Characterization and Complete Nucleotide Sequence of an Unusual Reptilian Retrovirus Recovered from the Order Crocodylia

Joanne Martin,* Peter Kabat,† Elisabeth Herniou, and Michael Tristem

Department of Biological Sciences, Imperial College, Silwood Park, Ascot, Berkshire SL5 7PY, United Kingdom

Received 17 July 2001/Accepted 4 February 2002

A novel group of retroviruses found within the order Crocodylia are described. Phylogenetic analyses demonstrate that they are probably the most divergent members of the Retroviridae described to date; even the most conserved regions of Pol show an average of only 23% amino acid identity when compared to other retroviruses.

The Retroviridae are a family of selfish genetic elements with a host range restricted to vertebrates (2, 5–7, 9, 22). They are currently subdivided into seven genera, all but one of which are harbored by mammalian or avian hosts (8, 11, 20, 21). There are currently no full-length retroviral sequences recovered from amphibians or reptiles, but analysis of PCR-amplified fragments from both vertebrate classes indicates that they harbor many elements that are only distantly related to other retroviruses (9, 18, 19).

We have previously described a large number of novel retroviruses via PCR amplification of approximately 1kb of the pol gene, followed by phylogenetic analysis (9, 14). During these studies, we characterized eight very unusual retroelements recovered from the order Crocodylia. The elements were present in all three extant families (the Alligatoridae, Crocodylidae, and Gavialidae (Table 1), and it is likely that similar elements remain to be found in many of the remaining 14 species constituting the order Crocodylia (1).

All eight elements were found to encode at least one in-frame stop codon or frameshift mutation, indicating they were endogenous in origin. We were unable to detect related elements in other organisms, either by PCR screening of other vertebrate taxa or by low-stringency hybridization of genomic DNA obtained from several birds, reptiles, amphibians, and fish. It is therefore likely that related elements are not widespread within vertebrate genomes.

To characterize these elements further, we constructed a genomic DNA library from liver tissue derived from a captive Nile crocodile (Crocodylus niloticus). Genomic DNA was digested with Sau3A and 10- to 15-kb fragments were ligated into BamHI-digested lambda EMBL3, packaged, and plated. Nine plaque-purified positive clones were identified; three, termed CnEVI to -III, were fully sequenced.

Open reading frame (ORF) maps of CnEVI to -III revealed that each carried multiple in-frame stop codons and frameshift mutations. Analysis of a consensus sequence constructed from all three elements indicated that their original genomic organization probably consisted of two large ORFs, corresponding to the major retroviral genes gag and pol, and a small third ORF immediately upstream of the 3' long terminal repeat (LTR).

CnEVI contains a 593-bp 5' LTR and a 585-bp 3' LTR which differ from each other by approximately 7%, indicating that integration occurred some time ago. We were unable to unambiguously identify promoter or polyadenylation signals, but a putative polyuridine tract and primer binding site were identified adjacent to the 3' and 5' LTRs, respectively. The primer binding site showed 14 of 18 matches to human tRNA (Ser).

The CnEVI Gag polyprotein is 526 residues in length and contains a consensus myristylation sequence (Met-Gly-X3-Ser) (21). No other obvious homology was observed with other retroviral Gag polyproteins. In particular, we were unable to identify either a Cys-His box or a major homology region (21). Translation of Pol requires a ~1 ribosomal frameshift and, as seen with other retroviruses, we identified a slippery sequence together with an associated hairpin loop and pseudoknot immediately upstream of the 3' end of gag (3, 4, 15). The predicted Pol polyprotein is 998 residues in length, with a gene order of protease, reverse transcriptase (RT), RnaseH, and integrase (Int). An additional ORF is located 3' of pol, although BLAST searches of its translated product failed to reveal any similarity with other proteins. There was no evi-

TABLE 1. Host species used in this study

<table>
<thead>
<tr>
<th>Family and species</th>
<th>Common name</th>
<th>Retroviral product</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Alligatoridae</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alligator sinensis</td>
<td>Chinese alligator</td>
<td>RV Chinese alligator I + II</td>
</tr>
<tr>
<td><em>Paleosuchus palpebrosus</em></td>
<td>Smooth-fronted caiman</td>
<td>RV smooth-fronted caiman</td>
</tr>
<tr>
<td><em>Caiman latirostris</em></td>
<td>Broad-nosed caiman</td>
<td>RV broad-nosed caiman</td>
</tr>
<tr>
<td><em>Crocodylidae</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crocodylus niloticus</td>
<td>Nile crocodile</td>
<td>RV Nile crocodile with CnEVI-III</td>
</tr>
<tr>
<td><em>Crocodylus intermedius</em></td>
<td>Orinoco crocodile</td>
<td>RV Orinoco crocodile</td>
</tr>
<tr>
<td><em>Gavialidae</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gavialis gangeticus</td>
<td>Gharial</td>
<td>Gharial II</td>
</tr>
</tbody>
</table>

*Corresponding author. Mailing address: Department of Biological Sciences, Imperial College, Silwood Park, Ascot, Berkshire SL5 7PY, United Kingdom. Phone: 44 (20) 75942306. Fax: 44 (20) 75942339. E-mail: j.martin01@ac.ic.uk.

†Present address: Institute of Virology, Slovak Academy of Sciences, 842 45 Bratislava, Slovak Republic.
ence of an env gene, as is the case with some other members of the Retroviridae (10, 13, 17).

Table 2 shows the average percentage of amino acid identities between CnEVI to -III and the corresponding region (RT domains 1 to 7 [22]) from other retroelements. Sequence identity varied from 20% (human immunodeficiency virus type I) to 27% (human spumavirus [HSV]) with members of the Retroviridae and 18 to 24% with the gypsy-type LTR retrotransposons. This unusually low level of sequence identity is under-scored by comparing both HSV and gypsy to the same retroviral isolates. HSV, one of the most divergent retroviruses described to date, shows a higher level of identity than CnEVI to -III in every comparison, and this is also generally the case for gypsy itself (which is not even a member of the Retroviridae).

We next constructed phylogenies based on a 340-residue alignment of retroelement Pol polyproteins. Phylogenies were constructed using both the neighbor-joining (Fig. 1a) and max-

![Phylogenetic tree](http://jvi.asm.org/DownloadedFrom/4652 NOTES J. VIROL. J. VIROL.)
imum parsimony (Fig. 1b) approaches with PAUP (for phylogenetic analysis using parsimony) (16). Clustering of the crocodilian elements gained strong bootstrap support, but their relationship to other retroviruses was not strongly supported. Although all trees placed the crocodilian virus group within the Retroviridae, the actual position of group members varied somewhat depending on the exact taxa within the data set. For this reason, we performed topological constraint analyses, forcing the crocodilian elements into certain locations and comparing the resulting tree to the best or minimum tree (length,

TABLE 2. Percentage of similarity between CnEVI to -III, prototypical members of the Retroviridae, and LTR retrotransposons

<table>
<thead>
<tr>
<th></th>
<th>MuLV</th>
<th>HSV</th>
<th>MMTV</th>
<th>MPMV</th>
<th>BLV</th>
<th>RSV</th>
<th>HIV-1</th>
<th>WDSV</th>
<th>gypsy</th>
<th>Del</th>
<th>RT412</th>
<th>copia</th>
<th>Ty1</th>
<th>Tat1</th>
</tr>
</thead>
<tbody>
<tr>
<td>CnEVI to -III</td>
<td>22</td>
<td>27</td>
<td>22</td>
<td>24</td>
<td>23</td>
<td>21</td>
<td>20</td>
<td>23</td>
<td>24</td>
<td>18</td>
<td>20</td>
<td>13</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>HSV</td>
<td>38</td>
<td>29</td>
<td>30</td>
<td>28</td>
<td>27</td>
<td>27</td>
<td>32</td>
<td>30</td>
<td>30</td>
<td>24</td>
<td>28</td>
<td>15</td>
<td>16</td>
<td>10</td>
</tr>
<tr>
<td>gypsy</td>
<td>28</td>
<td>30</td>
<td>24</td>
<td>24</td>
<td>28</td>
<td>28</td>
<td>25</td>
<td>26</td>
<td>33</td>
<td>42</td>
<td>16</td>
<td>17</td>
<td>13</td>
<td></td>
</tr>
</tbody>
</table>

*Calculated from amino acid alignment of RT domains 1 to 7, inclusive. Values are based on the calculation of average similarities observed for CnEVI to -III. MuLV, murine leukemia virus; MMTV, mouse mammary tumor virus; MPMV, Mason-Pfizer monkey virus; BLV, bovine leukemia virus; RSV, Rous sarcoma virus; HIV-1, human immunodeficiency virus type 1; WDSV, walleye dermal sarcoma virus.*
5,515 steps) (Fig. 1b). Only four to eight extra steps were required to place the crocodilian group in several locations within the phylogeny, usually clustering with (or close to) the spumaviruses or basal to all retroviral branch elements. A higher number of steps (5,532) were required to place the crocodilian elements as sister taxa to the gypsy-type LTR retrotransposons, with 5,545 steps needed to place these elements basal to a clade containing both the gypsy-type LTR retrotransposons and the Retroviridae. Only the last placement was significantly unlikely, using the Kishino-Hasegawa test (12), although some trees placing the crocodilian elements next to the gypsy-type LTR retrotransposons had scores of \( P < 0.1 \).

Thus, although our phylogenetic analyses were unable to determine the exact relationships of the crocodilian group to other retroelements, they did suggest that these elements are highly likely to lie within, or basal to, the Retroviridae. Trees placing the crocodilian elements outside the Retroviridae required at least 17 additional steps over the minimum tree shown in Fig. 1b, and such topologies were never observed during analyses. Furthermore, LTR retrotransposons almost never use a hairpin-mediated gag-pol frameshift and usually lack a myristylated Gag protein, and these features are both present within the CnEVI genome.

Despite their probable placement within the Retroviridae, clearly the crocodilian elements are very distantly related to other retroelements, they did suggest that these elements are highly likely to lie within, or basal to, the Retroviridae. Trees placing the crocodilian elements outside the Retroviridae required at least 17 additional steps over the minimum tree shown in Fig. 1b, and such topologies were never observed during analyses. Furthermore, LTR retrotransposons almost never use a hairpin-mediated gag-pol frameshift and usually lack a myristylated Gag protein, and these features are both present within the CnEVI genome.

Nucleotide sequence accession numbers. The PCR-amplified sequences and full-length elements have been submitted to the EMBL, GenBank, and DDBJ databases (accession numbers AJ438133 to AJ438138 for the PCR fragments and AJ438130 to AJ438132 for CnEVI to -III).

We thank J. Gatesy (American Museum of Natural History) for the crocodile samples. We thank A. Trnka for providing crocodile liver samples. Thanks also to C. Lynch and A. Burt for discussion.

REFERENCES