Chapparvoviruses occur in at least three vertebrate classes and have a broad biogeographic distribution

William Marcel de Souza,1,2 Marília Farignoli Romeiro,2 Marcílio Jorge Fumagalli,2 Sejal Modha,1 Jansen de Araujo,3 Luzia Helena Queiroz,4 Edison Luiz Durigon,3 Luiz Tadeu Moraes Figueiredo,2 Pablo Ramiro Murcia1 and Robert James Gifford1,*

Abstract

Chapparvoviruses are a highly divergent group of parvoviruses (family Parvoviridae) that have recently been identified via metagenomic sampling of animal faeces. Here, we report the sequences of six novel chapparvoviruses identified through both metagenomic sampling of bat tissues and in silico screening of published vertebrate genome assemblies. The novel chapparvoviruses share several distinctive genomic features and group together as a robustly supported monophyletic clade in phylogenetic trees. Our data indicate that chapparvoviruses have a broad host range in vertebrates and a global distribution.

Parvoviruses are small, non-enveloped viruses that have ssDNA genomes ~5 kb in length. They encode two gene cassettes: a non-structural replicase gene (NS) that encodes the enzymes required for replication and a capsid (VP) gene encoding structural proteins. Two parvovirus subfamilies are recognized: Densovirinae, which contains viruses that infect invertebrate hosts, and Parvovirinae, which contains viruses that infect vertebrate hosts. A total of eight genera have been recognized within the subfamily Parvovirinae [1–3]. Here, we report the identification via sequencing of six new members of the recently proposed genus Chapparvirus. We use these data to examine the genome structures and evolutionary relationships of these novel viruses.

All previously described chapparvoviruses have been detected by metagenomic sequencing. The prototypic member of the proposed genus, Eidolon helvum parvovirus 2 (EhPV-2), was identified in throat swabs taken from the fruit bat Eidolon helvum [4]. Three additional chapparvovirus sequences have been identified via metagenomic screening of turkey faeces [3], rat faeces [5] and rectal swabs of pigs [6]. It is currently not known whether these viruses are associated with disease.

The first of six novel chapparvoviruses identified in our study was recovered via metagenomic sequencing of tissue samples derived from common vampire bats (Desmodus rotundus). Kidney samples were obtained from eight D. rotundus individuals captured in a rural area of Araçatuba city, São Paulo State, Brazil, in June 2010. Pooled samples were used to generate cDNAs and prepared for high-throughput sequencing using TruSeq Universal Adapter (Illumina) protocols (RAPID module) and standard multiplex adaptors. A paired-end, 150-base-read protocol in RAPID module was used for sequencing on an Illumina HiSeq 2500 instrument as recommended by the manufacturer’s protocol. A total of 7,133,306 paired-end reads were generated with 78.12% of bases ≥Q30 (with a base call accuracy of 99.9%). Assembly of Illumina reads using metaViC [7] led to recovery of a sequence spanning a near-complete parvovirus genome (Fig. 1). Phylogenetic and genomic analysis established that this sequence represented a virus closely related to EhPV-2, which we refer to as Desmodus rotundus parvovirus (DrPV-1). The DrPV-1 genome is 4,284 nt in size and has a typical parvovirus genome organization (Fig. 1).

An additional five chapparvovirus sequences were identified by in silico screening of whole-genome shotgun
(WGS) sequence assemblies in various databases. The database-integrated genome screening tool [8] was used to screen WGS data of 281 vertebrate species (Table S1, available in the online Supplementary Material) for sequences homologous to parvoviral proteins and to tentatively classify these sequences into genera. This screen identified all previously identified parvovirus endogenous viral elements (EVEs) – all of which group closely with the Dependoparvovirus and Protoparvovirus genera – and a small number of novel ones. Most of the novel sequences disclosed homology to dependoparvoviruses, protoparvoviruses or amdoparvoviruses (Table S2), but unexpectedly, five disclosed homology to chapparvoviruses. Each of these was identified in a distinct species genome assembly. Two were identified in WGS assemblies of mammalian species, including a bat (Myotis davidii) and a New World primate, the white-headed capuchin (Cebus imitator). Further, chapparvovirus-related sequences were obtained from WGS assemblies of a reptile, the brown spotted pit viper (Protobothrops mucrosquamatus), and two avian species, the Atlantic canary (Serinus canaria) and the brown mesite (Mesitornis unicolor).

Sequences derived from parvoviruses are known to occur as EVEs in a wide range of animal genomes [9–12]. These sequences are thought to represent the remnants of ancient viruses that became fully or partially integrated into the germline of their hosts through non-homologous recombination events. However, all of the chapparvovirus-related sequences identified in WGS assemblies occurred within relatively short contigs, and since none contained any sequence that we could unambiguously identify as genomic, we could not definitively determine whether they represented integrated sequences (EVEs) or were sequences of exogenous viral DNAs that were present in the original DNA sample from which WGS genome data were generated. Notably, however, none of the sequences showed any evidence of a lengthy residence in the host germline (e.g. stop codons, frameshifting mutations in viral ORFs, transposable element insertions). In addition, all previously described parvovirus EVEs group within or close to the relatively closely related Dependoparvovirus, Protoparvovirus and Amdoparvovirus genera (Fig. 2). The Chapparvovirus genus is only distantly related to these two genera and is separated from them in phylogenies by three other genera (Bocaparvovirus, Tetraparvovirus and Erythroparvovirus).

Fig. 1. Genome structures of novel parvovirus reported here. The length of the determined nucleotide sequences of the viral sequences is shown in parentheses. Solid-lined boxes and dashed-lined arrows indicate complete or truncated sequence of ORFs, respectively. Truncated termini of ORFs are indicated by an arrow-shaped edge. ORFs were inferred by manual comparison of putative peptide sequences to those of closely related exogenous parvoviruses. Green and red arrowheads on NS1 indicate the position of conserved amino acid motifs of parvoviruses.
that do not appear to have generated any EVEs (based on current information). Together, these data suggest that the chapparvovirus sequences we identified in WGS assemblies are likely to be infectious viruses present in the DNA samples used for shotgun sequencing, rather than EVEs.

The amino acid sequence identities of novel chapparvovirus sequences shared to those previously published in GenBank were 34–75 % in replicase and 41–55 % in capsid. In contigs that included a complete replicate ORF, the predicted gene product was ~650–672 amino acids in length. Conserved amino acid motifs ‘HVH’ and ‘GPXNTGKS’, the putative endonuclease metal coordination motif ‘HIH’ and the helicase motif ‘GPASTGKS’ were all present (Fig. 1) [13, 14].

As shown in Fig. 1, all six sequences spanned at least part of the replicate gene, including a region that is relatively well conserved across all viruses in the subfamily Parvovirinae. We constructed a multiple sequence alignment spanning 113 residues within this region and containing representative Parvovirinae reference sequences in addition to novel sequences. A combination of automated procedures (MAFFT,
Table 1. Sample information names, sources, sample, locality and environment of viruses reported in this study

<table>
<thead>
<tr>
<th>Virus</th>
<th>Source</th>
<th>Sample</th>
<th>Location</th>
<th>Date</th>
<th>Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>DrPV-1</td>
<td>D. rotundus</td>
<td>Pool of kidney</td>
<td>Araçatuba city, São Paulo, Brazil</td>
<td>23 June 2010</td>
<td>Native</td>
</tr>
<tr>
<td>Cebus capucinus imitator chapparvirus</td>
<td>C. capucinus imitator (adult male)</td>
<td>Missing</td>
<td>Costa Rica</td>
<td>Missing</td>
<td>Missing – killed by a vehicle</td>
</tr>
<tr>
<td>Mesitornis unicolor chapparvirus</td>
<td>Mesitornis unicolor (female)</td>
<td>Missing</td>
<td>Madagascar</td>
<td>Missing</td>
<td>Native</td>
</tr>
<tr>
<td>Protobothrops muscosquamus chapparvirus</td>
<td>P. muscosquamus</td>
<td>Missing</td>
<td>Okinawa, Japan</td>
<td>2014</td>
<td>Missing</td>
</tr>
<tr>
<td>Myotis davidi chapparvirus</td>
<td>Myotis davidi</td>
<td>Spleen, kidney and small intestine</td>
<td>Taiyi Cave, Xianning, China</td>
<td>21 August 2011</td>
<td>Native</td>
</tr>
<tr>
<td>Serinus canaria chapparvirus</td>
<td>S. canaria</td>
<td>Missing</td>
<td></td>
<td>Missing</td>
<td>Missing</td>
</tr>
</tbody>
</table>

Complete capsid ORFs are present in two previously obtained chapparvirus genome sequences (rat parvovirus and porcine parvovirus 7) and two that were obtained in our study (DrPV-1 and Cebus capucinus parvovirus). Where complete capsid ORFs are present, they are significantly shorter than found in other members of the subfamily Parovirinae (i.e. ~500 amino acids as compared to ~700). Also, the predicted capsid proteins of newly characterized chapparviruses contained phospholipase A2 motifs in their N-terminal regions. These motifs, which are reportedly involved in intracellular trafficking and/or escape from endosomes, are found in many, but not all, members of the Paroviridae [20, 21]. Notably, they have been reported to be absent from previously reported chapparvirus sequences [5, 6]. Phylogenetic relationships between capsid sequences were inferred using the methodology described above for replicase, and those mirrored those obtained for replicase (Fig. 2b).

We noted that the replicase and capsid genes of chapparviruses often overlap slightly (~8–11 nucleotides; see Table S3), a trait that has only been observed in one other genus (Erythroparvovirus) within the Parovirinae. The relatively small size of the chapparvirus capsid protein, combined with the presence of overlap between the capsid and replicase genes, suggests a selection pressure for smaller genome size in these viruses. If we assume that the capsid gene found in these viruses shares a common origin with those found in other Parovirinae genera, then it appears that this genus has evolved a smaller overall genome size, reducing the size of the capsid gene, while the replicase gene has remained approximately unchanged. Interestingly, this goes against the well-established hypothesis that virus genome size is physically limited by length constraints on genes encoding icosahedral capsids [22–24]. However, since we could not identify any regions of unambiguous homology between the chapparvirus capsid proteins and those found in other Parovirinae genera, an alternative scenario can also be considered wherein the shorter chapparvirus capsid gene has a separate evolutionary origin to the one found in the other genera.

The clustering of chapparvirus sequences into host-class-specific sub-lineages (see Fig. 2) is consistent with their being derived from viruses that have been evolutionarily associated with their different hosts. We collected information on the location and context of sampling for samples that were used to generate the metagenomic and WGS sequence datasets, and mapped the biogeographic associations of samples onto the replicate phylogeny (Table 1, Fig. 2). These data show that chapparviruses have an extensive geographic distribution and likely have a worldwide distribution in many different hosts. In future studies, we expect that these viruses will be found in many other hosts – perhaps without causing disease in most cases.

Funding information
This work was supported by the Medical Research Council (grant no. MC_UU_12014/10) and by the Fundação de Amparo à Pesquisa do Estado de São Paulo, Brazil (grant no. 13/14929-1 and scholarships no. 12/24150-9, 15/05778-5, 14/20851-8, 16/01414-1 and 06/ 00572-0).

Acknowledgements
We thank Luiz Gustavo Betim Góes (Institute of Biomedical Sciences, University of São Paulo) and Cristiano de Carvalho (Faculty of Veterinary Medicine, São Paulo State University) for help with capture of bats. We also thank Colin R. Parrish (Cornell University) and Andrew Davison (MRC-University of Glasgow Centre for Virus Research) for their useful comments, which helped to improve this manuscript.

Conflicts of interest
The authors declare that there are no conflicts of interest.
References