Discovery and characterization of mammalian endogenous parvoviruses

Citation for published version:

Digital Object Identifier (DOI):
10.1128/JVI.01732-10

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Early version, also known as pre-print

Published in:
Journal of Virology

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Discovery and Characterization of Mammalian Endogenous Paroviruses

Amit Kapoor, Peter Simmonds and W. Ian Lipkin


Updated information and services can be found at:
http://jvi.asm.org/content/84/24/12628

**SUPPLEMENTAL MATERIAL**
Supplemental material

**REFERENCES**
This article cites 36 articles, 21 of which can be accessed free at: http://jvi.asm.org/content/84/24/12628#ref-list-1

**CONTENT ALERTS**
Receive: RSS Feeds, eTOCs, free email alerts (when new articles cite this article), more

Information about commercial reprint orders: http://journals.asm.org/site/misc/reprints.xhtml
To subscribe to another ASM Journal go to: http://journals.asm.org/site/subscriptions/
Discovery and Characterization of Mammalian Endogenous Parvoviruses

Amit Kapoor,1* Peter Simmonds,2 and W. Ian Lipkin1

Center for Infection and Immunity, Columbia University, New York, New York 10032,1 and University of Edinburgh, Edinburgh, Scotland, United Kingdom2

Received 16 August 2010/Accepted 5 October 2010

Public databases of nucleotide sequences contain exponentially increasing amounts of sequence data from mammalian genomes. Through the use of large-scale bioinformatic screening for sequences homologous to exogenous mammalian viruses, we found several sequences related to human and animal parvoviruses (PVs) in the Parvovirus and Dependovirus genera within genomes of several mammals, including rats, wallabies, opossums, guinea pigs, hedgehogs, African elephants, and European rabbits. However, phylogenetic analysis of these endogenous parvoviruses (EnPV) sequences demonstrated substantial genetic divergence from exogenous mammalian PVs characterized to date. Entire nonstructural and capsid gene sequences of a novel EnPV were amplified and genetically characterized from rat (Rattus norvegicus) genomic DNA. Rat EnPV sequences were most closely related to members of the genus Parvovirus, with >70% and 65% amino acid identities to nonstructural and capsid proteins of canine parvovirus, respectively. Integration of EnPV into chromosome 5 of rats was confirmed by PCR cloning and sequence analysis of the viral and chromosomal junctions. Using inverse PCR, we determined that the rat genome contains a single copy of rat EnPV. Considering mammalian phylogeny, we estimate that EnPV integrated into the rat genome less than 30 million years ago. Comparative phylogenetic analysis done using all known representative exogenous parvovirus (ExPV) and EnPV sequences showed two major genetic groups of EnPVs, one genetically more similar to genus Parvovirus and the other genetically more similar to the genus Dependovirus. The full extent of the genetic diversity of parvoviruses that have undergone endogenization during evolution of mammals and other vertebrates will be recognized only once complete genomic sequences from a wider range of orders, species, and types of animals become available.

Approximately 8% of the human genome comprises endogenous retroviruses (ERV) (10, 15–17, 21). Molecular characterization of ERV has provided insights into the origin and evolution of their exogenous counterparts and also of their hosts (9, 10, 16, 17). While ERV are classically regarded as “junk” or “selfish” DNA, the maintenance of gene expression of some integrated ERV proviral sequences led to the proposal that ERV may be co-opted in certain cellular functional roles. For example, expression of gag- and env-encoded proteins of human ERV K (HERV-K) in the placenta may play roles in preventing rejection of the fetus (25). Until the recent discovery of endogenous Borna disease virus- and filovirus-like sequences in mammalian genomes (5, 15, 33), ERV were the only endogenous viruses known. The process of endogenization, whereby viral sequences are integrated in the genomes of their hosts (12), occurs when viral nucleic acid as DNA or cDNA integrates into chromosomes of reproductive germ line cells. In contrast to integration of reverse-transcribed proviral sequences of retroviruses, which is part of the retrovirus replicative cycle, integration of viral sequences of other viruses is cell mediated and serendipitous and creates embedded, fossilized genomic elements, incapable of generating infectious viruses (5, 12, 15).

Exogenous parvoviruses (ExPVs) are ubiquitous and can cause a broad spectrum of diseases in animals, including enteritis, panleukopenia, hepatitis, erythrocyte aplasia, immune complex-mediated vasculitis, reproductive failure, and cerebellar ataxia (11). Vaccines against animal parvovirus infections are widely employed (2, 14, 27, 28, 31). The family Paroviridae as currently defined comprises two subfamilies, Densovirusinae and Parovirinae, that infect nonvertebrate and vertebrate hosts, respectively (11). The International Committee on Taxonomy of Viruses (ICTV) has further classified the subfamily Parovirinae into five genera primarily based on phylogenetic analysis: Dependovirus, Bocavirus, Erythrovirus, Parvovirus, and Ampovirus.

Members of the subfamily Parovirinae are small, nonenveloped icosahedral viruses with single-stranded linear DNA genomes that frequently infect animals through the fecal-oral route (11). The genomes of most parvoviruses are nearly 5,000 nucleotides (nt) in length and comprise two transcriptional units, one encoding the capsid proteins and the other the nonstructural proteins. Protein-coding sequences are flanked on each side by noncoding palindromic repeats, also known as inverted terminal repeat (ITR) sequences, that play an important role in viral DNA replication (22, 23). Based on their replication requirements, paroviruses can be classified as either autonomous paroviruses or dependoviruses; the latter require external factors for replication (6). The most extensively studied dependoviruses are the adeno-associated viruses (AAV) that are used as gene therapy vectors. During their replication in the nuclei of infected host cells, the genomes of wild-type AAV integrate in a site-specific manner (chromo-
some 19) (29, 37), resulting in latent infection of host cells. Several recent studies detected the presence of AAV genomes in tissues of humans and nonhuman primates as integrated virus in the host genome and/or in episomal closed circular form (8, 13, 30). As part of their replication cycle, all paroviruses must enter the nuclei of their host cells and generate a double-stranded monomer replicative form. Some animal paroviruses are known to cause persistent infections and long-term shedding (3, 4). Moreover, all parovirus genomes have sequences that would have been missed during similarity-based search to one or more parovirus proteins were considered candidate EnPVs. These candidate EnPV sequences were used to research the database for other homologous sequences that would have been missed during similarity-based search using EnPVs.

### Sequence acquisition and phylogenetic analysis of EnPV

After initial identification of potential EnPV sequences, genomic DNA of rats was used to test for the presence of EnPV (Fig. 1C). An outline of experiments to confirm the existence, integration, and genomic organization of rat EnPV is shown in Fig. S1 in the supplemental material. Rat DNA was obtained from the Rattus norvegicus lung epithelial cell line L2 (ATCC catalog number CCL-149) and from Rattus norvegicus kidney, liver, and brain using Trizol (Invitrogen). EnPV-specific PCR to detect the presence of rat EnPV in different samples targeted the VP gene.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Reference genomic sequence database (refseq_genomic)</th>
<th>NCBI Genomes/Chromosomes</th>
<th>Expressed sequence tags (est database)</th>
<th>Whole-genome shotgun reads (wgs database)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NS</td>
<td>NC_005101 (rat, ch-2), AC_000081 (rat, ch-13), NC_000069 (mouse, ch-3), NC_008806 (opossum, ch-6), NC_008803 (opossum, ch-3), NC_013678 (rabbit, ch-10), NC_009175 (horse, ch-X)</td>
<td>NC_005104 (rat, ch-5), NC_008803 (opossum, ch-3), NC_008806 (opossum, ch-6), NC_008808 (opossum, ch-8), C_013678 (rabbit, ch-10)</td>
<td>DW308476 (rat liver), B577555 (drosophila), C348320 (opossum liver), X610113 (pea aphid), V837559 (aphid), C387312 (pea aphid)</td>
<td>AAB00124553 (Rattus norvegicus), BQ001031875 (Macropus eugenii), BQ001059935 (Macropus eugenii), BQ001031620 (Macropus eugenii), AK02032906 (Cavia porcellus), AK02030352 (Cavia porcellus), AYI01487966 (Echinops telfairi), AGU03013549 (Laxodonta africana), APEF1512673 (Myotis lucifugus), BRN01263281 (Tarsonys transcanus), AAOE1015016 (Tetraodon nigroviridis), AAYZ01294085 (Ochotona princeps), BRP01170809 (Pteropus vampyrus)</td>
</tr>
<tr>
<td>VP1</td>
<td>NC_005104 (rat, ch-5), C_000069 (mouse, ch-3), C_008806 (opossum, ch-6), C_008808 (opossum, ch-8), C_008803 (opossum, ch-3), C_013678 (rabbit, ch-10)</td>
<td>NC_005104 (rat, ch-5), C_008806 (opossum, ch-6), C_008808 (opossum, ch-8), C_008803 (opossum, ch-3)</td>
<td>DW382746 (rat liver), W313020 (opossum tissues), O892248 (bovine brain), O888893 (bovine brain)</td>
<td>ABO01015994 (Macropus eugenii), ABO010033457 (Macropus eugenii), ABO0010193462 (Macropus eugenii), ABO010185939 (Macropus eugenii), AK02030352 (Cavia porcellus), AGV020719236 (Dasyurus novemcinctus)</td>
</tr>
</tbody>
</table>

### Materials and Methods

#### Identification of endogenous paroviruses

We used one reference genome sequence representing one species for each of the five genera included in subfamily Parovirinae to identify genetically related sequences in different NCBI sequence databases (Table 1). NCBI databases used to search for parovirus related sequences included refseq_genomic (genomic entries from NCBI’s Reference Sequence project), NCBI Genomes/Chromosomes (a database with complete genomes and chromosomes from the NCBI Reference Sequence project), est (a database of GenBank, EMBL, and DDBJ sequences from expressed sequence tags), and wgs (a database for whole-genome shotgun sequence entries). Default search criteria for NCBI genomic (protein query search against a translated nonredundant database) were used, except for changing the highest expect score (E = 10−10) to make the similarity search highly stringent. All exogenous parovirus (ExPV) sequences were excluded from the analysis. The animal genomic sequences that showed significant tblastn similarity to NS or VP proteins of paroviruses and respective hosts of origin are shown, ch-2, chromosome 2.

<table>
<thead>
<tr>
<th>Reference genomic sequence database (refseq_genomic)</th>
<th>NCBI Genomes/Chromosomes</th>
<th>Expressed sequence tags (est database)</th>
<th>Whole-genome shotgun reads (wgs database)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NS</td>
<td>NC_005101 (rat, ch-2), AC_000081 (rat, ch-13), NC_000069 (mouse, ch-3), NC_008806 (opossum, ch-6), NC_008803 (opossum, ch-3), NC_013678 (rabbit, ch-10), NC_009175 (horse, ch-X)</td>
<td>NC_005104 (rat, ch-5), NC_008803 (opossum, ch-3), NC_008806 (opossum, ch-6), NC_008808 (opossum, ch-8), C_013678 (rabbit, ch-10)</td>
<td>DW308476 (rat liver), B577555 (drosophila), C348320 (opossum liver), X610113 (pea aphid), V837559 (aphid), C387312 (pea aphid)</td>
</tr>
<tr>
<td>VP1</td>
<td>NC_005104 (rat, ch-5), C_000069 (mouse, ch-3), C_008806 (opossum, ch-6), C_008808 (opossum, ch-8), C_008803 (opossum, ch-3), C_013678 (rabbit, ch-10)</td>
<td>NC_005104 (rat, ch-5), C_008806 (opossum, ch-6), C_008808 (opossum, ch-8), C_008803 (opossum, ch-3)</td>
<td>DW382746 (rat liver), W313020 (opossum tissues), O892248 (bovine brain), O888893 (bovine brain)</td>
</tr>
</tbody>
</table>
on 2% agarose gel. PCR products showing positive bands of approximately 126 bp (Fig. 1C), corresponding to the VP gene fragment of rat EnPV, were purified using a PCR purification kit (Qiagen) and directly sequenced from both directions. To acquire the complete genome of rat EnPV, PCR primers were designed to amplify five different overlapping fragments of genome using nested PCR conditions (see Table S1 in the supplemental material). All reactions for genomic PCR were performed in PCR buffer suitable for amplification of GC-rich genomic sequences as specified by the manufacturer (Takara; catalog no. RR02AG). All PCRs for junction and genomic PCRs used conditions similar to those described for inverse PCR below, except that the annealing temperatures for PCRs were 6°C below the average melting temperatures of the two primers. PCR products were sequenced from both directions. All EnPV sequences generated by PCR and extracted from the NCBI database were translated to proteins in silico for phylogenetic analysis.

Integration site analysis and copy number. To confirm the integration of EnPV in the host genome, two different approaches were used (see Fig. S1 in the supplemental material). First, the sequence information available in the NCBI database was used to design the primer pairs targeting the junction of the viral gene and host chromosome. Amplification and sequencing of the junctional region (junctional PCR) were then pursued to define the integration site sequence (see Fig. S1 and Table S1 in the supplemental material). Second, to independently confirm the integration site sequence and to determine the copy number of EnPV per host genome, we used an inverse PCR approach (Fig. 1C; see Fig. S1 in the supplemental material). The genomic DNA of a rat was digested with multiple enzymes, rejoined under conditions designed to favor intramolecular ligation, and employed as the template for PCR amplification (Fig. 1C; see Fig. S1 in the supplemental material). Restriction enzymes used were AvrII, SpeI, BspHI, AflII, and KpnI. Restriction enzyme-digested PCR products were self-ligated in reaction conditions favoring intramolecular ligation (15). The first round of inverse PCR used an inverse forward primer targeting the rat EnPV NS gene (INP-NS-F1, 5'-T*T*C*ACAGCCAAACTCAGGGCTCCA CATAC-3' (H11032)) and an inverse reverse primer targeting the rat EnPV VP gene (INP-VP-R1, 5'-C*T*G*TTCCCTGGTCCCAGGTACTTGTA-3' (H11032)). Both first-round primers were made exonuclease resistant through phosphorothioation of the first three bases at the 5' end. Similarly, the second round of inverse PCR products were self-ligated in reaction conditions favoring intramolecular ligation (15). The first round of inverse PCR used an inverse forward primer targeting the rat EnPV NS gene (INP-NS-F1, 5'-T*T*C*ACAGCCAAACTCAGGGCTCCA CATAC-3') and an inverse reverse primer targeting the rat EnPV VP gene (INP-VP-R1, 5'-C*T*G*TTCCCTGGTCCCAGGTACTTGTA-3'). Both first-round primers were made exonuclease resistant through phosphorothioation (+) of the first three bases at the 5' end. Similarly, the second round of inverse PCR used primers INP-NS-F2 (5'-CTCAGGGCTCCACATACAGG-3') and INP-VP-R2 (5'-CCTGGTCCCCAGGTACTTGTA-3'). For the second round of nested PCR, 5 µl of restriction enzyme-digested and circularized DNA templates was mixed with 25 µl GC Buffer I (Takara; catalog no. RR02AG), 6 µl dNTP solution (10 mM), 25 pmol forward and reverse primers, 0.5 µl LA Tag DNA polymerase (Takara; catalog no. RR002A), and 11.5 µl DEPC-treated water, in a total reaction volume of 50 µl. The reaction was performed using initial
en successful amplification of the targeted EnPV sequence in each PCR. 1.5% agarose gels. All PCRs included a non-DNase-treated positive control to classification products were visualized after size fractionation by electrophoresis in 1.5% agarose gel (Fig. 1C) and sequenced from both ends.

**EnPV expression analysis.** Total nucleic acids were extracted from all tissue samples of rats using Trizol (Invitrogen). To determine the presence of EnPV-derived transcript RNA, total nucleic acid was digested with DNase followed by PCR targeting the amino terminus of the VP protein of the rat EnPV. Amplification products were visualized after size fractionation by electrophoresis in 1.5% agarose gels. All PCRs included a non-DNase-treated positive control to confirm successful amplification of the targeted EnPV sequence in each PCR.

**RESULTS**

**Identification of endogenous paroviruses.** The sequence of the structural protein (VP) of canine parvovirus (NC_041400) was used as the template in a tblastn search to identify genetically related sequences in the NCBI database. Our first sequence similarity search results indicated the presence of parovirus VP-like sequences in several mammalian species (Table 1). All the genomic sequences were extracted from the database, conceptually translated, and aligned to VP protein sequences of exogenous paroviruses (ExPV) for identifying the different fragments of the VP gene in each EnPV (Fig. 1A). Results of our analysis suggested that rat (*Rattus norvegicus*) chromosome 5 (NC_005104) contained the coding sequence for almost the entire structural protein and that this sequence was similar to those encoding VP proteins of different members of genus Parovirus (Fig. 1B).

**Characterization of EnPV in rats (rat EnPV in chromosome 5).** The tblastn search using canine parvovirus (NC_001539) nonstructural (NS) protein against the NCBI Genomes/Chromosomes database showed the presence of a highly similar sequence in *Rattus norvegicus* chromosome 5. The translated sequence comprised 281 amino acids (aa) with >70% identity to the N-terminal half of canine parvovirus NS proteins and an expected score of 2e-98. To determine the location of the rat EnPV nonstructural protein coding sequence in chromosome 5, the corresponding nucleotide sequence was extracted and used to perform a BLAT search against databases available at the University of California Santa Cruz Web server (http://genome.ucsc.edu/) (20). DNA BLAT works by keeping an index of the entire genome in memory and is designed to quickly find sequences of 95% and greater similarity containing 25 bases or more (19). The BLAT results showed the location of rat EnPV integration between chromosomal loci 5q22 and 5q24, nearly in the middle of chromosome 5 (Fig. 2A). Nucleotide alignment of the homologous chromosome in mice, humans, and dogs showed this region to be nonconserved among these species. Detailed analysis of the neighboring sequence resulted in characterization of three other large open reading frames (ORFs), whose products showed highly significant similarity to virion proteins (VP) of canine parvovirus (Fig. 2A). The NS and VP coding sequences of rat EnPV were found in close proximity (within 4,000 bp) on chromosome 5. However, in contrast to the genomic organization of known ExPVs, the NS protein ORF of rat EnPV was in reverse orientation with respect to the capsid ORFs (Fig. 2A). Phylogenetic analysis of the 281-aa rat EnPV NS sequence showed the coding sequence to be equidistant from the NS coding sequences of canine, porcine, mouse, and rat ExPVs. The NS protein of rat EnPV showed >70% amino acid identity to the NS proteins of different paroviruses (Fig. 2B). Protein sequences encoded by three different ORFs of rat EnPVs related to the parovirus VP protein were aligned to known ExPV proteins and combined to deduce a 744-aa protein representing the product of the complete structural gene ORF (Fig. 1B). Phylogenetic analysis of this 744-aa protein showed that the full-length VP protein of rat EnPV was most closely related to the capsid protein of canine parvovirus, with >65% protein identity (Fig. 1B and 2B).
EnPV sequences related to VP protein were more common in various rodents and marsupials than in other mammalian groups like primates, canines, and ruminants (Table 1). Whereas in mice, rats, rabbits, and armadillos only one EnPV sequence was detected, multiple diverse EnPV sequences were found in opossums, wallabies, and guinea pigs. Although mice and rats are the most closely related species in which EnPVs were found, the mouse EnPV was genetically closest to members of the Dependovirus genus, in contrast to the rat EnPV (genetically closest to members of the Parvovirus genus).

**Phylogeny of mammalian EnPV.** To determine the phylogenetic relationships of EnPV sequences to exogenous parvoviral sequences, all genomic sequences were translated in silico and joined manually using reference protein sequences of the most closely related known ExPV. All protein alignments were generated using the ClustalW program implemented in MEGA4.1. To accommodate the missing partial protein sequences of EnPV and the presence of stop codons and high genetic diversity between sequences, we used the BLOSUM protein weight matrix after reducing the gap opening penalty to 5 and gap extension penalty to 0.1 for both pairwise and multiple sequence alignments. Phylogenetic analysis done using default criteria also resulted in a tree topology similar to that generated using the modified criteria mentioned above.

Minimal manual editing of EnPV sequences was performed to ensure an unbiased analysis. Although many of the protein sequences are of different lengths in both the NS and VP protein alignments, it is now well known that sequences of different lengths can be accurately placed in phylogenetic trees (35).

On the basis of their genetic relatedness to paroviruses, the EnPV sequences extracted from the NCBI database were divided into two groups. The first group included the genomic sequences that showed significant similarity to VP genes of paroviruses, and the second group included genomic sequences that showed similarity to NS genes of paroviruses (Table 1). EnPV sequences similar to VP genes of paroviruses were further classified into two heterogeneous phylogenetic clusters: one genetically closer to members of the Parvovirus genus, named EnPV group A (EnPV-A), and the other genetically closer to members of the Dependovirus genus, named EnPV group B (Fig. 3A). EnPV-A included virus sequences derived from genomes of wallabies, guinea pigs, opossums, and rats. All the members of EnPV-A, excluding the rat EnPV, clustered together to form a distinct clade whose members were genetically closer to each other than to known ExPVs. The capsid protein of rat EnPV was most closely related to the capsid protein of canine parvovirus and showed 65% amino acid identity to the...
canine parvovirus protein. The genus Parvovirus includes ExPVs known to cause diseases in mouse, rat, porcine, and canine hosts. Intriguingly, the rat EnPV capsid protein was more closely related to those of canine ExPVs than to those of rat or mouse ExPVs (Fig. 1B and 2B).

The NS gene-related EnPV sequences could be classified into four heterogeneous clusters, tentatively named EnPV groups A to D (Fig. 3B). Two NS protein clusters, EnPV-A and -B, were similar to VP protein clusters described above and were genetically closest to members of genera Parvovirus and Dependovirus, respectively. EnPV sequences that clustered together with the NS genes of human parvovirus B19 and human parvovirus 4 are in EnPV-C and included EnPV sequences found in bats, horses, and puffer fish. A few EnPV NS protein sequences that were highly diverse and equidistant from those of all known ExPV genera formed an outlier group, which we designated EnPV group D. Like the phylogeny based on VP protein sequences, the phylogeny based on NS proteins of EnPV in rats, opossums, wallabies, and guinea pigs indicated genetic similarity to members of genus Parvovirus (EnPV-A). Although not conclusive, this finding is compatible with endogenization of both parvovirus genes together from the same exogenous virus.

Expression of EnPV in hosts. EnPV sequences were found in the NCBI expressed sequence tag (EST) database, indicating that they were expressed at the mRNA level in host animals. Indeed, in rats we found expression of sequence tags representing two different genomic regions of rat EnPV (Table 1). To independently test for the expression of EnPV, we looked for the presence of viral RNA transcripts in rat tissues. Total nucleic acid was extracted, and a fraction of the sample was digested with RNase-free DNase to ensure amplification of only RNA. Reverse transcription-PCR (RT-PCR) assays targeting the structural gene of rats showed no amplification products. Additionally, no EnPV transcripts were found in actively replicating rat lung epithelial cells (rat L2 cells). How-
ever, results of these experiments cannot rule out the transient or more-tissue-restricted expression of EnPV in rats.

**DISCUSSION**

Our data indicate that parvoviruses are integrated into the genomes of a wide range of hosts, including rats, wallabies, opossums, guinea pigs, hedgehogs, African elephants, and European rabbits. The three other viruses known to be integrated into host genomes, retroviruses, Bornavirus (5, 15), and filoviruses (5, 33), have RNA genomes and require DNA replication intermediates for integration. Although parvoviruses have single-stranded DNA genomes, their replication within the host cell nucleus requires synthesis of a double-stranded monomer replicative form that could facilitate integration into host genomes (6). Consistent with this model, several EnPVs characterized here showed close proximity in the integration sites of NS and VP gene sequences, suggesting endogenization of complete viral DNA genomes rather than independent RNA transcripts encoding different genes. However, we failed to detect the noncoding sequences of genetically related parvoviruses near capsid or nonstructural genes of EnPV found in rats. Interestingly the endogenization of EnPVs is likely to be more recent (<20 to 30 million years ago) than that of filoviruses, since the EnPVs from genetically related host species (rats and mice) are not orthologous like their host animals (1, 5, 33).

Endogenous viruses are fossilized exogenous viruses (5, 12). Over time host genetic mechanisms such as mutation and genomic rearrangements lead to the eventual degeneration of endogenous virus genes; attrition of their capacity to express viral gene products may indeed be a benefit to the host (5). Detailed analysis of rat EnPV genetic diversity and genomic organization suggests that rat EnPV may represent an early progenitor of the genus Parovirus. The absence of rat EnPV-like sequences in the canine and porcine genomes suggests that either rats were the only natural host of this virus or endogenization of this virus in the genomes of the other related mammalian species was unsuccessful. All known parvoviruses have similar genomic organizations wherein the NS and VP genes are present in the same orientation. In rat EnPV the NS and VP genes are oriented oppositely in spite of being in close proximity, which suggests that, after the integration of an intact genome, small-scale, local genome rearrangements disrupted the integrated parvoviral sequence. The presence of a LINE-like repeat sequence near the truncated NS gene in rat EnPV suggests a role for these elements in integration of viral genes, similar to the role these elements play in generating cellular pseudogenes (5). However, the presence of both parovirus NS and VP genes in close proximity in several EnPVs suggests that direct insertion of double-stranded viral DNA replication intermediates into the chromosome is also plausible.

Endogenous retroviruses may protect the host from infection and disease by similar exogenous viruses (7), a resistance mechanism that has also recently been proposed to underlie the integration of poxvirus and dicistrovirus genome sequences in insects (26). None of the endogenous viruses described thus far has a genetically identical exogenous counterpart (9, 10). Similarly, we cannot ascertain whether the presence of rat EnPV protects rats from infection by closely related exogenous viruses as no genetically similar exogenous virus has been discovered. However, we do know that the presence of rat EnPV does not provide protection against several known rat paroviruses that naturally infect rats and whose VP and NS proteins are >25 to 35% different in amino acid sequence (Fig. 3). Although we did not detect expression of rat EnPV mRNA in tissue samples from rat organs or a continuous cell line, expression may be transient. We also found sequences of expressed sequence tags in the NCBI database, suggesting the expression of EnPV in rats, opossums, and aphids (Table 1).

The observed genetic diversity between NS (>25%) and VP (>35%) proteins of rat EnPV and known paroviruses is consistent with faster evolution rate of the parovirus VP protein by as much as 10%. Based on comparative phylogenetic analysis, most of the EnPVs are either genetically close to members of the genus Parovirus or Dependovirus (24). In the opossum we observed that all three different EnPV VP coding sequences integrated in different chromosomes belong to EnPV-A. In the absence of a mechanism for expression, reverse transcription, and reintegration of viral sequences (as documented for ERV and LINE), this suggests multiple endogenization events over time. Indeed, in the wallaby, we found integrated VP sequences of both the EnPV-A and EnPV-B clusters. Unlike other endogenous viruses reported (5, 9, 15, 34), the EnPV-like sequences were not detected in genomes of primates and were more prevalent in small mammals and marsupials. The genetic diversity of the combined data set of EnPV sequences examined and lack of homologous integrated sequences in related rodent species suggest that multiple endogenization events occurred after the speciation of rodents 25 to 35 million years ago (1). The genetic relatedness between the EnPVs and their host species is also discordant, consistent with endogenization after host speciation. For example, the EnPVs of rats and mice are genetically more distant than EnPVs of mice and rabbits. Considering that rats and mice separated 20 to 40 million years ago (1), the endogenization time of rat EnPV in chromosome 5 appears to be more recent. Interestingly we found no EnPV genetically related to members of erythroviruses, avian paroviruses, or bocaviruses, with the exception of the NS gene sequences of EnPVs found in puffer fish chromosomes that were genetically closest to parovirus B19 (Fig. 3B). Bocaviruses are unique among paroviruses in that they contain a middle ORF called NP1 between the NS and VP coding regions (18, 24). The tblastn search using the bocavirus NP1 protein did not detect any significantly related sequence in the NCBI database.

In summary, we report paroviruses as the first small DNA viruses that exist as endogenous viruses integrated in the host chromosomes and demonstrate that they are widespread among several mammalian species. Our observations based on genetic diversity suggest that EnPVs were generated as results of multiple endogenization events. We believe the EnPVs are likely to be more widely dispersed than described here and that their complete diversity will be recognized only as genomic sequences of more mammalian species become available.
ACKNOWLEDGMENTS

We thank Brian Hjelle for helpful comments, Mady Hornig for critical reagents, and Natasha Mehta and Natasha Qaisar for technical assistance.

Our work was supported by National Institutes of Health grant awards (AI090196A, AI079231, AI57158, AI070411, and EY017404) and by an award from the Department of Defense.

REFERENCES


Downloaded from jvi.asm.org on June 8, 2013 by UNIVERSITY OF EDINBURGH