JVI Journals.ASM.org

Multiple Groups of Endogenous Epsilon-Like Retroviruses Conserved across Primates

Katherine Brown,[a] Richard D. Emes,[a,b] Rachael E. Tarlinton[a]

School of Veterinary Medicine and Science[a] and Advanced Data Analysis Centre,[b] University of Nottingham, Sutton Bonington Campus, Loughborough, Leicestershire, United Kingdom

ABSTRACT

Several types of cancer in fish are caused by retroviruses, including those responsible for major outbreaks of disease, such as walleye dermal sarcoma virus and salmon swim bladder sarcoma virus. These viruses form a phylogenetic group often described as the epsilonretrovirus genus. Epsilon-like retroviruses have become endogenous retroviruses (ERVs) on several occasions, integrating into germ line cells to become part of the host genome, and sections of fish and amphibian genomes are derived from epsilon-like retroviruses. However, epsilon-like ERVs have been identified in very few mammals. We have developed a pipeline to screen full genomes for ERVs, and using this pipeline, we have located over 800 endogenous epsilon-like ERV fragments in primate genomes. Genomes from 32 species of mammals and birds were screened, and epsilon-like ERV fragments were found in all primate and tree shrew genomes but no others. These viruses appear to have entered the genome of a common ancestor of Old and New World monkeys between 42 million and 65 million years ago. Based on these results, there is an ancient evolutionary relationship between epsilon-like retroviruses and primates. Clearly, these viruses had the potential to infect the ancestors of primates and were at some point a common pathogen in these hosts. Therefore, this result raises questions about the potential of epsilonretroviruses to infect humans and other primates and about the evolutionary history of these retroviruses.

IMPORTANCE

Epsilonretroviruses are a group of retroviruses that cause several important diseases in fish. Retroviruses have the ability to become a permanent part of the DNA of their host by entering the germ line as endogenous retroviruses (ERVs), where they lose their infectivity over time but can be recognized as retroviruses for millions of years. Very few mammals are known to have epsilon-like ERVs; however, we have identified over 800 fragments of endogenous epsilon-like ERVs in the genomes of all major groups of primates, including humans. These viruses seem to have circulated and infected primate ancestors 42 to 65 million years ago. We are now interested in how these viruses have evolved and whether they have the potential to infect modern humans or other primates.

Epsilonretroviruses are a genus of retrovirus usually associated with fish (1). Several common proliferative diseases in commercially important fish species are caused by these viruses. In the walleye (*Sander vitreus*), a species of perch that is an important source of sport fishing revenue in Canada and the northern United States (2), up to 30% of some populations are affected annually by skin lesions resulting from the epsilonretrovirus walleye dermal sarcoma virus (WDSV) and up to 10% by skin lesions resulting from the epsilonretrovirus walleye epidermal hyperplasia virus (WEHV) (3). Outbreaks of sarcoma in the Atlantic salmon (*Salmo salar*), a species that makes up almost 2.5% of worldwide aquaculture production, have been attributed to Atlantic salmon swim bladder sarcoma virus (SSSV), which is genetically similar to the epsilonretroviruses (4, 5). Other diseases in fish and amphibians have also been provisionally linked to epsilon-like retroviruses (6, 7). However, no epsilon-like retroviruses causing disease in mammals or birds have been identified.

To date, evidence from endogenous retroviruses (ERVs) has confirmed these viruses as primarily causing infections of fish. ERVs are retroviruses that have integrated into germ line, rather than somatic, cells and are therefore transmitted vertically from parents to offspring and can become a permanent part of the genome of their host. ERVs are degraded over time by mutation and become inactive but remain detectable in their host genome millions of years after integration. This means they provide valuable insight into the retroviruses that a species has been exposed to deep in its evolutionary history. Epsilon-like ERVs have been found in a diverse range of fish and amphibian genomes, suggesting a longstanding relationship with both these groups (8–10). These retroviruses are thought to be the result of multiple integration events taking place over many millions of years, including several relatively recent insertions (8–10).

Genome-wide screening for all genera of retroviruses has been performed in many species of mammals and birds (11–13), revealing a rich diversity of gammaretroviruses, a genus closely related to epsilonretroviruses. However, epsilon-like ERVs have not been identified in most mammals. Some epsilon-like insertions have previously been found in the human genome. Tristem (14) iden-

Received 9 April 2014 **Accepted** 23 July 2014

Published ahead of print 20 August 2014

Editor: K. L. Beemon

Address correspondence to Rachael E. Tarlinton, rachael.tarlinton@nottingham.ac.uk.

Supplemental material for this article may be found at http://dx.doi.org/10.1128/JVI.00966-14.

doi:10.1128/JVI.00966-14

tified a group of approximately 70 highly degenerate sequences clustering with nonmammalian retroviruses in the human genome, named the HERV.HS49C23 group and later subdivided into the HERV-L(b), HERV-R(c), HERV(AC0956774), and ERV(AC018462) families (15). These insertions were described as being more closely related to WDSV than to the gammaretroviruses. Oja et al. (16) identified 12 epsilon-like insertions in the human genome, and in our previous work (17), we characterized a group of epsilon-like ERVs in the horse genome, using a newly developed bioinformatics pipeline.

We have now screened 32 species of primates, rodents, lagomorphs (rabbits and pikas), and birds for epsilon-like ERVs using this pipeline, and unexpectedly, we have identified several groups of epsilon-like ERVs that appear to be ubiquitous in primates. The integration patterns and phylogeny of these primate epsilon-like (PE) ERVs suggest that they entered the genome of a common ancestor of Old and New World monkeys at least 40 million years ago. These results raise several important questions about the origin and evolutionary history of the epsilonretroviruses and their relatives, their relationship with gammaretroviruses, and their potential for cross-species transmission.

MATERIALS AND METHODS

Genome screening. A database of 382 *gag*, 670 *pol*, and 356 *env* amino acid sequences was built to represent the diversity of known exogenous and endogenous retroviruses. The viruses included in this data set are listed in full in Table S1 in the supplemental material. Details of the genomes screened in this analysis are listed in Table S2 in the supplemental material. All genomes were downloaded on 3 August 2013 from RefSeq release 57, NCBI Genome, or Ensembl release 70. For genomes not assembled into chromosomes, scaffolds were concatenated into approximately chromosome-length strings for ease of analysis and later traced back to their original scaffold. Candidate ERV regions were identified using the Exonerate algorithm (18) and formatted using the Perl pipeline (https://github.com/ADAC-UoN/predict.genes.by.exonerate.pipeline) under the protein2genome model with a minimum hit length of 200 amino acids without introns. When predicted genes overlapped, the gene with the highest Exonerate score was selected.

ERV DNA fragments predicted by Exonerate were verified using a TBLASTX (19) search of the untranslated version of the input database described above. Sequences producing an alignment greater than 100 amino acids in length and with greater than 40% amino acid identity with a sequence in the input database (the thresholds are based on reference 20) were classified as ERVs. These sequences were aligned individually with each of the original untranslated input sequences listed in Table S1 in the supplemental material using EMBOSS water (21), which is based on the Smith-Waterman algorithm (22) and finds regions of local similarity among otherwise dissimilar sequences. Sequences were categorized into genera according to their highest alignment score. Sequences that showed the highest similarity to the epsilon and epsilon-like retroviruses were assigned to a provisional epsilon-like data set. All the sequences in this data set were divided by host, and their nucleotide sequences were aligned with those of 34 known epsilon and epsilon-like retroviruses and 41 diverse gammaretroviruses using the localpair setting of MAFFT (23) with 1,000 iterations (these sequences are highlighted in Table S1 in the supplemental material). This alignment technique and these settings were also used for all subsequent multiple-sequence alignments. Maximum-likelihood phylogenetic trees were built for the alignments using PHYML (24) under the GTR model with aLRT branch support, no invariable sites, optimized across site rate variation, and with optimized tree topology. PHYML and these settings were also used for all subsequent tree building. Only sequences clustering within a monophyletic group of epsilon and epsilon-like retroviruses, distinct from the gammaretroviruses, with branch support greater than 75% were kept in the data set.

Comparison between primate genomes. The Compara EPO six-primate alignment (C6P) (Ensembl release 74), an alignment of the DNA sequences of human, chimpanzee, gorilla, orangutan, rhesus macaque, and marmoset genomes, was screened for loci containing an epsilon-like ERV *pol* gene fragment in at least one host, and sequences from these loci were extracted. If there was at least 75% sequence identity between the ERV sequence and the sequence of any host within the ERV region, excluding gaps, the ERV was considered to be present in the host. All ERV sequences for each locus were extracted to form a data set of epsilon-like ERV fragments in these six primates. Sequences from all hosts at each locus were aligned, and PHYML phylogenetic trees were built for each locus. A consensus supertree representing all loci was built using CLANN (25). This analysis was repeated with loci divided according to the families described below.

Consensus nucleotide sequences for each locus from the C6P were generated using the alignments described above and the ambigcons function of EMBOSS (21). Ambiguous characters were then replaced in equal proportions with each of the bases represented by the character. Sites with gaps in the majority of sequences were excluded from the consensus. This method was also used to build all subsequent consensus sequences. All consensus sequences were combined into a 7,426-bp multiple-DNA alignment (including multiple gaps due to the degeneracy of the sequences). This alignment was used to build a phylogenetic tree, and the sequences were grouped according to this phylogeny. Each group was aligned and used to build a group consensus sequence. All group consensus DNA sequences were aligned with those of 38 known epsilon and epsilon-like retroviruses, with human ERV I, the closest known gammaretrovirus to the epsilonretroviruses (10), as the outgroup, forming a 5,510-bp multiple alignment. A phylogeny was built from this alignment.

Candidate Exonerate sequences from species outside the six primate species in the Compara six-primate alignment were aligned one by one with these group consensus sequences using EMBOSS water and assigned to a group according to their highest alignment score.

Genome characterization. To isolate long terminal repeats (LTRs), 8,000 bp on either side of the *pol* gene region from each host at each locus was extracted. The regions from the two sides were then aligned to each other using EMBOSS water (21), which was then used to identify the subsection of this alignment with the highest alignment score. Sequences within this subsection from either side of the *pol* gene that shared 75% sequence similarity, were between 6,000 and 15,000 bp apart, and were between 300 and 1,500 bp in length were isolated as candidate LTRs. These thresholds are based on the range of retroviral genome sizes and LTR lengths listed in reference 26. These candidate regions were classified using CENSOR (27). Sequence pairs classified as ERV LTRs were then used as query sequences and aligned back to all the 8,000-bp regions flanking *pol* genes, again using EMBOSS water, and any new sequences identified were added to the data set. Loci were dated using the following equation: $t = k/2N$, where t is time, k is divergence (the number of sites at which the LTRs differ over the LTR alignment length), and N is the neutral substitution rate of the host, assumed here to be the human neutral substitution rate of 4.5×10^{-9} substitutions per site per year. This is a common ERV dating technique (used, for example, in references 1, 11, and 28). For loci with recognizable LTRs, human sequences were extracted and aligned to each other and clustered using a PHYML phylogenetic tree. The human LTRs identified here were used as probes for a genome-wide BLAT search of the human genome (29), using the UCSC server and a threshold of greater than or equal to 75% sequence identity and 300 bp in length (as described above).

For the loci with recognizable LTRs, the 5′ and 3′ limits of the LTR provide the full span of the ERV, meaning other features of the ERVs could be identified and characterized. The regions between the LTRs were translated in all six reading frames to identify any potential open reading frames (ORFs). The regions between the LTRs and the *pol* regions were

TABLE 1 Number of epsilon-like endogenous retroviruses of each type (PE1 to PE3) identified in each host species[a]

Species	Group	No. ofepsilon-likeERVs			
		PE1	PE2	PE3	Total
Human	Ape	50	25	6	81
Bonobo	Ape	33	26	4	63
Chimpanzee	Ape	45	23	6	74
Gorilla	Ape	46	22	5	73
Orangutan	Ape	38	20	6	64
Gibbon	Ape	19	26	4	49
Baboon	Old World monkey	29	26	2	57
Crab-eating macaque	Old World monkey	21	23	3	47
Rhesus macaque	Old World monkey	39	20	6	65
Marmoset	New World monkey	31	15	4	50
Squirrel monkey	New World monkey	21	13	2	36
Tarsier	Prosimian	1	8	0	9
Aye-aye	Prosimian	39	49	25	113
Lemur	Prosimian	16	15	8	39
Bushbaby	Prosimian	0	3	3	6
Chinese tree shrew	Tree shrew	5	11	0	16
Northern tree shrew	Tree shrew	8	4	0	12
Total		441	329	84	854

[a] Details of hosts and genome builds can be found in Table S2 in the supplemental material. The highlighted species are those included in the C6P alignment.

also compared, using BLASTX (19), to the UNIPROT database to identify any candidate *gag* or *env* genes and to a local database containing the WSDV accessory gene sequences (from GenBank accession number NC_001867) to identify sequences resembling these genes. All regions showing significant similarity to any Gag, Env, or accessory gene sequences were examined individually, aligned with the appropriate gene from WDSV, and aligned with each other to establish if any degenerate ERV-derived sequences were present.

Comparison with other mammals. The positions of the *pol* genes in the genomes of humans and chimpanzees with recognizable LTRs identified in all six primate species were compared to the Compara 37-mammalian-genome alignment (C37M) (Ensembl release 74) to ascertain if the loci were conserved in nonsimian primates or outside the primates (as described above for the C6P alignment). The regions of all genomes aligning with the human and chimpanzee epsilon-like *pol* gene fragments were extracted. For each host, the percentage of sites in each genome with a base identical to that of the ERV was calculated. For each species where no ERV was apparent, a 16,000-bp fragment of the alignment was isolated from each locus, encompassing the site where the ERV was expected and the flanking sequence. A TBLASTN analysis was performed on these fragments, using the consensus LTR sequences, *pol* gene sequences, and *env* sequence as probes, to identify solo LTRs or any other ERV fragments that might suggest deletion of the ERV.

RESULTS

Our analysis identified 854 *pol* gene sequences (821 using the Exonerate pipeline and 33 more in the locus-by-locus analysis) that form a reliable phylogenetic cluster within the epsilon and epsilon-like retroviruses. The sequences ranged from 568 to 2,798 nucleotides in length, with a mean of 993 bp. These sequences were all found in primates and tree shrews (Table 1). Primates are generally divided into four major groups, as follows: apes (humans, chimpanzees, gorillas, orangutans, and gibbons), Old World monkeys (monkeys native to Africa and Asia), New World monkeys (monkeys native to Central and South America), and prosimians (tarsiers, lemurs, bushbabies, and lorises) (30). Tree shrews are the closest living relatives of modern primates (30). Epsilon-like insertions were identified in all of these groups (Table 1). No epsilon-like insertions were found in rodents, lagomorphs, or birds.

The C6P alignment allows comparisons between specific loci in the genomes of 6 of the 15 species of primates screened here: four apes, one Old World monkey, and one New World monkey. The 407 epsilon-like ERV sequences we identified in these six species fell at 87 loci. The retrovirus was found in the same position in all six C6P species at 36 of these loci and in three or more species at 75 loci. For the remainder, some species had the retrovirus and some did not; however, there was insufficient information to distinguish between empty ERV insertion sites, solo LTRs, and a lack of sequence data, due to poor alignment quality at and around the locus.

For each of the 87 loci identified in the C6P analysis, a consensus sequence representing the locus was produced. Phylogenetic analysis showed that these consensus sequences fall into three clear families, provisionally named primate epsilon-like 1 to primate epsilon-like 3 (PE1 to PE3) (Fig. 1). A consensus sequence was generated for each family based on this information, and then, sequences from the non-C6P species were assigned to these families using sequence similarity to the consensus. PE1, PE2, and PE3 were all present in all the major primate groups (Table 1). PE3 was not identified in tree shrews; however, the total number of ERVs found in tree shrews was relatively small.

The majority of previously described epsilon-like ERVs in the human genome were identified using our pipeline and are labeled in Table S3 in the supplemental material. We identified a total of 81 insertions in the human genome, consistent with the 70 ERVs clustering with nonmammalian ERVs identified by Tristem (14). Our PE2 group appears to encompass Oja et al.'s "upper" group of epsilon-like ERVs and our PE1 group their "lower" group. The HERV-AC018462 and HERV-L(b) groups of Katzourakis and Tristem (15) fell into our PE1 group and their HERV-R(c) group into our PE2 group. Three previously described sequences were not identified in our study: the type member of the HERV-AC096774 group described by Katzourakis and Tristem (15) and the chr1_684233 and chr17_47535521 groups described by Oja et al. (16). Five thousand base pairs from either side of human chr1_684233 (which corresponds to chr1_594413 in the most recent genome build) were analyzed using BLASTX against the nr database and by alignment with known epsilonretroviral *pol* genes, but nothing resembling a *pol* gene could be identified. Oja et al.'s chr17_47535521 was in the raw output from Exonerate but fell short of the quality threshold during our BLAST verification step, with the closest match to a known ERV being a 64-amino-acid segment sharing 54% identity with WDSV. HERV-AC096774 was not identified using Exonerate; however, as stated by Katzourakis and Tristem (15), this sequence is very degenerate. Both of these sequences are most similar to our PE1 group.

The consensus sequences of PE1, PE2, and PE3 were incorporated into a phylogeny of known epsilon and epsilon-like retroviruses (Fig. 2). Mammalian epsilon-like *pol* insertions in this phylogeny are the PE1, PE2, and PE3 consensus sequences; horse epsilon-like ERV fragments from our previous work (17); an example epsilon-like virus from Oja et al.; and one chimpanzee ERV lineage previously categorized only as "class I" (11). PE1, PE2, and PE3 form a moderately supported potential phylogenetic cluster with these known mammalian ERVs and the reptilian epsilon-like

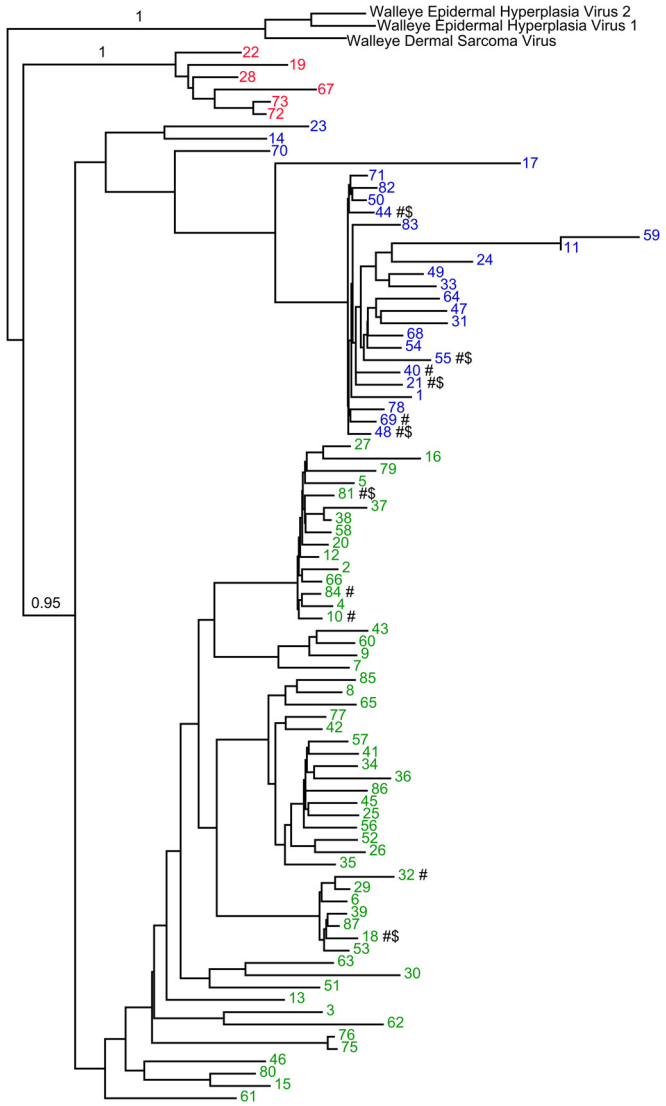

FIG 1 PhyML phylogenetic tree based on a 7,426-nucleotide multiple alignment of the consensus sequences for 87 epsilon-like *pol* gene fragments found in primates, showing the clustering of primate epsilonretroviral loci into three major phylogenetic groups. PE1 is shown in green, PE2 in blue, and PE3 in red. The numbers represent locus numbers, which were assigned arbitrarily. The 11 sequences with recognizable LTRs are labeled (#), and the six sequences with recognizable LTRs that are conserved in the Compara six-primate alignment species are also marked ($). Walleye dermal sarcoma virus and walleye epidermal sarcoma viruses 1 and 2 were used as an outgroup. Details of each locus are provided in Table S3 in the supplemental material. The branch support values are aLRT values calculated in PHYML. Branch support values are shown only for the three major clades.

ERVs. PE3 seems to be more closely related to the reptile epsilon-like ERVs than to the other mammalian insertions.

Potential LTRs were identified flanking 11 of the 87 PE loci; the remainder were too degenerate for reliable LTR sequences to be detected. Dating based on LTR similarity at these loci gave a mean integration date of 34.43 million years ago, with values ranging from 16.48 to 90.49 million years. LTRs clustered into four types, designated type 1 to type 4. PE2 loci had type 1 or type 4 LTRs, and PE1 loci had type 2 or type 3. No LTRs were identifiable at PE3 loci. These results are summarized in Table 2. Type 4 LTRs were identified only at loci with a median age greater than 34 million years.

Six loci had recognizable LTRs and were identified in all six C6P species. The C37M alignment was used to establish if these specific loci are found in all primates and if they are found outside the primates. The sequences were identifiable at the same positions in all apes, Old World monkeys, and New World monkeys in the alignment. However, at these positions, no ERV sequences were identifiable in prosimian primates or any nonprimates, including tree shrews. TBLASTN analysis did not identify any retroviral LTRs or *pol* or *env* gene fragments in these regions or the surrounding sequence in prosimians or nonprimates. Therefore, it appears that the insertion of epsilon-like ERVs at these specific sites occurred after the split between tarsiers and Old/New World primate ancestors (65 million years ago) but before the split between the ancestors of Old and New World monkeys (42 million years ago) (30). These dates are broadly consistent with the estimates above based on LTR divergence. Given that epsilon-like ERV fragments were absent at these loci in prosimians and tree shrews, the prosimian and tree shrew epsilon-like ERV fragments we identified appear to be the result of separate integration events at different integration sites than those in apes, Old World monkeys, and New World monkeys.

Using the human LTR sequences identified here as probes against the human genome, 777 further potential LTRs were identified. Fourteen pairs were identified between 8,000 and 15,000 bp apart, suggesting that the ERV sequence between the LTRs has not been deleted but is too degenerate to recognize. The remaining 749 are likely to be solo LTRs, the result of recombination between the two LTRs flanking an ERV sequence. This gives a ratio of 749 solo LTRs to 95 ERV sites that have not recombined in the human genome (including the 81 identified with Exonerate and the 14 pairs encompassing unrecognizable ERVs). In mice, the half-life for an ERV to recombine and form a solo LTR is estimated at 0.8 million years (13). The recombination rate of mice is around half that of humans per generation (31), but the mouse generation time is much shorter, with about 50 mouse generations for every human generation (32), giving an estimated ERV-to-solo LTR half-life of 20 million years in humans. At this rate, it would take approximately 60 million years to go from 844 ERV sites to 95 ERV sites and 749 solo LTRs, which is within our predicted range of insertion dates.

For the 11 loci with recognizable LTRs, the 5′ and 3′ limits of the LTR provided the full span of the ERV, meaning other features of the ERVs could be identified and characterized (see Table S4 in the supplemental material). WDSV is the type species for the epsilonretroviruses (33) and the only epsilonretrovirus with a reference sequence (GenBank accession number NC_001867) and so was used for comparisons. Apart from two endonuclease gene insertions, likely to be the result of later retrotransposition events by non-LTR retrotransposons, in humans at locus 84 and chimpanzees at locus 48, the longest ORF was a 296-amino-acid, or 888-bp, fragment at locus 32, starting within the 5′ LTR and ending within the region where *gag* would be expected. The protein encoded by this ORF shows no homology to any known retroviral protein (determined using BLASTP) and is considerably shorter than any major retroviral protein (WDSV has a 582-amino-acid Gag, a 1,171-amino-acid Pro-Pol, and a 1,225-amino-acid Env). Therefore, it is very unlikely that any of these ORFs could produce functional viral proteins. BLAST searches identified small *gag*

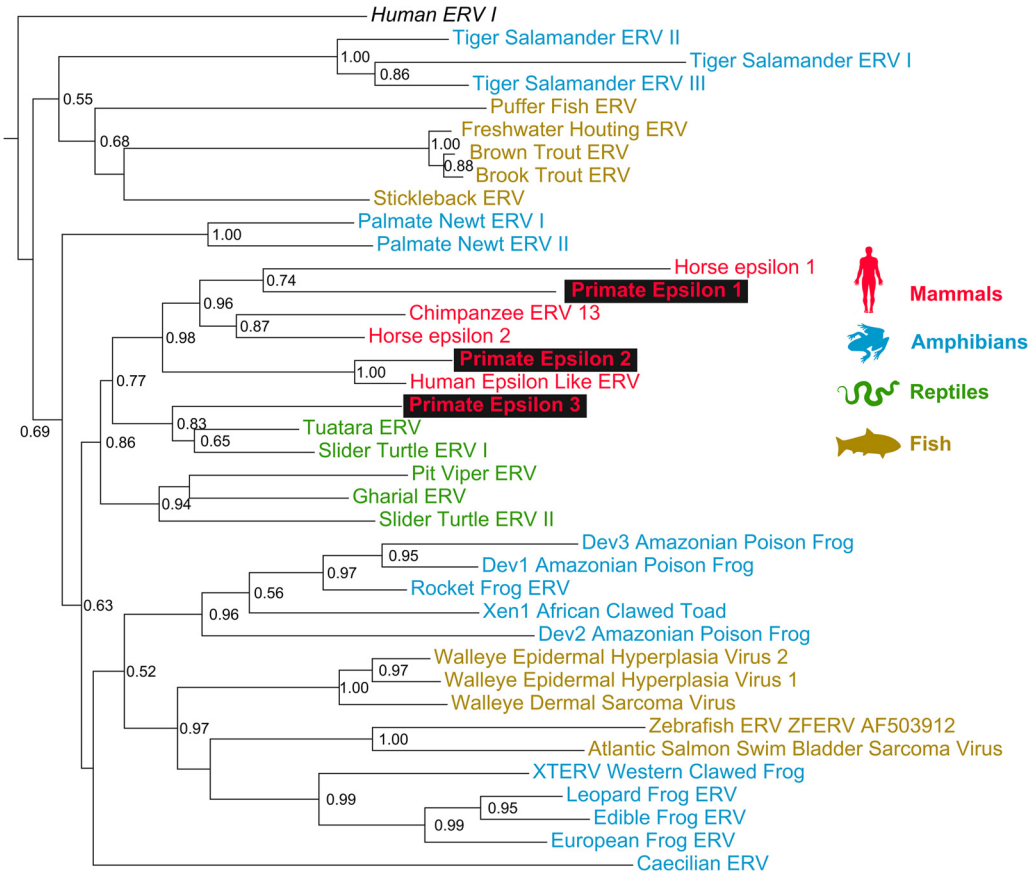

FIG 2 PhyML phylogenetic tree based on a 5,510-bp multiple alignment of the consensus sequences of three phylogenetic groups of primate epsilon-like *pol* gene fragments and known epsilon and epsilon-like retroviruses. Mammalian, amphibian, reptile, and fish epsilonretroviruses are color coded as shown. Newly identified sequences are highlighted. Full details of known epsilonretroviruses in the tree are provided in Table S1 in the supplemental material. Human ERV I is human endogenous retrovirus I, a gammaretrovirus. Branch support values are aLRT values calculated in PHYML; values below 0.5 are not shown.

fragments (less than 400 bp) with homology to WDSV between *pol* and the 5′ LTR of loci 18, 21, and 44 and *env* fragments sufficient to combine into a 1,330-bp consensus at loci 10 and 81 (see Table S4 in the supplemental material). These *gag* and *env* sequences, however, were too degenerate for meaningful phylogenetic analysis. No sequences with homology to the three WDSV accessory genes, *orf-A*, *orf-B*, and *orf-C*, were identified. A partial genome structure for the PE group was deduced from these results and is shown in Fig. 3. If accessory genes are excluded, the length of the PE genome and the positions of the *pol* gene and *env* fragment are consistent with WDSV, and the gaps between these regions are sufficient for the remainder of a functional epsilon-like ERV to have been present at some point in the evolutionary history of the host.

TABLE 2 Phylogenetic group, LTR type, proportion of sites at which LTRs are not identical to each other, and median age of each of the 11 epsilon-like ERV loci flanked by two recognizable LTRs

Locus	Group	LTR type	LTR divergence	Median age (million yr)
loc_18	PE1	3	0.078	17,319,367
loc_10	PE1	1	0.088	19,586,308
loc_81	PE1	2	0.100	22,173,007
loc_44	PE2	1	0.104	23,052,162
loc_69	PE2	1	0.107	23,772,610
loc_48	PE2	1	0.117	26,073,350
loc_84	PE1	2	0.139	30,939,030
loc_55	PE2	4	0.155	34,500,254
loc_21	PE2	4	0.176	39,089,995
loc_32	PE1	3	0.181	40,322,514
loc_40	PE2	4	0.185	41,044,747

A supertree representing the evolutionary relationships between sequences from each host at each locus was generated (data not shown). This tree is identical to the consensus host phylogeny, based on 17 host genes, available through the 10k trees project (34). If the loci are divided by family, PE1 and PE2 show this relationship with 100% support for all branches, while PE3 shows ambiguity in the relationship between human, gorilla, and chimpanzee, a relationship that is also sometimes ambiguous in evolutionary analyses of the host (35).

DISCUSSION

These results confirm the presence of a group of endogenous epsilon-like ERVs in the 14 primate species and in two species of tree shrew, the closest living relatives of the primates. The sequenced primates are from diverse geographical regions and represent all major primate taxonomic groups, so the identification of PE insertions in all of these hosts suggests that PE is found in all pri-

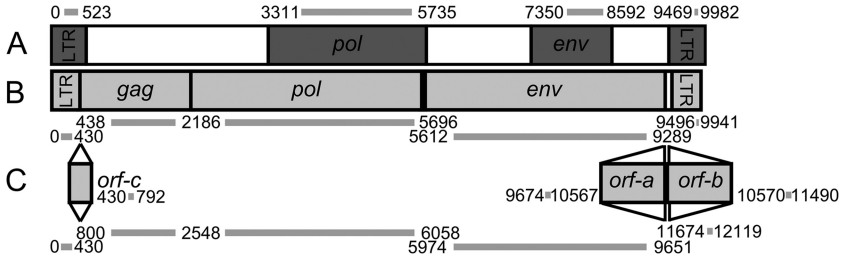

FIG 3 Comparison of identified regions of the PE genome (A) and the reference genome of WDSV (GenBank accession number NC_001867) with *orf-a*, *orf-b*, and *orf-c* excluded (B) and included (C) in the genome length and gene position calculations. The positions for PE are means across all loci with identifiable LTRs.

mates. By looking at individual PE loci in six primate species, we have confirmed that PE is likely to have entered the genome of a common ancestor of apes, Old World monkeys, and New World monkeys, while PE insertions in prosimian primates and tree shrews are likely to represent separate integration events in ancestors of these species. Many of these ERVs have not been identified previously. This is most likely due to the degree of degeneration of these sequences and the diversity of our input data set of known retroviruses, which is considerably more comprehensive that those that are generally used.

Mammals, reptiles, and birds make up a distinct group in vertebrate phylogeny known as amniotes (36). The phylogenetic tree shown in Fig. 2 suggests that all three families of PE insertions may form part of a group of epsilon-like ERVs unique to the amniotes, along with several previously characterized mammalian and reptilian epsilon-like ERVs. The known human epsilon-like ERVs (14–16) seem to represent members of our PE1 and PE2 families, and chimpanzee endogenous retrovirus lineage 13 (11) appears to be a member of PE1. PE3 clusters robustly with a group of reptilian ERVs. Our previously identified horse epsilon-like ERVs (17) fall within this provisional amniote ERV group.

The shared insertion sites in New and Old World monkeys provide a minimum age for circulation of the exogenous versions of these epsilon-like ERVs of 42 million years ago, and the absence of these shared insertion sites in prosimians provides a maximum age of 65 million years (30). Only one amphibian epsilon-like ERV currently has an estimated integration date, an insertion in *Xenopus tropicalis* dated at 41 million years old (1). This date is consistent with the relationships between amphibian retroviruses shown in Fig. 3. Therefore, amniote and amphibian retroviruses appear to have been circulating during approximately the same time period. The structure of the epsilon-like ERV phylogeny is best explained by a member of a group of circulating amphibian retroviruses 40 to 60 million years ago entering amphibian genomes multiple times and forming two distinct phylogenetic groups and a single strain crossing into amniotes and then diversifying to infect different amniote species.

All known endogenous fish epsilon-like ERVs are considerably more modern than this, with the oldest estimated at 3.79 million years old (8). This long gap between the ancient amphibian/amniote viruses and the modern fish viruses raises questions about the evolution of epsilon-like ERVs. The degeneration seen in amphibian and primate endogenous epsilon-like ERVs means they are unlikely to have had the potential to produce functional viral particles recently enough to be responsible for these integrations into fish. If exogenous members of the PE or horse epsilon-like ERV families had remained infectious throughout this period, there would most likely be more modern integrations detectable in our genome screens, though the possibility remains that other mammals have as yet unidentified epsilon-like ERVs, particularly as horses and primates are quite divergent host species. The remaining explanation is that exogenous epsilon-like retroviruses have been circulating throughout this period in another host or group of hosts and later crossed into fish. Significantly more screening would be needed to identify this host. The three distinct groups of fish/amphibian insertions in Fig. 2 suggest that cross-species transmissions into fish have occurred at least three times. As all three phylogenetic groups of fish epsilon and epsilon-like retroviruses are more similar to amphibian ERVs than amniote ERVs, amphibians could be a candidate. Screening of amphibians for ERVs to date has also been minimal. It is also possible that epsilon-like retroviruses have been circulating among fish throughout this time and that there are considerably more epsilon-like ERVs in fish that are yet to be discovered.

The exogenous fish epsilonretroviruses WDSV and WEHV encode three accessory proteins, Rv-cyclin (encoded by *orf-a*), Orf-B, and Orf-C (3) (Fig. 3). We did not identify the genes encoding these proteins at any PE locus or in the horse epsilon-like ERVs. Rv-cyclin and Orf-B are involved in tumor development, while Orf-C is involved in apoptosis, tumor regression, and tumor development (3). These genes are essential for WDSV proliferation and dissemination (3). However, the genes are not universal in fish retroviruses; for example, they are absent in zebrafish endogenous retrovirus (37) and Atlantic salmon swim bladder sarcoma virus (5), so they are likely to represent a later acquisition in the lineage leading to WDSV and the WEHVs.

We did not identify any epsilon-like ERVs in any of the 11 rodent species or two lagomorphs we screened. Rodents and lagomorphs are known to carry many endogenous and exogenous gammaretroviruses and appear to have high vulnerability to retroviral infection (12, 38, 39), so it is surprising that their closest sequenced relatives have endogenous epsilon-like ERVs but they do not. One possible explanation for this is that one of the diverse gammaretroviruses infecting rodents offered a protective effect against epsilon-like retroviruses. The use of existing endogenous retroviruses as restriction factors against exogenous pathogens is a known mechanism used by some hosts (40). Alternatively, the epsilon-like retroviral host range may depend on a combination of host restriction factors and viral accessory genes similar to that of simian immunodeficiency viruses (SIVs). For example, it has been demonstrated that macaques were unable to contract SIV from sooty mangabeys until the *vif* accessory gene of the virus adapted to counteract the macaque APOBEC3G restriction factor (41). A similar phenomenon may have prevented epsilon-like retrovi-

ruses from entering rodent genomes. Finally, it is possible that rodents and lagomorphs lack a receptor that epsilon-like retroviruses require and that is present in primates and horses. The two bird species screened here also lacked epsilon-like ERVs. Birds have an unusual complement of ERVs compared to mammals, which again might have acted as a barrier to epsilon-like retrovirus infection. It is also possible that there are epsilon-like ERVs in other bird species that were not analyzed here.

As fish still have active epsilonretroviruses and primate ancestors have clearly been susceptible to epsilon-like retroviruses in the past, it is not inconceivable that fish epsilonretroviruses could enter the human genome again. Further research is needed to establish if the lack of modern infections in mammals is due to a restriction factor or if mammals remain vulnerable to epsilon or epsilon-like retroviruses. Any restriction factor identified may be of interest to the aquaculture industry in terms of its potential in the control of WDSV and WEHV. The degree to which all the identified PE insertions have degenerated and the lack of functional *gag* and *env* genes make it very improbable that these loci could generate an active epsilon-like retrovirus, even by recombination.

In conclusion, epsilon-like ERVs appear to be common to all primate genomes and are likely to be widespread among mammals, although they are absent in rodents and lagomorphs. Amniote epsilon-like ERVs may form a distinct group within the epsilon and epsilon-like retrovirus phylogeny and are most likely to be the result of diversification of a cross-species transmission of viruses circulating 40 to 65 million years ago. Epsilon-like retroviruses appear to have continued to circulate since this time and have most recently invaded the genomes of fish, but further research is needed to establish whether these viruses originated in fish or other hosts.

ACKNOWLEDGMENTS

We thank John Brookfield, Elizabeth Hellen, Frank Wessely, and the University of Nottingham HPC for bioinformatics assistance and Ed Louis and Stephen Dunham for comments on the manuscript.

Funding for this project was provided by the University of Nottingham. The funding body had no role in the execution and analysis of the study.

REFERENCES

1. **Sinzelle L, Carradec Q, Paillard E, Bronchain OJ, Pollet N.** 2011. Characterization of a Xenopus tropicalis endogenous retrovirus with developmental and stress-dependent expression. J. Virol. **85:**2167–2179. http://dx.doi.org/10.1128/JVI.01979-10.
2. **VanDeValk AJ, Adams CM, Rudstam LG, Forney JL, Brooking TE, Gerken MA, Young BP, Hooper JT.** 2002. Comparison of angler and cormorant harvest of walleye and yellow perch in Oneida Lake, New York. Trans. Am. Fish Soc. **131:**27–39. http://dx.doi.org/10.1577/1548-8659(2002)131<0027:COAACH>2.0.CO;2.
3. **Rovnak J, Quackenbush SL.** 2010. Walleye dermal sarcoma virus: molecular biology and oncogenesis. Viruses **2:**1984–1999. http://dx.doi.org/10.3390/v2091984.
4. **Statistics and Information Service, Fisheries and Aquaculture Department.** 2012. FAO yearbook. Fishery and aquaculture statistics. Statistics and Information Service, Fisheries and Aquaculture Department, Food and Agriculture Organization of the United Nations, Rome, Italy. http://www.fao.org/3/a-i3740t/index.html.
5. **Paul TA, Quackenbush SL, Sutton C, Casey RN, Bowser PR, Casey JW.** 2006. Identification and characterization of an exogenous retrovirus from Atlantic salmon swim bladder sarcomas. J. Virol. **80:**2941–2948. http://dx.doi.org/10.1128/JVI.80.6.2941-2948.2006.
6. **Lepa A, Siwicki A.** 2011. Retroviruses of wild and cultured fish. Pol. J. Vet. Sci. **14:**703–709. http://dx.doi.org/10.2478/v10181-011-0106-8.
7. **Masahito P, Nishioka M, Ueda H, Kato Y, Yamazaki I, Nomura K, Sugano H, Kitagawa T.** 1995. Frequent development of pancreatic carcinomas in the Rana nigromaculata group. Cancer Res. **55:**3781–3784.
8. **Basta HA, Cleveland SB, Clinton RA, Dimitrov AG, McClure MA.** 2009. Evolution of teleost fish retroviruses: characterization of new retroviruses with cellular genes. J. Virol. **83:**10152–10162. http://dx.doi.org/10.1128/JVI.02546-08.
9. **Betancur-R R, Broughton RE, Wiley EO, Carpenter K, Lopez JA, Li C, Holcroft NI, Arcila D, Sanciangco M, Cureton JC, II, Zhang F, Buser T, Campbell MA, Ballesteros JA, Roa-Varon A, Willis S, Borden WC, Rowley T, Reneau PC, Hough DJ, Lu G, Grande T, Arratia G, Orti G.** 2013. The tree of life and a new classification of bony fishes. PLoS Curr. **5:**ecurrents.tol.53ba26640df0ccaee75bb165c8c26288. http://dx.doi.org/10.1371/currents.tol.53ba26640df0ccaee75bb165c8c26288.
10. **Herniou E, Martin J, Miller K, Cook J, Wilkinson M, Tristem M.** 1998. Retroviral diversity and distribution in vertebrates. J. Virol. **72:**5955–5966.
11. **Polavarapu N, Bowen NJ, McDonald JF.** 2006. Identification, characterization and comparative genomics of chimpanzee endogenous retroviruses. Genome Biol. **7:**R51. http://dx.doi.org/10.1186/gb-2006-7-6-r51.
12. **Stocking C, Kozak C.** 2008. Endogenous retroviruses. Cell. Mol. Life Sci. **65:**3383–3398. http://dx.doi.org/10.1007/s00018-008-8497-0.
13. **Nellaker C, Keane T, Yalcin B, Wong K, Agam A, Belgard TG, Flint J, Adams D, Frankel W, Ponting C.** 2012. The genomic landscape shaped by selection on transposable elements across 18 mouse strains. Genome Biol. **13:**R45. http://dx.doi.org/10.1186/gb-2012-13-6-r45.
14. **Tristem M.** 2000. Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the Human Genome Mapping Project database. J. Virol. **74:**3715–3730. http://dx.doi.org/10.1128/JVI.74.8.3715-3730.2000.
15. **Katzourakis A, Tristem M.** 2005. Phylogeny of human endogenous and exogenous retroviruses, p 186–203. *In* Sverdlov ED (ed), Retroviruses and primate genome evolution. Landes Bioscience, Austin, TX.
16. **Oja M, Sperber GO, Blomberg J, Kaski S.** 2005. Self-organizing map-based discovery and visualization of human endogenous retroviral sequence groups. Int. J. Neural Syst. **15:**163–179. http://dx.doi.org/10.1142/S0129065705000177.
17. **Brown K, Moreton J, Malla S, Aboobaker AA, Emes RD, Tarlinton RE.** 2012. Characterisation of retroviruses in the horse genome and their transcriptional activity via transcriptome sequencing. Virology **433:**55–63. http://dx.doi.org/10.1016/j.virol.2012.07.010.
18. **Slater G, Birney E.** 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics **6:**31. http://dx.doi.org/10.1186/1471-2105-6-31.
19. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local alignment search tool. J. Mol. Biol. **215:**403–410.
20. **Coffin JM, Hughes SH, Varmus HE.** 1997. Retroviruses. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
21. **Rice P, Longden I, Bleasby A.** 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. **16:**276–277. http://dx.doi.org/10.1016/S0168-9525(00)02024-2.
22. **Smith TF, Waterman MS.** 1981. Identification of common molecular subsequences. J. Mol. Biol. **147:**195–197. http://dx.doi.org/10.1016/0022-2836(81)90087-5.
23. **Katoh K, Misawa K, Kuma K, Miyata T.** 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. **30:**3059–3066. http://dx.doi.org/10.1093/nar/gkf436.
24. **Guindon S, Gascuel O.** 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. **52:**696–704. http://dx.doi.org/10.1080/10635150390235520.
25. **Creevey CJ, McInerney JO.** 2005. Clann: investigating phylogenetic information through supertree analyses. Bioinformatics **21:**390–392. http://dx.doi.org/10.1093/bioinformatics/bti020.
26. **Bannert N, Fiebig U, Hohn O.** 2010. Retroviral particles, proteins and genomes, p 71–106. *In* Kurth R, Bannert N (ed), Retroviruses. Molecular biology, genomics and pathogenesis. Caister Academic Press, Norfolk, United Kingdom.
27. **Jurka J, Klonowski P, Dagman V, Pelton P.** 1996. CENSOR—a program for identification and elimination of repetitive elements from DNA sequences. Comput. Chem. **20:**119–121. http://dx.doi.org/10.1016/S0097-8485(96)80013-1.
28. **Gifford RJ, Katzourakis A, Tristem M, Pybus OG, Winters M, Shafer RW.** 2008. A transitional endogenous lentivirus from the genome of a basal pri-

mate and implications for lentivirus evolution. Proc. Natl. Acad. Sci. U. S. A. **105:**20362–20367. http://dx.doi.org/10.1073/pnas.0807873105.
29. **Kent WJ.** 2002. BLAT—the BLAST-Like Alignment Tool. Genome Res. **12:**656–664. http://dx.doi.org/10.1101/gr.229202.
30. **Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, Moreira MAM, Kessing B, Pontius J, Roelke M, Rumpler Y, Schneider MPC, Silva A, O'Brien SJ, Pecon-Slattery J.** 2011. A molecular phylogeny of living primates. PLoS Genet. **7:**e1001342. http://dx.doi.org/10.1371/journal.pgen.1001342.
31. **Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, Roskin KM, Chen C-F, Thomas MA, Haussler D, Jacob HJ.** 2004. Comparative recombination rates in the rat, mouse, and human genomes. Genome Res. **14:**528–538. http://dx.doi.org/10.1101/gr.1970304.
32. **Keightley PD, Eyre-Walker A.** 2000. Deleterious mutations and the evolution of sex. Science **290:**331–333. http://dx.doi.org/10.1126/science.290.5490.331.
33. **International Committee on Taxonomy of Viruses.** 2013. Virus taxonomy: 2013 release. http://www.ictvonline.org/virusTaxonomy.asp.
34. **Arnold C, Matthews LJ, Nunn CL.** 2010. The 10kTrees website: a new online resource for primate phylogeny. Evol. Anthropol. **19:**114–118. http://dx.doi.org/10.1002/evan.20251.
35. **Chen F-C, Li W-H.** 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am. J. Hum. Genet. **68:**444–456. http://dx.doi.org/10.1086/318206.
36. **Meredith RW, Janečka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simão TLL, Stadler T, Rabosky DL, Honeycutt RL, Flynn JJ, Ingram CM, Steiner C, Williams TL, Robinson TJ, Burk-Herrick A, Westerman M, Ayoub NA, Springer MS, Murphy WJ.** 2011. Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. Science **334:**521–524. http://dx.doi.org/10.1126/science.1211028.
37. **Shen C-H, Steiner LA.** 2004. Genome structure and thymic expression of an endogenous retrovirus in zebrafish. J. Virol. **78:**899–911. http://dx.doi.org/10.1128/JVI.78.2.899-911.2004.
38. **Baillie GJ, van de Lagemaat LN, Baust C, Mager DL.** 2004. Multiple groups of endogenous betaretroviruses in mice, rats, and other mammals. J. Virol. **78:**5784–5798. http://dx.doi.org/10.1128/JVI.78.11.5784-5798.2004.
39. **McCarthy E, McDonald J.** 2004. Long terminal repeat retrotransposons of Mus musculus. Genome Biol. **5:**R14. http://dx.doi.org/10.1186/gb-2004-5-3-r14.
40. **Arnaud F, Caporale M, Varela M, Biek R, Chessa B, Alberti A, Golder M, Mura M, Zhang Y-P, Yu L, Pereira F, DeMartini JC, Leymaster K, Spencer TE, Palmarini M.** 2007. A paradigm for virus-host coevolution: sequential counter-adaptations between endogenous and exogenous retroviruses. PLoS Pathog. **3:**e170. http://dx.doi.org/10.1371/journal.ppat.0030170.
41. **Krupp A, McCarthy KR, Ooms M, Letko M, Morgan JS, Simon V, Johnson WE.** 2013. APOBEC3G polymorphism as a selective barrier to cross-species transmission and emergence of pathogenic SIV and AIDS in a primate host. PLoS Pathog. **9:**e1003641. http://dx.doi.org/10.1371/journal.ppat.1003641.