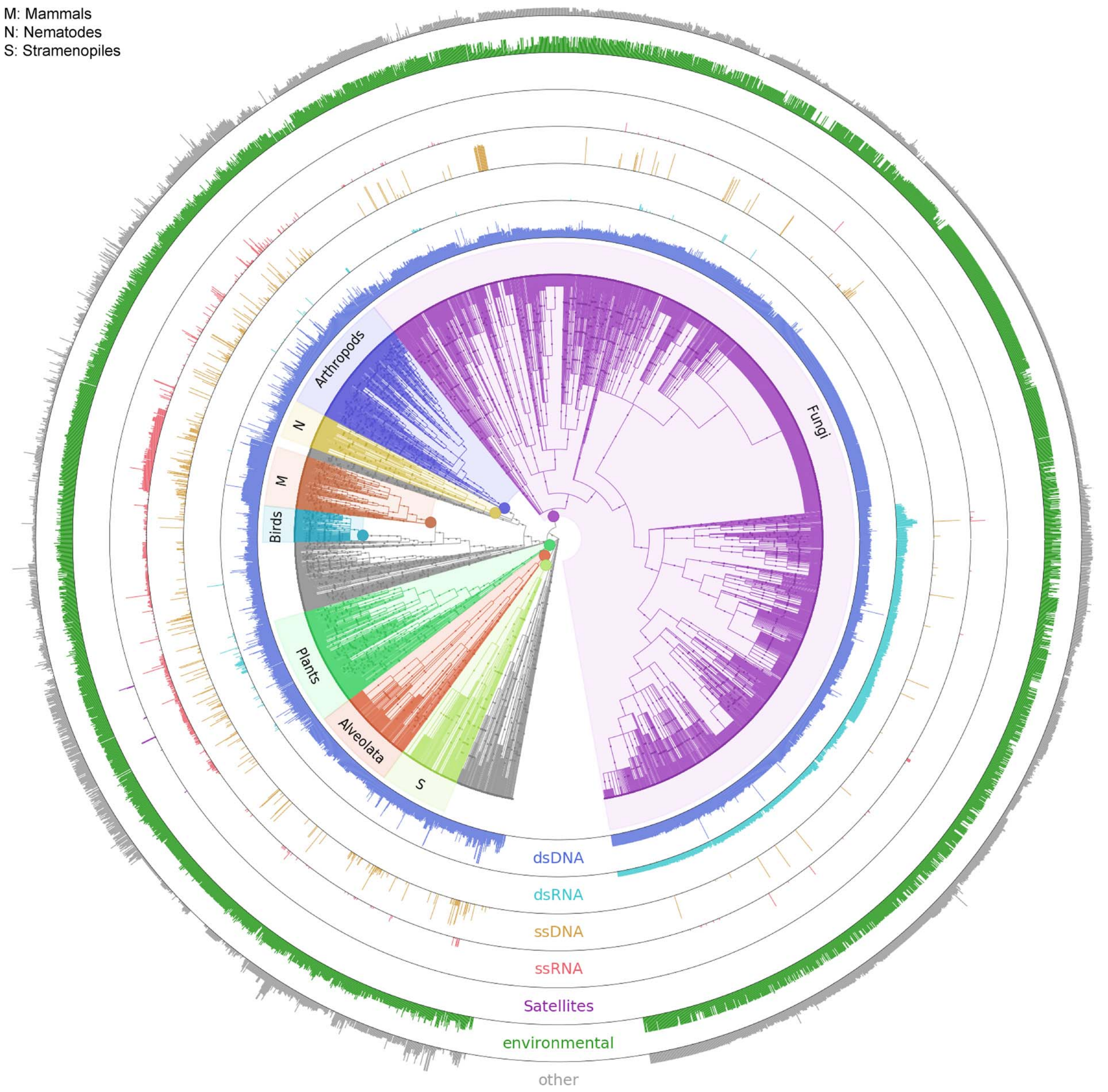


Fig. 3. Taxonomic tree of the 4120 eukaryotes labeled and colored according to the eukaryotic categories. The bars in each circle indicate the maximum BLAST bit scores for hits of each eukaryote to genomes in seven categories of viruses. For details of labeling and coloring, please see Fig. 2. For the dsDNA viruses, more detail categories are shown in Supplementary Fig. S6.

Table 1
The number of viral genomes that have at least one eukaryotic genome hit in the search.

Virus	The number of viral genomes	Total length (bp)	The number of viral genomes containing at least one eukaryotic genome hit
dsDNA	2697	214,619,795	2353 (87.2%)
dsRNA	284	3,147,306	45 (15.8%)
ssDNA	911	3,293,031	595 (65.3%)
ssRNA	1624	16,692,894	547 (33.7%)
other	1491	15,546,295	316 (21.2%)
All	7007	253,299,321	3856 (55.0%)

negative-sense RNA viruses, such as bornaviruses and filoviruses (Horie et al., 2010; Belyi et al., 2010a,b; Katzourakis and Gifford, 2010). We found both bornavirus- and filovirus-like EVEs in the genomes of several mammalian lineages, such as primates, rodents, elephants, and bats. Filovirus-like EVEs have previously been found in the genomes of shrews, tenrecs, and marsupials (Belyi et al., 2010a,b; Katzourakis and Gifford, 2010). Our analysis also detected these sequences in mammalian genomes; however, unlike previous studies, we did not detect any bornavirus-like EVE in primate genomes or filovirus-like EVE in bat, shrew, and tenrec genomes. Even with a relaxed threshold (e-value < 1E-5), no hits were obtained for these genomes. This may have been because of differences in the methods used to detect EVE-like sequences between our study and previous studies. Indeed, using the TBLASTX program, we could detect all hits for bornavirus- and filovirus-like sequences in these genomes, as previously reported. This suggests that the BLASTN program is suitable only for comprehensive analyses of EVEs recently integrated into eukaryotic genomes. However, to improve the detection sensitivity of EVE-like sequences derived from non-retroviral DNA/RNA viruses in eukaryotic genomes, we intend to add data predicted by the TBLASTX program in the future.

Although detection sensitivity is limited for sequences with low similarity to virus genomes, our comprehensive analysis could identify novel EVE-like sequences, such as enterovirus-like sequences in parasitic roundworms (*Trichinella*), glossinavirus-like sequences in tsetse

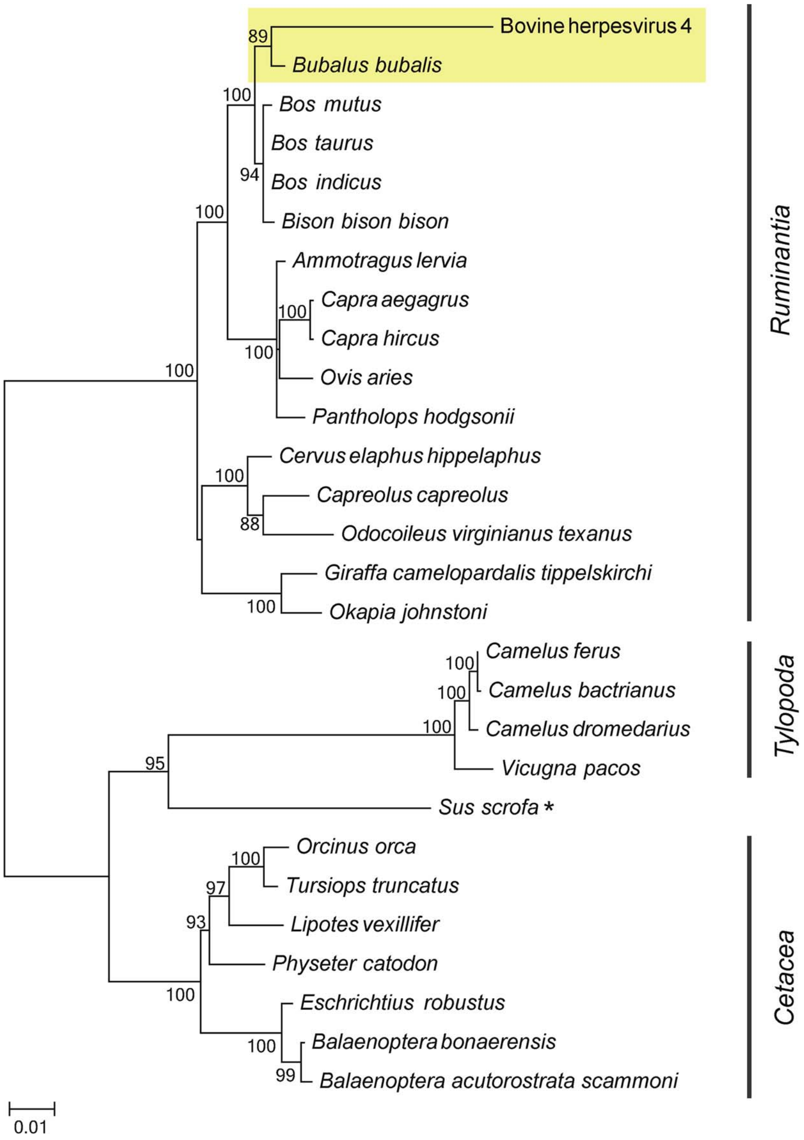


Fig. 4. Phylogenetic tree for bovine herpesvirus 4-like sequences of genomes of cetartiodactyla. Bovine herpesvirus 4 is also included. Bootstrap values based on 1000 replicates are shown on the branches for scores of ≥80. The cluster of BHV4 and the Asian water buffalo (*Bubalus bubalis*) is highlighted in yellow. Names of sub-/infra-orders are on the right side of the tree. The *Suina* infraorder is indicated by an asterisk.

flies (*Glossina*), and hubei virga-like virus-like sequences in anopheles mosquitoes (*Anopheles*).

Herpesvirus is a large family of dsDNA viruses containing diverged virus species containing three sub-families. Among these, human herpesvirus 6 in the sub-family *Betaherpesvirinae* is known to be integrated into the human genome (Arbuckle et al., 2010). Other herpesvirus-like EVEs have been reported in the genomes of non-human primates such as the aye-aye (*Daubentonia madagascariensis*), Philippine tarsier (*Tarsius syrichta*), and bonobo (*Pan paniscus*) (Aswad and Katzourakis, 2014). We confirmed these results in our analyses; furthermore, we obtained several hits for novel herpesviruses in all sub-families for eukaryotic genomes, such as yeasts (*Saccharomycetes*), birds (*Neognathae*), rodents (*Muriadae*), and bats (*Chiroptera*).

It has been shown that host-to-virus horizontal transfers have often occurred in large dsDNA viruses (Holzerlandt et al., 2002; Hughes and Friedman, 2003, 2005; Filee and Chandler, 2010; Gilbert et al., 2016). One interesting example is the sequence homologous to BoHV4 in gammaherpesvirinae. Our homology searches found BoHV4-like sequences in the genomes of mammalian species in the super-orders *Boreotheria* and *Xenarthra* with high scores (> 1000). Especially, Ruminantia and Cetacea had large scores (> 2400 and > 2000, respectively). The coordinate of BoHV4 sequences homologous to mammalian genomes corresponds with BoHV4gp79, which is a viral beta-1,6-N-acetylglucosaminyltransferase gene. Protein BLAST searches of the BoHV4gp79 sequence against the NCBI non-redundant protein database revealed that the viral gene was similar to the beta-1,3-galactosyl-*O*-glycosyl-glycoprotein beta-1,6-N-acetylglucosaminyltransferase 3 gene of the bovine family. We extracted BoHV4-like sequences from mammalian genomes and constructed a phylogenetic tree (Fig. 4). This showed that BoHV4 formed a cluster with the sequence of the Asian water buffalo (*Bubalus bubalis*), suggesting the possibility that the viral gene BoHV4gp79 was transferred from the genome of a buffalo or of other bovine species.

Another example of transferred sequences from eukaryotic genomes to viruses was found in the sequences similar to MpV, which is a double-strand DNA virus infecting green alga, *Micromonas pusilla* (Mayer and Taylor, 1979). We found that the hit sequence in the MpV genome was a part of the coding sequence of the heat shock protein 70 (Hsp70). It has been reported that Hsp70 in viral genomes is horizontally transferred from the host species (Moreau et al., 2010; Liu et al., 2010). We conducted a TBLASTN sequence similarity search of all viral genomes in the GenomeSync database using the amino acid sequence of Hsp70 as the query and found that genomes of several other marine viruses also contained Hsp70 sequences: *Bathycoccus sp. RCC1105 virus BpV1*, *Chrysochromulina ericina virus isolate CeV-01B*, *Acanthamoeba polyphaga mimivirus*, *Mimivirus terra2*, *Cafeteria roenbergensis virus BV-PW1*, *Megavirus chiliensis*, and *Acanthamoeba polyphaga moumouvirus*. These results demonstrated that our comprehensive datasets facilitated a detailed analysis of the evolutionary history of candidate sequences and the identification of not only EVE-like sequences but also sequences that had been potentially horizontally transferred from eukaryotic genomes to viral genomes.

One interesting question is whether exchange of DNA occurs more often between a virus and its known host. We compared the BLASTN scores of EVEs with the distances between actual host and EVE-harboring host. Overall, we found nearly no correlation – that is, a virus exchanges DNA with remote hosts about as readily as with its own natural host. Only very few viruses show host preference (Figs. S7, S8).

3.2. Genome contamination

Kryukov and Imanishi reported that human contamination in the public genome assemblies (Kryukov and Imanishi, 2016). Therefore, it is possible that some EVE-like sequences found in this study could also be contamination. Although it is difficult to conclude that some EVE-like sequences are really contamination based only on the sequence comparison, there is a high possibility that EVE-like sequences are contaminations if they satisfy the following conditions: 1) EVE-like sequences are found only in one genome even if several genomes of closely related species are used for the search. 2) Phylogenetic analyses show that EVE-like sequences are very close to a specific virus species. 3) A contig (a set of overlapping sequences representing a consensus genomic region) contains only EVE-like sequences. 4) There are no duplicated EVE-like sequences in the genome. 5) The species of the genomes containing a specific virus like sequences are not the host of the virus.

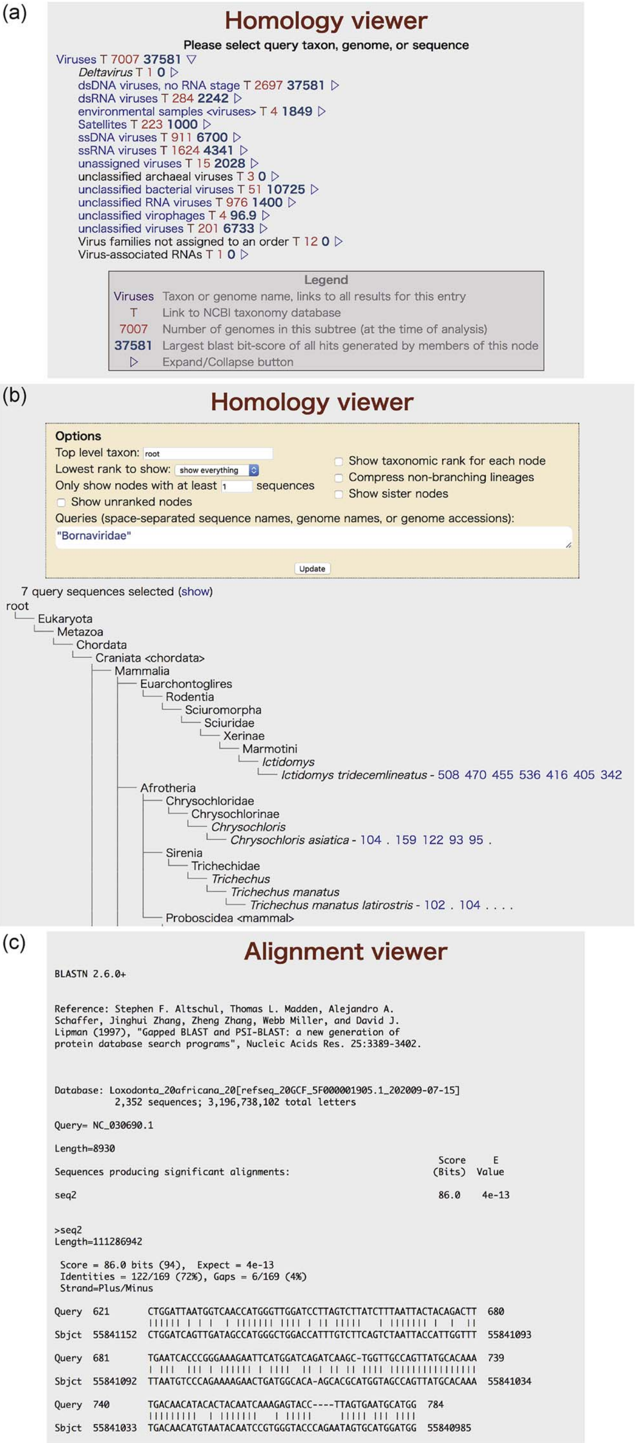
For example, in the Atlantic salmon (*Salmo salar*) genome (RefSeq ID: GCF_000233375.1), there are six contigs containing Influenza A virus-like sequences, each of which corresponds to the sequence of different segment of the Influenza A virus genome (Table 2). Further, we found the following evidences that the Influenza A virus-like sequence found in the Atlantic salmon genome could be contamination; 1) An entire region of each Atlantic salmon contig corresponds to the sequence of different segment (1–5, and 7) of the influenza A virus genome, 2) No similar sequences were found in the other two salmon species (*Oncorhynchus mykiss* and *Oncorhynchus kisutch*) in the subfamily Salmoniae, and 3) A BLASTN search of the Atlantic salmon genome including these contigs against all available virus genomes in the family Orthmyxoviridae show high similarity only to influenza A with E-value ≤1E-10, whereas salmon anemia (infectious to Atlantic salmons) viruses belonging to Orthmyxoviridae was not found in this search. Therefore, it can be concluded that the Influenza A virus-like sequences in the Atlantic salmon genome are contamination. Further analyses are required to solve the contamination problem.

Table 2
Sequence comparisons between Atlantic salmon and Influenza A virus genomes.

Atlantic salmon genome	Influenza A virus genome	Identity (aligned length)
NW_012393421.1 (2343 nt)	Segment 1, H3N2/ NC_007373.1 (2341 nt)	97.9% (2291 nt)
NW_012403001.1 (1958 nt)	Segment 2, H3N2/ NC_007372.1 (2341 nt)	97.5% (1909 nt)
NW_012396048.1 (2206 nt)	Segment 3, H3N2/ NC_007371.1 (2233 nt)	97.6% (2152 nt)
NW_012560585.1 (1067 nt)	Segment 4, H1N1/ CY035030.1 (1737 nt)	99.9% (1067 nt)
NW_012492408.1 (1262 nt)	Segment 5, H1N1/ CY039994.1 (1566 nt)	100.0% (1262 nt)
NW_012566800.1 (1027 nt)	Segment 7, H1N1/ EU742637.2 (1027 nt)	99.1% (1026 nt)

3.3. Database

We constructed a web database—pEVE (http://peve.med.u-tokai.ac.jp)—to visualize our search results (Fig. 5a); all viruses analyzed in this study are listed according to their taxonomy. When the name of a



Homology viewer

Please select query taxon, genome, or sequence

Viruses T 7007 **37581** ▽
Deltavirus T 1 **0** ▷
dsDNA viruses, no RNA stage T 2697 **37581** ▷
dsRNA viruses T 284 **2242** ▷
environmental samples <viruses> T 4 **1849** ▷
Satellites T 223 **1000** ▷
ssDNA viruses T 911 **6700** ▷
ssRNA viruses T 1624 **4341** ▷
unassigned viruses T 15 **2028** ▷
unclassified archaeal viruses T 3 **0** ▷
unclassified bacterial viruses T 51 **10725** ▷
unclassified RNA viruses T 976 **1400** ▷
unclassified virophages T 4 **96.9** ▷
unclassified viruses T 201 **6733** ▷
Virus families not assigned to an order T 12 **0** ▷
Virus-associated RNAs T 1 **0** ▷

Legend

Viruses	Taxon or genome name, links to all results for this entry
T	Link to NCBI taxonomy database
7007	Number of genomes in this subtree (at the time of analysis)
37581	Largest blast bit-score of all hits generated by members of this node
▷	Expand/Collapse button

Homology viewer

Options
Top level taxon: root
Lowest rank to show: show everything
Only show nodes with at least 1 sequences
□ Show unranked nodes
□ Show taxonomic rank for each node
□ Compress non-branching lineages
□ Show sister nodes
Queries (space-separated sequence names, genome names, or genome accessions):
"Bornaviridae"

7 query sequences selected (show)

```
root
└── Eukaryota
    └── Metazoa
        └── Chordata
            └── Craniata <chordata>
                ├── Mammalia
                │   ├── Euarchontoglires
                │   │   └── Rodentia
                │   │       └── Sciuromorpha
                │   │           └── Sciuridae
                │   │               └── Xerinae
                │   │                   └── Marmotini
                │   │                       └── Ictidomys
                │   │                           └── Ictidomys tridecemlineatus - 508 470 455 536 416 405 342
                │   ├── Afrotheria
                │   │   ├── Chrysochloridae
                │   │   │   └── Chrysochlorinae
                │   │   │       └── Chrysochloris
                │   │   │           └── Chrysochloris asiatica - 104 . 159 122 93 95 .
                │   │   ├── Sirenia
                │   │   │   └── Trichechidae
                │   │   │       └── Trichechus
                │   │   │           └── Trichechus manatus
                │   │   │               └── Trichechus manatus latirostris - 102 . 104 . . . .
                │   │   └── Proboscidea <mammal>
```

Alignment viewer

```
BLASTN 2.6.0+


Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro A.
Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J.
Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs", Nucleic Acids Res. 25:3389-3402.



Database: Loxodonta_20africana_20[refseq_20GCF_5F000001905.1_202009-07-15]
           2,352 sequences; 3,196,738,102 total letters

Query= NC_030690.1

Length=8930
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

seq2                                                                  86.0    4e-13


>seq2
Length=111286942

 Score = 86.0 bits (94),  Expect = 4e-13
 Identities = 122/169 (72%), Gaps = 6/169 (4%)
 Strand=Plus/Minus

Query  621       CTGGATTAATGGTCAACCATGGGTTGGATCCTTAGTCTTATCTTTAATTACTACAGACTT  680
                 |||||| | | |  | ||||||| |||| | || |||||   | ||||||| |  |  ||
Sbjct  55841152  CTGGATCAGTTGATAGCCATGGGCTGGACCATTTGTCTTCAGTCTAATTACCATTGGTTT  55841093

Query  681       TGAATCACCCGGGAAAGAATTCATGGATCAGATCAAGC-TGGTTGCCAGTTATGCACAAA  739
                 | |||  ||| | |||||| | ||||  || | || || |||| ||||||||||||||||
Sbjct  55841092  TTAATGTCCCAGAAAAGAACTGATGGCACA-AGCACGCATGGTAGCCAGTTATGCACAAA  55841034

Query  740       TGACAACATACACTACAATCAAAGAGTACC----TTAGTGAATGCATGG  784
                 |||||||||  | |||||||   | |||||     ||||| ||| ||||
Sbjct  55841033  TGACAACATGTAATACAATCCGTGGGTACCCAGAATAGTGCATGGATGG  55840985
```

Fig. 5. Example screenshots of the interface for the pEVE website (http://peve.med.u-tokai.ac.jp). (a) The top page of the pEVE database presents the virus taxonomy and statistical scores. (b) The taxonomical view shows all species containing similar sequences in their genomes with their BLAST bit scores. Several options are available to change the view. (c) An example of the local alignment of eukaryotic and viral sequences.

viral taxonomy or species is selected, the system shows all the species that contains similar sequences in their genomes with their BLAST bit scores (Fig. 5b). The default viewer hides unranked and/or sister nodes of viral taxonomies, but these can be revealed if needed. The detailed alignments of these two sequences can also be viewed (Fig. 5c), and the BLASTN search results can be downloaded.

3.4. Perspective

This report describes a procedure to identify EVE candidate sequences in all available eukaryotic genomes using the GenomeSync database with GSTK script. The search scripts used in this study can be downloaded via the pEVE website, allowing anyone to conduct similar comprehensive searches of EVE sequences using all currently available eukaryotic genomes with any homology search software and parameter. The number of genome sequences continues to markedly increase (Fig. 1), and we believe that this system will prove useful for the comprehensive identification of EVE sequences. Previously, we have developed a database (http://geve.med.u-tokai.ac.jp) that provides coding sequences in EVEs obtained from 20 genomes of 19 mammalian species (Nakagawa and Takahashi, 2016). Compared with that database, the pEVE database provides comprehensive EVE-like sequence information, although it is not annotated to the same extent. Further phylogenetic analyses of EVE-like sequences in the pEVE database are required for a greater understanding of the detailed distribution of EVEs in eukaryotes.

Acknowledgments

This work was supported by grants from Scientific Research on Innovative Areas (16H06429, 16K21723, 17H05823 to S.N.), Challenging Exploratory Research (17K19359 to M.T.U.), Research Activity Start-up (17H07123 to K.K.), and the Strategic Research Foundation at Private Universities (to S.N. and M.T.U.) from the Ministry of Education, Culture, Science, Sports, and Technology (MEXT) of Japan. The super-computing resource was provided by Human Genome Center, the Institute of Medical Science, the University of Tokyo.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.virusres.2018.02.002.

References

Aiewsakun, P., Katzourakis, A., 2015. Endogenous viruses: connecting recent and ancient viral evolution. Virology 479–480, 26–37. http://dx.doi.org/10.1016/j.virol.2015.02.011.

Arbuckle, J.H., Medveczky, M.M., Luka, J., Hadley, S.H., Luegmayr, A., Ablashi, D., Lund, T.C., Tolar, J., De Meirleir, K., Montoya, J.G., Komaroff, A.L., Ambros, P.F., Medveczky, P.G., 2010. The latent human herpesvirus-6A genome specifically integrates in telomeres of human chromosomes in vivo and in vitro. Proc. Natl. Acad. Sci. U. S. A. 107, 5563–5568. http://dx.doi.org/10.1073/pnas.0913586107.

Asnicar, F., Weingart, G., Tickle, T.L., Huttenhower, C., Segata, N., 2015. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ 18 (3), e1029. http://dx.doi.org/10.7717/peerj.1029.eCollection.2015.

Aswad, A., Katzourakis, A., 2014. The first endogenous herpesvirus, identified in the tarsier genome, and novel sequences from primate rhadinoviruses and lymphocryptoviruses. PLoS Genet. 10http://dx.doi.org/10.1371/journal.pgen.1004332. e1004332-17.

Belyi, V.A., Levine, A.J., Skalka, A.M., 2010a. Unexpected inheritance: multiple integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate genomes. PLoS Pathog. 6, e1001030. http://dx.doi.org/10.1371/journal.ppat.1001030.

Belyi, V.A., Levine, A.J., Skalka, A.M., 2010b. 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. 84, 12458–12462. http://dx.doi.org/10.1128/JVI.01789-10.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2017. GenBank. Nucleic Acids Res. 45, D37–D42. http://dx.doi.org/10.1093/nar/gkw1070.

Federhen, S., 2015. Type material in the NCBI taxonomy database. Nucleic Acids Res. 43, D1086–1098. http://dx.doi.org/10.1093/nar/gku1127.

Feschotte, C., Gilbert, C., 2012. Endogenous viruses: insights into viral evolution and impact on host biology. Nat. Rev. Genet. 13, 283–296. http://dx.doi.org/10.1038/nrg3199.

Filee, J., Chandler, M., 2010. Gene exchange and the origin of giant viruses. Intervirology 53, 354–361. http://dx.doi.org/10.1159/000312920.

Frith, M.C., 2011. A new repeat-masking method enables specific detection of homologous sequences. Nucleic Acids Res. 39, e23. http://dx.doi.org/10.1093/nar/gkq1212.

Fujino, K., Horie, M., Honda, T., Merriman, D.K., Tomonaga, K., 2014. Inhibition of Borna disease virus replication by an endogenous bornavirus-like element in the ground squirrel genome. Proc. Natl. Acad. Sci. U. S. A. 111, 13175–13180. http://dx.doi.org/10.1073/pnas.1407046111.

Gilbert, C., Peccoud, J., Chateigner, A., Moumen, B., Cordaux, R., Herniou, E., 2016. Continuous influx of genetic material from host to virus populations. PLoS Genet. 12, e1005838. http://dx.doi.org/10.1371/journal.pgen.1005838.

Holzerlandt, R., Orengo, C., Kellam, P., Alba, M.M., 2002. Identification of new herpesvirus gene homologs in the human genome. Genome Res. 12, 1739–1748. http://dx.doi.org/10.1101/gr.334302.

Horie, M., Honda, T., Suzuki, Y., Kobayashi, Y., Daito, T., Oshida, T., Ikuta, K., Jern, P., Gojobori, T., Coffin, J.M., Tomonaga, K., 2010. Endogenous non-retroviral RNA virus elements in mammalian genomes. Nature 463, 84–87. http://dx.doi.org/10.1038/nature08695.

Hughes, A.L., Friedman, R., 2003. Genome-wide survey for genes horizontally transferred from cellular organisms to baculoviruses. Mol. Biol. Evol. 20, 979–987. http://dx.doi.org/10.1093/molbev/msg107.

Hughes, A.L., Friedman, R., 2005. Poxvirus genome evolution by gene gain and loss. Mol. Phylogen. Evol. 35, 186–195. http://dx.doi.org/10.1016/j.ympev.2004.12.008.

Imakawa, K., Nakagawa, S., Miyazawa, T., 2015. Baton pass hypothesis: successive incorporation of unconserved endogenous retroviral genes for placentation during mammalian evolution. Genes Cells 20, 771–788. http://dx.doi.org/10.1111/gtc.12278.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780. http://dx.doi.org/10.1093/molbev/mst010.

Katzourakis, A., Gifford, R.J., 2010. Endogenous viral elements in animal genomes. PLoS Genet. 6, e1001191. http://dx.doi.org/10.1371/journal.pgen.1001191.

Kobayashi, Y., Horie, M., Nakano, A., Murata, K., Itou, T., Suzuki, Y., 2016. Exaptation of bornavirus-like nucleoprotein elements in afrotherians. PLoS Pathog. 12, e1005785. http://dx.doi.org/10.1371/journal.ppat.1005785.

Kryukov, K., Imanishi, T., 2016. Human contamination in public genome assemblies. PLoS One 11, e0162424. http://dx.doi.org/10.1371/journal.pone.0162424.

Liu, H., Fu, Y., Jiang, D., Li, G., Xie, J., Cheng, J., Peng, Y., Ghabrial, S.A., Yi, X., 2010. Widespread horizontal gene transfer from double-stranded RNA viruses to eukaryotic nuclear genomes. J. Virol. 84, 11876–11887. http://dx.doi.org/10.1128/JVI.00955-10.

Mager, D.L., Stoye, J.P., 2015. Mammalian endogenous retroviruses. Microbiol. Spectr. 3http://dx.doi.org/10.1128/microbiolspec.MDNA3-0009-2014. MDNA3-0009-2014.

Mayer, J.A., Taylor, F.J.R., 1979. A virus which lyses the marine nanoflagellate Micromonas pusilla. Nature 281, 299–301. http://dx.doi.org/10.1038/281299a0.

Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., Hingamp, P., Goto, S., Ogata, H., 2016. Linking virus genomes with host taxonomy. Viruses 8, 66. http://dx.doi.org/10.3390/v8030066.

Moreau, H., Piganeau, G., Desdevises, Y., Cooke, R., Derelle, E., Grimsley, N., 2010. Marine prasinovirus genomes show low evolutionary divergence and acquisition of protein metabolism genes by horizontal gene transfer. J. Virol. 84, 12555–12563. http://dx.doi.org/10.1128/JVI.01123-10.

Nakagawa, S., Takahashi, M.U., 2016. gEVE: a genome-based endogenous viral element database provides comprehensive viral protein-coding sequences in mammalian genomes. Database (Oxford). http://dx.doi.org/10.1093/database/baw087. 2016. baw087.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., Pruitt, K.D., 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44, D733–D745. http://dx.doi.org/10.1093/nar/gkv1189.

Parrish, N.F., Fujino, K., Shiromoto, Y., Iwasaki, Y.W., Ha, H., Xing, J., Makino, A., Kuramochi-Miyagawa, S., Nakano, T., Siomi, H., Honda, T., Tomonaga, K., 2015. piRNAs derived from ancient viral processed pseudogenes as transgenerational sequence-specific immune memory in mammals. RNA 21, 1691–1703. http://dx.doi.org/10.1261/rna.052092.115.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313. http://dx.doi.org/10.1093/bioinformatics/btu033.