Endogenous hepadnaviruses, bornaviruses and circoviruses in snakes

C. Gilbert1, J. M. Meik2, D. Dashevsky3, D. C. Card4, T. A. Castoe4 and S. Schaack3,5,6

1Université de Poitiers, UMR CNRS 7267, Ecologie et Biologie des Interactions, Equipe Ecologie Evolution Symbiose, Poitiers, France
2Department of Biological Sciences, Tarleton State University, Stephenville, TX, USA
3Department of Biology, Reed College, Portland, OR, USA
4Department of Biology, The University of Texas at Arlington, Arlington, TX, USA
5Biosciences Eastern and Central Africa, International Livestock Research Institute, Nairobi, Kenya
6Centre for Bioinformatics and Biotechnology, University of Nairobi, Nairobi, Kenya

We report the discovery of endogenous viral elements (EVEs) from Hepadnaviridae, Bornaviridae and Circoviridae in the speckled rattlesnake, *Crotalus mitchellii*, the first viperid snake for which a draft whole genome sequence assembly is available. Analysis of the draft assembly reveals genome fragments from the three virus families were inserted into the genome of this snake over the past 50 Myr. Cross-species PCR screening of orthologous loci and computational scanning of the python and king cobra genomes reveals that circoviruses integrated most recently (within the last approx. 10 Myr), whereas bornaviruses and hepadnaviruses integrated at least approximately 13 and approximately 50 Ma, respectively. This is, to our knowledge, the first report of circo-, borna- and hepadnaviruses in snakes and the first characterization of non-retroviral EVEs in non-avian reptiles. Our study provides a window into the historical dynamics of viruses in these host lineages and shows that their evolution involved multiple host-switches between mammals and reptiles.

1. Introduction

Endogenous viral elements (EVEs) are entire or fragmented viral genomes that have been integrated into the genome of their hosts and are therefore vertically inherited in a stable manner [1,2]. Although only retroviruses are known to require integration of their genomes into their host genome to complete their replication cycle, EVEs from all types of viruses have now been identified in eukaryotic genomes [3]. These elements result from endogenization events that occurred in the past, the genetic 'fossils' of which can be detected in whole genome sequence data millions of years later. Endogenization of viruses is not rare; in fact, it appears to be a recurrent and on-going process in the evolution of eukaryotic genomes [4,5].

The study of EVEs fossilized in their host genomes has opened new avenues to better understand the biology of viruses and the evolutionary history of host–virus interactions [6]. For example, EVE discovery can expand the past and current known host range of viral families [7–9], contribute to the characterization of the viral flora infecting a specific organism or lineage [10,11] and help to identify reservoir species and/or reconstruct complex transmission routes of zoonotic viruses [12–14]. Furthermore, endogenization events can be dated and used to calibrate long-term evolutionary patterns of modern virus families [15–18]. Finally, studies in palaeovirology are increasingly revealing how viruses have contributed to the gene repertoires of their hosts. Indeed, although relatively few EVE-derived host genes have been characterized in detail so far [19–22], numerous EVEs are transcribed and/or have evolved under purifying selection after endogenization [10,11,23–25], suggesting they may have been domesticated and now fulfill cellular functions in the host.

Non-retroviral EVEs have been identified in the genomes of species from several branches of the eukaryotic tree, including metazoans, fungi, plants,
chromalveolates and excavates [1,3,11]. However, most branches have not been examined, either because of taxonomic biases in sequencing efforts or owing to a lack of screening. As a result, most vertebrate EVEs have been identified in mammalian genomes (in approx. 20 species), with only a few cases known in other host species; one EVE found in the genomes of two species of birds (Hepadnaviridae [8,18,26]), mononegavirus-like sequences found in two species of teleostean fishes and in the lamprey [12,24], and a circovirus- and nanovirus-like sequence reported in the frog *Xenopus tropicalis* [27]. No non-retroviral EVEs have been comprehensively characterized in non-avian reptiles (but see [28]), which is probably owing to there being only one whole genome sequence available until very recently (the Carolina anole, *Anolis carolinensis* [29]).

Among non-avian reptiles, snakes represent a large and diverse radiation of over 3000 species, for which genomic data are just starting to become available [30,31]. Among snakes, the rattlesnakes (which include two genera, *Crotalus* and *Sistrurus*) represent a clade of approximately 40 species of highly venomous snakes in the family Viperidae endemic to the Western Hemisphere. Owing partly to their unique morphology, ecology and physiology, as well as their cultural and medical importance, rattlesnakes have become one of the most extensively studied groups of non-avian reptiles [32]. Here, we search whole genome sequence data obtained from a speckled rattlesnake (*Crotalus mitchellii*) for endogenous viruses and identified multiple EVEs belonging to the Hepadnaviridae, Bornaviridae and Circoviridae families in this and several other species of pitvipers. In addition, we use the genomes of two additional snake species (Burmese python and king cobra [30,31]) to further estimate the timing of viral endogenization. Our results show that the three families of viruses that were previously not known to occur in the genomes of squamate reptiles have recurrently infiltrated the germline of snakes over the past 50 Myr.

### 2. Material and methods

#### (a) Whole genome sequencing and assembly of *Crotalus mitchellii*

The speckled rattlesnake (*C. mitchellii*) species complex includes species endemic to arid regions of western North America and Mexico [33,34]. The single specimen used for genome sequencing is shown on figure 1. It is a subadult female individual of *C. m. pyrrhus* (JMM 685; UTA R-60292) collected on 11 June 2012 from Palm Canyon, Kofa Mountains, La Paz County, AZ, USA (33.360 N, 114.106 W; 650 m).

Genomic DNA was extracted from snap-frozen liver tissue, stored at −80°C, using standard lysis and proteinase digestion followed by phenol–chloroform–isoamyl alcohol methods. A genomic shotgun sequencing library was prepared by shearing genomic DNA using a Covaris ultrasonicator, and further preparation using an Illumina Truseq DNAseq library kit. This shotgun library was sequenced on a single lane of an Illumina HiSeq 2000 using 100 bp paired-end reads. Sequencing produced 408 million total paired reads, which were quality trimmed using the Trim Sequences tool in CLC Genomics Workbench (CLC Bio/Qiagen) with a quality score limit of 0.05 and a maximum of two ambiguities. Trimmed reads were assembled de novo using the De Novo Assembly tool in CLC Genomics Workbench with automatic word and bubble sizes, a minimum contig length of 200 bp, paired-read scaffolding, and reads mapped back to contigs for gap-filling (mixmatch cost = 2, insertion cost = 3, deletion cost = 3, length fraction = 0.05, similarity fraction = 0.8). The total length of the resulting 478,598 contigs was 1.14 Gb, providing approximately 40× coverage of the genome, with an N50 scaffold size of 5.3 kb (GenBank accession number: JPMF0000000).

#### (b) Screening for endogenous viral elements in the genome of *Crotalus mitchellii*

To screen for the presence of EVEs in the rattlesnake genome, we used all viral genomes available in GenBank excluding retroviruses (*n* = 2048 as of November 2013) as queries to perform a TBLASTX search of *C. mitchellii* contigs with a 0.00001 e-value cut-off. Our study only focused on non-retroviral EVEs because endogenous retrovirus (ERV) EVEs have been extensively characterized in the genomes of many vertebrates [39] and various studies have shown that snake genomes, like all vertebrate genomes, harbour dozens or hundreds of ERVs [30,40–42]. Our TBLASTX library did include the recently characterized snake arenaviruses, presumed to be the cause of inclusion body disease in various snake species [43], but no hits were recovered matching the genome of these viruses.

All hits longer than 50 amino acids (aa) resulting from the TBLASTX search were then used as queries to perform a reciprocal BLASTP search of the GenBank non-redundant protein database (accessed in November 2013). Based on preliminary TBLASTN searches, we found that snake genomes contain numerous ERVs, as found in all other vertebrates (data not shown). In addition to searching for non-retroviral endogenous viruses in *C. mitchellii*, we searched with the same library in two additional snake genomes that are publicly available: the king cobra, *Ophiophagus hannah* [31], and the Burmese python, *Python molurus bivittatus* [30].

#### (c) Estimating the timing of endogenization events

In order to estimate the timing of the endogenization events producing the various EVEs that we detected, we followed an approach previously used in palaeovirology studies [18,26,44]. We used EVE-specific PCR primers to screen for the presence or absence of orthologous EVEs in vipersid snakes with various degrees of relatedness to *C. mitchellii*. When possible, primer pairs included a forward primer anchored in the upstream region flanking a given EVE and a reverse primer anchored in the downstream region flanking the EVE. This type of primer pair allowed amplification of orthologous loci, irrespective of whether the EVE was present or absent. Other primer pairs included one primer anchored within the EVE and another primer anchored either in the upstream or downstream region flanking the EVE. With these types of primer pairs, orthologous loci could only be amplified if the EVE was present. The primers are listed in the electronic supplementary material, table S1, and their sequences are provided in dataset S2.

For use in PCR reactions, snake DNA was extracted using the DNaseeasy blood and tissue kit (Qiagen) according to the manufacturer’s protocol. PCR reactions were conducted using the following thermal cycling profile: initial denaturation at 94°C for 5 min, followed by 30 cycles of denaturation at 94°C for 30 s, annealing at 54°C for 30 s and elongation at 72°C for 1 min, ending with a 10 min elongation step at 72°C. Fragments from the PCR were visualized on a 1–2% agarose gel. Purified PCR products were directly sequenced using ABI BigDye sequencing mix (1.4 μl template PCR product, 0.4 μl BigDye, 2 μl manufacturer supplied buffer, 0.3 μl primer and 6 μl H2O). Sequencing reactions were ethanol precipitated and run on an ABI 3730 sequencer. The list of snake species screened using this approach, together with corresponding museum voucher numbers for these specimens, are provided in the electronic supplementary material,
In order to assess the phylogenetic relationships of the various EVEs detected in each species, we conducted a series of selection analyses on orthologous EVEs detected in each species (table 1; electronic supplementary material, datasets S4–S12). These analyses are based on alignments of orthologous EVEs only and do not include exogenous viruses. The codon-based Z-test in MEGA 6 [51] was used with the Nei–Gojobori method and Jukes–Cantor correction and 500 bootstrap (BP) replicates. We also calculated long-term substitution rates of all three families following the approach used in [26] in order to compare them to short-term rates that have been calculated based on analyses of exogenous viruses [52,53]. As illustrated in the details in Box 2 of [3], fig. 4 of [5] and fig. 4 of [26], we first calculated the distance between each EVE and their closest extant virus relative (Kimura-2P distance). This distance corresponds to the sum of the mutations accumulated in EVEs at the host rate since endogenization and the mutations accumulated at the viral rate on the exogenous viral branch. In order to estimate the long-term viral rate, we subtracted the mutations accumulated at the host rate since endogenization, which we were able to estimate by calculating the distance between orthologous EVEs. We then divided the final distance, corresponding only to mutations accumulated at the viral rate, by the age of each EVE.

3. Results

(a) Multiple endogenous viral elements in the speckled rattlesnake

Our search for EVEs in the genome of the speckled rattlesnake (C. mitchelli) revealed that there are nine fragments with significant similarity to sequences from three extant virus families, each characterized by a different genome type: Circoviridae (ssDNA), Bornaviridae (-ssRNA) and Hepadnaviridae (dsDNA-RT). We named these EVEs eSCV1–5 for endogenous snake circovirus fragments 1 through to 5 and eSHBV1–2 for endogenous hepadnaviruses and endogenous bornaviruses.
Table 1. Characteristics of the snake EVEs identified in this study. (+, positive PCR yielding a band of the size expected for presence of a specific EVE locus; —, positive PCR yielding a band of the size expected for absence of a specific EVE locus; neg. PCR, negative PCR, either because the primer pair does not allow amplification of empty site or because orthologous loci are not sufficiently conserved; (seq), the PCR product was sequenced (see the electronic supplementary material, datasets S4–S12). See also the electronic supplementary material, table S1, for more details.)

<table>
<thead>
<tr>
<th>name of EVE</th>
<th>Grooviridae (ssDNA)</th>
<th>Bornaviridae (-ssRNA)</th>
<th>Hepadnaviridae (dsDNA-RT)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>eSCV1</td>
<td>eSCV2</td>
<td>eSCV3</td>
</tr>
<tr>
<td>Crotalus mitchelli</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>contig</td>
<td>224 958</td>
<td>224 959</td>
<td>9 220</td>
</tr>
<tr>
<td>length of EVE (aa)</td>
<td>167</td>
<td>199</td>
<td>241</td>
</tr>
<tr>
<td>best BLASTP hit</td>
<td>porcine circovirus capsid</td>
<td>porcine circovirus capsid</td>
<td>bat circovirus Rep</td>
</tr>
<tr>
<td>% aa similarity</td>
<td>58</td>
<td>53</td>
<td>82</td>
</tr>
<tr>
<td>K2P nucleotide distance</td>
<td>0.64</td>
<td>0.67</td>
<td>0.49</td>
</tr>
<tr>
<td>frameshift/stops</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
</tr>
</tbody>
</table>

*Sequences used to calculate K2P distances between orthologous snake EVEs.*

*bHost substitution rates were calculated by dividing the K2P distance between orthologous snake EVEs by the inferred age of the EVEs.*

*cSee methods for a detailed explanation of how long-term viral rates were inferred.*
endogenous snake hepadnavirus fragments 1 and 2 (table 1). For bornavirus EVEs, we follow Horie et al. [28] and use EBLN1–2 for endogenous bornavirus-like N nucleoprotein fragments 1 and 2. The nucleotide sequences of all snake EVEs identified in this study are provided in the electronic supplementary material, dataset S3, and a schematic alignment of these sequences with a representative species of each family is illustrated in the electronic supplementary material, figure S1.

The eSCVs identified are 130–241 aa in length; two of them correspond to the viral capsid protein (53–58% aa similarity, K2P nucleotide distance = 0.64–0.67, over 85% of the length of the porcine circovirus 1 capsid), and three are homologous to the viral replication initiator (Rep) protein (80–82% aa similarity, K2P distance = 0.45–0.49 over 75% of the length of a bat circovirus Rep). Both circovirus genes have been found endogenized in various metazoan genomes [11,15]. Overall, eSCVs cover about 75% of the length of a typical circovirus genome (e.g. porcine circovirus 1 = 1768 bp).

Snake EBLNs are 345 and 368 aa long and 61–63% similar (aa level) to the avian bornavirus nucleoprotein (N) gene over its entire length (K2P distance = 0.74–0.75). In addition to a nucleoprotein, the genome of bornaviruses encodes five other proteins (a phosphoprotein, a matrix protein, a glycoprotein, an RNA-dependent RNA polymerase and an accessory protein). Several endogenous fragments of the matrix, RNA-polymerase and glycoprotein genes have been discovered in the genomes of a number of other vertebrate species [1,12,28,44]. Most endogenous bornaviruses reported to date correspond to the nucleoprotein, which may be owing to the existence of a 3′–5′ transcription gradient of bornavirus genomes, resulting in a higher abundance of nucleoprotein mRNA [28].

Previous studies on endogenous hepadnaviruses characterized cases in two bird species, the zebrafinch Taeniopygia guttata and the budgerigar Melopsittacus undulatus, each of which contained multiple partial copies (n > 10), as well as a full-length or nearly full-length copy of the hepadnaviral genome [8,18,26,54]. In snakes, the eSHBVs are 316–370 aa long, with 48–67% aa similarity (K2P distance = 0.63–0.83) to the avianhepadnavirus polymerase and cover about 37% of the duck hepadnavirus genome, which is 3021 bp long. EVE scans in the cobra and python genomes additionally yielded a 53 aa fragment in the cobra showing 60% similarity to the core protein of the parrot hepatitis B virus, in addition to eSHBV1. In the python, we also discovered a 58 aa fragment showing 72% similarity to the Rep protein of the bat circovirus and a 326 aa fragment showing 67% similarity to the nucleoprotein of the orangutan bornavirus (electronic supplementary material, table S1).

Importantly, all snake EVEs discovered in this study are flanked by different sequences, suggesting that they all derive from independent endogenization events. Furthermore, although it has been shown that endogenization could be catalysed by host retrotransposons [44], we did not detect any molecular signature (such as target-site duplications) indicating involvement of such mobile elements in snake EVEs endogenization.

(b) Estimates of the timing of endogenization events

The results of PCR and bioinformatics screens for orthologous EVEs are depicted in figure 1 and table 1 (see also the electronic supplementary material, table S1). In three instances (eSCV2, eSCV4 and eSCV5), we were able to amplify both the orthologous EVE-containing loci in species closely related to C. mitchelli and the empty EVE loci (i.e. EVEs absent) in more distantly related species, yielding both an upper and lower time limit for the endogenization events. For these three loci, we can therefore confidently conclude that endogenization took place in the ancestor of the (C. mitchelli + Crotalus tigris + Crotalus atrox) clade, between 6 and 8 Ma (figure 1). For the other loci, we were able to amplify only EVE-containing loci. This is because either our PCR design did not allow us to amplify empty sites (see Material and methods) or, irrespective of whether the orthologous loci contained a given EVE or not, they were not sufficiently conserved to allow amplification. In these instances, we could still estimate the lower bound date of endogenization, but only speculate on the upper limit based on the absence of amplification. Thus, our screening revealed that cmEBLN1, eSHBV2, eSCV1 and eSCV3 are at least 8 Myr old and that cmEBLN2 is at least 13 Myr old. Furthermore, we also found that eSHBV1 was present at the orthologous locus in the king cobra genome, showing that this EVE is at least 50 Myr old (electronic supplementary material, dataset S11).

To assess whether the rates of evolution of EVEs are consistent with genome-wide substitution rates in snakes, we calculated the distance between orthologous EVEs and divided this distance by the minimum age of each EVE to obtain a substitution rate (substitutions per site per year). This yielded rate estimates ranging from 4.1 × 10^-10 to 2.5 × 10^-9 subst. site^-1 year^-1, with an average of 1.16 × 10^-9 subst. site^-1 year^-1 across EVEs (s.d. = 7.5 × 10^-10, table 1). These rates are close to those recently estimated for fourfold degenerate third codon positions based on 44 genes in viperids (1.4 × 10^-9 subst. site^-1 year^-1) and elapids (1.6 × 10^-9 subst. site^-1 year^-1; [30]), suggesting that our inferences about the age of snake EVEs are consistent with what is currently known about the evolutionary dynamics in snake genomes, more broadly. These data also suggest that these elements are evolving at approximately the background neutral rate of the host genomes and are probably non-functional. Overall, the snake EVEs uncovered in this study provide evidence for multiple endogenization events in these lineages over at least the last 50 Myr.

(c) Phylogenetic analyses of snake endogenous viral elements

The phylogenetic relationships of all snake EVEs, together with their most closely related known exogenous and endogenous viruses, are shown in figures 2–4 (see also the electronic supplementary material, figure S2, for the circovirus capsid gene tree). In all trees, snake EVEs are characterized by relatively long branches and are distantly related to well-characterized exogenous viruses. Considering that the phylogenetic distance between snake EVEs and known exogenous viruses (e.g. between eSHBVs and other hepadnaviruses) is as large or larger than the distance separating extant viruses known to belong to distinct lineages infecting widely divergent hosts (e.g. mammalian hepadnaviruses versus avian hepadnaviruses), we contend that snake EVEs belong to new lineages that may be considered new viral species or genera. Interestingly, eSCV3, eSCV4 and eSCV5 are separated by very short branches in the tree, figure 2, reflecting minimal divergence between these sequences, potentially indicating that these
EVEs result from repeated endogenization of the same virus. Together, these three EVEs form a fairly well-supported cluster (maximum-likelihood BP = 81; Bayesian posterior probability (PP) > 0.9) with an exogenous circovirus sequenced as part of a recent viral metagenomic study conducted on Rhinolophus ferrumequinum bats (figure 2 [55]). This grouping, also recovered in the capsid phylogeny (electronic supplementary material, figure S2), suggests circoviruses may have jumped between mammals and snakes during their evolutionary history. In the phylogeny of bornaviruses (figure 3), it is noteworthy that the python EBLN3 is more closely related to a subset of mammalian EBLNs (BP = 80; PP > 0.9) than it is to the rattlesnake EBLNs. Interestingly, none of the newly described snake EBLNs are closely related to the bornavirus transcript of the gaboon viper reported in [44] which seems to be more closely related to a clade including exogenous bird and human bornaviruses, as well as endogenous squirrel bornavirus (although this group is not well supported in our analysis). Overall, these results suggest repeated endogenization events of widely divergent bornaviruses in snakes and, potentially, multiple transfers between mammals and reptiles.

Interestingly, the cobra and rattlesnake eSHBV1 group tightly together (BP = 100; PP > 0.9), as expected given that the two sequences are orthologous. It is noteworthy that in this grouping, the length of the terminal branches of the two sequences represents the amount of evolution that took place after endogenization at the host genome’s substitution rate. In the hepadnavirus tree (figure 4, rooted using the reverse transcriptase of three different retrotransposons), snake eHBVs appear paraphyletic, with eSHBV2 falling sister to a clade including eSHBV1 and mammalian exogenous hepadnaviruses and both avian exogenous and endogenous hepadnaviruses. Although the (eSHBV1 + mammalian hepadnaviruses + avian hepadnaviruses) clade receives moderate BP support (BP = 80) and the position of eSHBV1 within this clade is unresolved (BP < 70), the likely paraphyly of snake EVEs based on this tree indicates that the hepadnavirus phylogeny is not fully congruent with that of their hosts. This presents further evidence that, like bornaviruses and circoviruses, hepadnaviruses have probably jumped between snakes and other vertebrates during their evolution.

Our tests for selection on endogenized viral sequences did not reveal any evidence of purifying or positive selection (p > 0.5 in all tests), suggesting all EVEs have been evolving neutrally in snake genomes after endogenization. We infer long-term viral substitution rates ranging from 1.3 × 10⁻⁸ subst. site⁻¹ year⁻¹ for eSHBV1 to 1.1 × 10⁻⁷ subst. site⁻¹ year⁻¹ for eSCV2, with an average of 6.7 × 10⁻⁸ subst. site⁻¹ year⁻¹ (s.d. = 2.6 × 10⁻⁸) (table 1).

4. Discussion

A large number of viruses have been characterized in reptiles [58–60], including at least 11 families reported from snakes. Hepadnaviruses, bornaviruses and circoviruses, however, have not been reported in this group, with the exception of one transcript of a truncated bornavirus nucleoprotein gene (94 aa) found in a Gaboon viper transcriptome [44] and a TBLASTN hit corresponding to our python EBLN3 reported in a supplementary table of Horie et al. [28]. In fact, bornaviruses and hepadnaviruses have never been observed in non-avian reptiles, and only one member of the Circoviridae is known from a non-avian reptile (green sea turtle [61]). The multiple endogenization
events we report across taxa, however, show that snakes have been repeatedly exposed to divergent members of these three viral families during the last 50 Myr (table 1). While these data do not imply that snakes are currently interacting with these viruses, they do indicate that they have interacted in the past. Our results broadly illustrate the potential of palaeovirology for shedding light on past host–virus interactions, for developing and testing hypotheses regarding host-switching, and as a guide for targeting plausible host taxa in future searches of circulating viruses. In addition to the identification of previously unreported endogenized viruses in snakes, our phylogenetic analysis reveals a clear lack of congruence between host and virus phylogenies, suggesting the three virus families have probably jumped between mammals and reptiles one or multiple times in the past (figures 2–4).

In the case of circoviruses, the absence of overlap in the geographical distribution of rattlesnakes (restricted to the Americas) and *R. ferrumequinum* bats (restricted to Europe and Asia) precludes any straightforward hypothesis regarding the mechanism of transfer between these hosts. Similarly, although eSCV3 and eSCV4 have been present in rattlesnake genomes for more than 6 Myr (figure 1), it is not possible to determine whether this transfer between hosts took place before or after endogenization. Our phylogenetic analyses indicate that, like mammals, snakes have also been exposed to widely divergent lineages of bornaviruses during their evolutionary history. In addition to the fact that human exogenous bornaviruses and squirrel endogenous bornaviruses are closer to bird exogenous bornaviruses than to other mammalian EBLNs, the polyphyly of snake EBLNs suggests that, since their origin > 93 Ma [1], bornaviruses have repeatedly jumped between mammals and reptiles (including birds).

It has been shown that the evolution of exogenous hepatnaviruses is characterized by frequent recent host-switching [62,63]. Our rooted tree indicates that snake hepatnaviruses are probably paraphyletic and that bird hepatnaviruses (endogenous + exogenous) are closer to mammalian than to snake hepatnaviruses, which is incongruent with the host phylogeny (in which birds are closer to snakes than to mammals [64]). This suggests that these viruses probably jumped between vertebrate hosts as well. Interestingly, eSHBVs appear to fall outside of the known diversity of bird and mammalian hepatnaviruses (although the position of eSHBV1 has poor BP support) and indicate that these viruses might be characterized as hepatna-like viruses belonging to a new family. However, a unique characteristic of Hepadnaviridae is that their genomes are extremely compact and code for at least two overlapping open reading frames over more than 60% of their length. This characteristic is shared by the two snake endogenous hepatnavirus sequences, both of which encode a polymerase in the +1 frame, as well as the expected overlapping surface protein in the +3 frame (electronic supplementary material, figure S1). Based on this observation, we propose snake eHBVs be considered basally diverging stem-group members of the Hepadnaviridae family.

There are multiple examples where it appears that EVE-derived genes have been exapted by their host, and currently fulfill a variety of host cellular functions [19–22,65]. Our results indicate that snake EVEs show no strong signatures of purifying or positive selection, but instead are evolving neutrally. This indicates that if they play or have played a functional role in the host, it is unlikely to be at the protein level. Thus, the absence of nonsense mutations in eSCV1–5 and eSHBV2 most likely reflects the recent integration of these EVEs into snake genomes, rather than functional conservation owing to purifying selection. A future challenge in palaeovirology studies will be to test whether EVEs could have played and/or still play a role at the RNA level, notably in antiviral immunity [66–69].

**Figure 3.** Unrooted tree of the bornavirus nucleoprotein sequences. Asterisks indicate endogenous viruses. Numbers on branches correspond to bootstrap support values more than 70%. Symbols # indicate posterior probabilities more than 0.9. Mammalian endogenous bornaviral sequences were taken from Horie et al. [44]. The alignment used to reconstruct this tree is provided in the electronic supplementary material, dataset S1. (Online version in colour.)
Finally, our study provides yet another stunning illustration of the fact that, in spite of the extremely high mutation rates characterizing most modern viruses, the fossils of ancient viral genomes can be identified in their host genomes even after millions, or tens of millions of years [1,12,15]. As observed in previous studies [17,26,70,71], the long-term substitution rates we have inferred for hepadnaviruses and circoviruses (between $1.2 \times 10^{-8}$ and $7.7 \times 10^{-8}$ subst. site$^{-1}$ year$^{-1}$ and between $5.9 \times 10^{-8}$ and $7.6 \times 10^{-8}$ subst. site$^{-1}$ year$^{-1}$, respectively), although clearly approximate, are three to five orders of magnitude slower than short-term rates calculated using sequences of circulating viruses (e.g. $1.2 \times 10^{-3}$ subst. site$^{-1}$ year$^{-1}$ for circoviruses [53] and $7.9 \times 10^{-5}$ subst. site$^{-1}$ year$^{-1}$ for hepadnaviruses [52]). It is known that many factors may influence viral substitution rates and their variation over time [72]. Various methodological biases probably influence the calculation of both short- and long-term rates and therefore could explain the large difference between these rates [3,5,73,74], which has also been observed in cellular organisms [75]. Specifically, it has been recently shown that pervasive purifying selection could be one of the major obstacles in accurately estimating the ancient age of recent pathogens [76], and the development of substitution models accounting for this phenomenon are beginning to yield more realistic results [76,77]. It is also important to consider that short-term rates based on pathogenic viruses (which most likely jumped relatively recently) may reflect the ongoing evolutionary arms race with their host, as opposed to their long-term evolutionary history [26]. Furthermore, viral evolution may be constrained owing to both intrinsic structural features of the virus and the diverse immunological strategies of the host [78]. We believe combining palaeovirology and metagenomic studies of modern viruses holds exciting promise for understanding diverse modes of viral evolution and to better grasp the dynamics of long- and short-term host virus interactions.

The specimen was collected in accordance with a scientific collecting permit (State of Arizona, M30207246) and registered Institutional Animal Care and Use Committee protocols (University of Texas at Arlington).

Acknowledgements. C.G. thanks all the technical staff of the UMR CNRS 7267 for assistance in the laboratory and Richard Cordaux and Julien Thézé for insightful discussions. We thank two anonymous reviewers for helpful comments.
References


