Novel bat adenoviruses with low G+C content shed new light on the evolution of adenoviruses

Bing Tan,¹,² Xing-Lou Yang,¹ Xing-Yi Ge,¹ Cheng Peng,¹ Hai-Zhou Liu,¹ Yun-Zhi Zhang,³ Li-Biao Zhang⁴ and Zheng-Li Shi¹,²,*

Abstract

Bats have been reported to carry diverse adenoviruses. However, most bat adenoviruses have been identified on the basis of partial genome sequences, and knowledge on the evolution of bat adenoviruses remains limited. In this study, we isolated and characterized four novel adenoviruses from two distinct bat species, and their full-length genomes were sequenced. Sequence analysis revealed that these isolates represented three distinct species of the genus Mastadenovirus. However, all isolates had an exceptionally low G+C content and relatively short genomes compared with other known mastadenoviruses. We further analysed the relationships among the G+C content, 5′-C-phosphate-G-3′ (CpG) representation and genome size in the family Adenoviridae. Our results revealed that the CpG representation in adenoviral genomes depends primarily on the level of methylation, and the genome size displayed significant positive correlations with both G+C content and CpG representation. Since ancestral adenoviruses are believed to have contained short genomes, those probably had a low G+C content, similar to the genomes of these bat strains. Our results suggest that bats are important natural reservoirs for adenoviruses and play important roles in the evolution of adenoviruses.

INTRODUCTION

The family Adenoviridae comprises five genera, Mastadenovirus, Aviadenovirus, Siadenovirus, Atadenovirus and Ichtaadenovirus [1], which are distinguished by genus-common genes and genus-specific genes in their genome organization and by levels of sequence similarity. Members of the genera Mastadenovirus and Aviadenovirus are highly host-specific and infect mammals and birds, respectively, whereas members of Siadenovirus have been reported in both birds and amphibians. Atadenoviruses also display a wide host range, including mammals, birds and reptiles. Only one adenovirus (AdV) has been isolated from fish and is currently assigned to the genus Ichtaadenovirus.

AdVs contain a linear, dsDNA genome flanked by an inverted terminal repeat (ITR) on each end. The genome sizes and G+C contents of AdVs vary widely within the viral family and show high variation among different genera: genomes of atadenoviruses that have been fully sequenced range from 27 751 to 33 213 bp, with G+C contents of 33.6–53.5 %; those of members of Siadenovirus range from 26 163 to 26 340 bp, with G+C contents of 34.2–38.5 %; those of mastadenoviruses range from 30 536 to 37 860 bp, with relatively high G+C contents of 43.6–63.9 %; and those of aviadenoviruses range from 42 940 to 45 810 bp, with G+C contents of 44.7–66.9 % (Table 1). Most recently, an unusually low G+C content of 36 % was reported for the genome of Californian sea lion mastadenovirus-1 (CSLAdV-1), which infects Zalophus californianus [2].

It has been suggested that the AdV genome evolved with an increase in its length [3]. Based on this hypothesis, the ancestral genome might have contained a minimum of essential genes required for replication and later enlarged through multiple mechanisms, such as recombination, gene acquisition, gene duplication, etc. One of the examples is the E3 transcription unit from members of the species Human mastadenovirus D (HAdV-D). This region encodes eight proteins, of which three share conserved region 1. These genes are not present in non-primate AdVs, but their
One of the best-known mutation biases, from a methylated
believed to contribute to the variation to some extent [7, 8].

The reason for the great G+C content variations in AdV
genomes is not well understood. In vertebrates, natural
selection facilitating gene regulation has been proposed as
an explanation for such variations due to the associations
observed between G+C content and important genomic
properties [6]. Alternatively, neutral mechanisms, including
mutation biases and GC-biased gene conversions, are
believed to contribute to the variation to some extent [7, 8].

One of the best-known mutation biases, from a methylated
5′-C-phosphate-G-3′ (CpG) dinucleotide to TpG (CpA on
the complementary strand), is due to a high methylcytosine
deamination frequency [9]. Since CpG is a predominant tar-

get of DNA methyltransferase activity, CpG suppression is
commonly observed in the genomes of vertebrates and con-
tributes to the bias in the G+C content [10]. Similarly, sup-
pressed CpG has been commonly observed in RNA viruses
and small DNA viruses (<30 kb) infecting vertebrates but
not in the majority of large DNA viruses (>30 kb) [11]. It is
not clear if viruses share the mechanism of CpG suppression
with their hosts, but in the family Parvoviridae CpG deple-
tion has also been associated with the level of methylation
[12]. In AdVs, the representation of CpG has been reported
in limited cases and was found to fall in the normal range
but tended to be slightly suppressed [13]. However, CpG
representation has not been studied in most AdVs, and its
relative effects on the base composition remain to be
clarified.

Previously, we and other investigators have isolated nine
AdVs from bats: bat adenovirus-1 (BtAdV-1) FBV1 from
Pteropus dasymallus yayeyamae [14], BtAdV-2 PPV1 from
Pipistrellus pipistrellus [15], BtAdV-3 TJM from Myotis rick-
etti [16], BtAdV 1050597 from Rousettus leschenaulti [17],
Eidolon helvum AdV-1 [18], BtAdVs WIV9–11 from
Rhinolophus sinicus [19] and the very recently identified
BtAdV 250-A from Corynorhinus rafinesquii [20]. Besides,
PCR screening of AdVs in different bat species has revealed
their high prevalence and great genetic diversity [16, 21–
26]. However, full-length genome sequences were deter-
mined for only TJM, PPV1, WIV9–11 and 250-A [16, 19,
20, 27]. In this study, we describe the isolation and genomic
characterization of four novel bat AdVs with relatively short
genomes and unusually low G+C content and analyse the
relationships among CpG representation, methylation level
and genome size.

### RESULTS

**Isolation and identification of four novel AdVs from bats**

Four virus strains were successfully isolated, two from Mini-
opterus schreibersii samples infecting MsIn cells and two
from R. leschenaultii samples infecting RlKi cells. Based on
the sequences of the 261 bp DNA polymerase (pol) genes
amplified by PCR [16], the M. schreibersii strains shared
74% nucleotide similarity, and both displayed the highest
nucleotide sequence similarity (71 and 72%, respectively) to
a bat mastadenovirus (GenBank accession no. KC692425)
infesting Pteropus giganteus [24]. On the other hand,
the R. leschenaultii strains displayed 99% nucleotide
sequence similarity and showed the highest similarity (88 %)
to isolate 1050597 [17]. According to the naming order for
live viruses isolated in our laboratory at the Wuhan Institute
of Virology (WIV), the M. schreibersii strains were tenta-

tively named BtAdV WIV12 and 13, and the R. leschenaultii
strains BtAdV WIV17 and 18.

**Full-length genome characterization of the novel bat AdVs**

The sequences of these bat AdV genomes were determined
using next-generation sequencing, with coverage depth
of >500 times achieved for WIV12 and 13 and >200 times

### Table 1. Genomic characterization of novel AdV strains WIV12, 13, 17 and 18 and comparison with known AdVs

<table>
<thead>
<tr>
<th>Virus</th>
<th>Host</th>
<th>Genome size (bp)</th>
<th>ITR (bp)</th>
<th>G+C content (mol %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BtAdV 250-A</td>
<td>Corynorhinus rafinesquii</td>
<td>31 484</td>
<td>239</td>
<td>49.8</td>
</tr>
<tr>
<td>BtAdV PPV1</td>
<td>Pipistrellus pipistrellus</td>
<td>31 616</td>
<td>137</td>
<td>53.5</td>
</tr>
<tr>
<td>BtAdV TJM</td>
<td>Myotis ricketti</td>
<td>31 806</td>
<td>128</td>
<td>56.9</td>
</tr>
<tr>
<td>BtAdV WIV9–11</td>
<td>Rhinolophus sinicus</td>
<td>37 545–38 073</td>
<td>50, 51</td>
<td>55.0</td>
</tr>
<tr>
<td>BtAdV WIV12</td>
<td>Miniopterus schreibersii</td>
<td>29 581</td>
<td>73</td>
<td>34.2</td>
</tr>
<tr>
<td>BtAdV WIV13</td>
<td>Miniopterus schreibersii</td>
<td>29 162</td>
<td>61</td>
<td>31.3</td>
</tr>
<tr>
<td>BtAdV WIV17</td>
<td>Rousettus leschenaulti</td>
<td>29 923</td>
<td>178</td>
<td>34.3</td>
</tr>
<tr>
<td>BtAdV WIV18</td>
<td>Rousettus leschenaulti</td>
<td>29 812</td>
<td>177</td>
<td>34.2</td>
</tr>
<tr>
<td>CSLAdV-1</td>
<td>California sea lion</td>
<td>31 709</td>
<td>109</td>
<td>36.0</td>
</tr>
<tr>
<td>Other mastadenoviruses</td>
<td>Mammals</td>
<td>30 536–37 860</td>
<td>93–371</td>
<td>43.6–63.9</td>
</tr>
<tr>
<td>Aviadenoviruses</td>
<td>Birds</td>
<td>42 940–45 810</td>
<td>54–95</td>
<td>44.7–66.9</td>
</tr>
<tr>
<td>Atadenoviruses</td>
<td>Birds, mammals and reptiles</td>
<td>27 751–33 213</td>
<td>46–118</td>
<td>33.6–53.5</td>
</tr>
<tr>
<td>Siadenoviruses</td>
<td>Amphibians and birds</td>
<td>26 163–26 340</td>
<td>29–39</td>
<td>34.2–38.5</td>
</tr>
</tbody>
</table>
achieved for WIV17 and 18. The full-length genomes of these isolates ranged in size from 29,162 to 29,923 bp, with ITRs of 61–178 bp (Table 1 and Fig. 1). The G+C contents of these isolates ranged from 31.3 to 34.3%, i.e. fell into the lowest range among all known AdVs. The global pairwise comparison of these genomes using mVISTA revealed an average sequence similarity of 73% between WIV12 and 13 and 90% between WIV17 and 18 (Fig. 1).

Twenty-eight to 30 ORFs were predicted as protein-encoding genes in the four genomes, with a typical genomic organization of the genus *Mastadenovirus* (Fig. 1). Most of the putative gene products displayed sequence similarities ranging from 20 to 84% to known homologues, including those from members of *Mastadenovirus* infecting bovines, canines, humans, ovines, porcines, simians, sea lions and tree shrews.

Except for the E3 and E4 transcription units, these isolates displayed genomic structures highly similar to those of known bat AdVs that have been fully sequenced [16, 19, 20, 27]. The E3 units of PPV1, TJM and 250-A are ~2 kb in size and harbour the 12.5K and ORFA genes. Those of WIV9-11 are >6 kb in size with the 12.5K gene and two unique genes encoding polypeptides of ~1500 and 99 amino acids. By contrast, each E3 of the novel isolates obtained in this study was ~1 kb in size and carried a single gene encoding a 14.7K homologue. In the E4 unit, the reported bat AdVs

---

**Fig. 1.** Genomic characterization of newly identified bat AdVs. The genomes of BtAdV WIV12 (a) and 18 (b) are represented by a double line with the scale indicated under the line. Predicted ORFs and the ITR sequences are shown as arrows and rectangles, respectively. Exons of spliced transcripts are linked with broken lines. The E4 genes are shown with their ORF numbers. WIV13 and 17 displayed identical genome organization to that of WIV12 and 18, respectively, except that WIV17 lacked E4 ORF1 in the genome. Below the genome maps of WIV12 and 18 are global pairwise comparisons of homology to related AdV strains. The height of each point along the y-axis indicates 50–100% identity of genome pairs. The nucleotide positions noted along the x-axis correspond to those in the genomes of WIV12 (a) and 18 (b).
have two conserved (34K and ORF6/7) and four or five variable genes. By contrast, WIV12 and 13 harboured only four E4 genes, 34K, two unique ORFs and a dUTPase homologue. WIV17 and 18 both contained genes 34K, ORF6/7 and unique ORFs 1–3, with an additional ORF1’ present in the E4 region of WIV18.

**Phylogenetic analysis of the novel bat AdVs**

Phylogenetic analysis was performed to better understand the evolutionary relationships between the four bat AdVs and known strains. Based on the analysis of the full-length pol genes, four main branches were formed representing the main genera, *Aviadenovirus*, *Siadenovirus*, *Atadenovirus* and *Mastadenovirus* (the fish AdV was not included in this analysis) (Fig. 2a). Similar to the previously reported bat AdVs, the four novel strains fell into the *Mastadenovirus* group. WIV12 and 13 were in sister clades diverging from the WIV17 and 18 branch. The distances among WIV12, 13 and WIV17/18 warrant the establishment of distinct species, while that between WIV17 and 18 does not [1]. TJM and

---

**Fig. 2.** Phylogenetic analysis of bat AdVs. (a) Analysis was conducted based on the full-length amino acid sequences of DNA polymerase (pol). GenBank accession numbers for the sequences of strains included in the compressed subtrees are as follows: bovine atadenovirus-4 (AF306902), bovine atadenovirus-6 (JQ345700), duck atadenovirus-1 (AC_000004), ovine atadenovirus-7 (NC_004037), snake atadenovirus-1 (DQ106414), fowl aviadeno-virus-1 (AC_000014), fowl aviadeno-virus-4 (GU188428), fowl aviadeno-virus-5 (KC493644), fowl aviadeno-virus-8 (NC_014969), fowl aviadeno-virus-9 (NC_000899), goose aviadeno-virus-4 (JF510462), turkey aviadeno-virus-1 (GU936707), frog aviadeno-virus-1 (AF224336), raptor aviadeno-virus-1 (EU715130), skua aviadeno-virus-1 (HM853533), turkey aviadeno-virus-3 (AC_000016). (b) Analysis was conducted based on a 219 bp fragment of the pol gene. The GenBank accession number of each strain is as follows: bovine atadenovirus-6 (JQ345700), duck atadenovirus-1 (AC_000004), ovine atadenovirus-7 (AC_000004), snake atadenovirus-1 (DQ106414), fowl aviadeno-virus-1 (AC_000014), fowl aviadeno-virus-4 (GU188428), fowl aviadeno-virus-5 (KC493644), fowl aviadeno-virus-8 (NC_014969), fowl aviadeno-virus-9 (NC_000899), goose aviadeno-virus-4 (JF510462), turkey aviadeno-virus-1 (GU936707), frog aviadeno-virus-1 (AF224336), raptor aviadeno-virus-1 (EU715130), skua aviadeno-virus-1 (HM853533), turkey aviadeno-virus-3 (AC_000016).
PPV1 have been formally recognized as representing the species *Bat mastadenovirus A* and *Bat mastadenovirus B*, and the recently identified strains WIV9–11 and 250-A clearly represent two additional species, tentatively named as *Bat mastadenovirus C* and *Bat mastadenovirus D*. In this study, we propose to classify WIV12, 13 and 17/18 as representing the tentative species *Bat mastadenovirus E*, *Bat mastadenovirus F* and *Bat mastadenovirus G*, respectively.

When partial *pol* gene sequences available for all bat AdVs were used, WIV12 and 13 still formed independent sister clades. By contrast, WIV17 and 18 were found to cluster with several types infecting fruit bats, 1050597 from India, FBV1 from Japan and two *P. giganteus* strains from Bangladesh (Fig. 2b).

**CpG representation in AdV genomes is responsible for G+C contents**

CpG representation in AdV genomes is a well-known discrepancy in the genomes of vertebrates, and it may affect the G+C content [10]. To test this parameter in the family *Adenoviridae*, the ratios of observed/expected CpG (O/E<sub>CpG</sub>) were calculated and analysed for all available full-length AdV genomes, and O/E<sub>GpC</sub> ratios were also calculated for comparison (detailed in Methods). The O/E<sub>CpG</sub> ratios varied widely, between 0.373 and 1.196 (Fig. 3a), whereas the O/E<sub>GpC</sub> ratios ranged from 0.903 to 1.287, with the exception of 1.395 found for bovine AdV-2 (Fig. 3b). The O/E<sub>CpG</sub> ratios of most AdVs tended to be less than 1; however, murine AdV-2, tree shrew AdV-1 and most aviadenoviruses were the exceptions. The O/E<sub>CpG</sub> ratios of the previously studied bat AdVs ranged from 0.542 to 0.855, whereas those of the novel bat types ranged from 0.373 to 0.455.

Notably, a strong positive correlation between the G+C content and O/E<sub>CpG</sub> was observed within the family *Adenoviridae* based on Pearson’s correlation coefficient (Fig. 3c). In contrast, the G+C content showed only a weak negative correlation with O/E<sub>GpC</sub> (Fig. 3d).

**CpG depletion in AdV genomes is affected by methylation**

To investigate if the CpG representation in AdV genomes was affected by a methylation-induced CpG mutation bias, O/E values of CpG and TpG+CpA were calculated and analysed. Consistent with the mutational bias [9], a strong negative correlation was observed between the values of CpG and TpG+CpA in the family *Adenoviridae* (Fig. 4a). In particular, the correlation observed between the average deviation frequency of CpG and TpG+CpA was very strong (Fig. 4b).

---

**Fig. 3.** Correlation analysis of the G+C content and CpG representation in the family *Adenoviridae*. Distribution of CpG (a) and GpC (b) representations. The O/E ratios of CpG varied widely and tended to be plotted at the low side, except for the members of the genus *Aviadenovirus*. By contrast, those of GpC were confined. At, Avi, Mast, Sia, Pri-AdVs and Oth-Mast indicate members of the genera *Atadenovirus*, *Aviadenovirus*, *Mastadenovirus* and *Siadenovirus*, primate AdVs, and other mastadenoviruses, respectively. The G+C content was plotted versus the O/E ratios of CpG (c) and GpC (d). The correlations were tested using Pearson’s correlation coefficient.

---

*Tan et al., Journal of General Virology 2017;98:739–748*
Suppressed G+C and CpG are not associated with a recent host switch in AdVs

Based on the recognition of unmethylated CpG motifs by the host immune system, suppressed G+C content and CpG representation in viral genomes have been suggested to indicate recent host-jumping events [2, 28]. However, the phylogenetic analysis of WIV12, 13, 17 and 18 indicated co-evolution between the viruses and their present hosts. We further investigated if the host switch affects the base composition of primate AdV genomes as primate strains are the best-studied AdVs and comprise >300 members (Fig. 5). Our results clearly indicated that the G+C content and O/E<sub>CpG</sub> ratios were specific within a given AdV species, and no decreases in either value were observed for any strains that potentially evolved through recent host-jumping events.

G+C and CpG contents are linked to the genome size in AdVs

The genomes of Siadenovirus and Atadenovirus members are relatively short and characterized by a low G+C content within the family Adenoviridae. Similar characteristics were also observed in the isolates obtained in this study. These observations prompted us to question whether the size of an AdV genome was correlated with its G+C content. To test this hypothesis, the sizes of all available full-length genomes of AdV members were plotted against their G+C contents (Fig. 6a). Based on Pearson’s correlation coefficient, a moderate positive correlation between the G+C content and genome size was observed in the family. Even in the genera Aviadenovirus (r<sup>2</sup>=0.68, P=0.008) and Mastadenovirus (r<sup>2</sup>=0.56, P<0.001), the correlations were also observed.

Since the relative abundance of CpG is strongly associated with the G+C content in the family Adenoviridae, we further tested the correlation between O/E<sub>CpG</sub> and the genome size. As a result, a strong positive correlation was observed (Fig. 6b).

DISCUSSION

Previously, we reported a high prevalence and genetic diversity of AdVs in bats [16, 19]. In this study, we have described the isolation and characterization of four novel bat AdVs. These isolates are substantially different from the previously reported AdVs and are considered to represent three distinct species within the genus Mastadenovirus according to taxonomic classification criteria. In addition, these strains have relatively short genomes with a significantly low G+C content compared with other known mastadenoviruses and are thus close to atadenoviruses and siadenoviruses. Together with the data from previous reports, bat AdVs found to date show a great diversity not only in genome sequences but also in genome sizes (from 29,162 to 38,073 bp) and G+C contents (from 31.3 to 56.9%), which almost cover the entire ranges for the family Adenoviridae.

Low G+C contents are one of the characteristics of the genera Siadenovirus and Atadenovirus. In this study, the newly discovered bat AdVs were found to share this characteristic. Based on the partial sequences of previously reported AdVs, low G+C contents are probably also characteristic of the genomes of strains FBV1, 1050597 and E. helvum AdV-1 [14, 17, 18]. In addition, a low G+C content has also been reported for the genome of CSLAdV-1 [2]. Thus, low G+C contents seem to be characteristic of not only members of Siadenovirus and Atadenovirus but are also common in the genus Mastadenovirus. Although the genus Atadenovirus was named based on the high percentage (>60 %) of A+T in the genomes of its initial members [3], subsequent studies revealed non-biased nucleotide compositions in some other members of this genus [28–30]. Therefore, the name ‘atadenovirus’ now appears somewhat misleading, and a more appropriate nomenclature is required for this genus.

Consistent with observations in vertebrates [10], AdVs display a wide range of G+C contents along with variable CpG representation. The novel bat isolates were in the

---

Fig. 4. Correlation analysis of the representations of CpG and TpG+CpA in AdV genomes. (a) The O/E ratios of CpG were plotted versus those of TpG+CpA. (b) The average deviation frequencies of these dinucleotides were also plotted. The correlations were tested using Pearson’s correlation coefficient. At, Avi, Mast and Sia indicate members of the genera Atadenovirus, Aviadenovirus, Mastadenovirus and Siadenovirus, respectively.
lowest range of both G+C content and CpG relative abundance. In the family Adenoviridae, most members, except avian adenoviruses, tend to show low CpG representation. By contrast, the O/E_{GpC} ratios of AdVs were slightly >1.

The overestimation of O/E_{GpC} values could be explained by CpG depletion as the decreased frequency of C (and G in the complementary strand) led to underestimation of the frequency of expected GpC. The strong correlation...
observed between O/E_{CpG} and the G+C content revealed that the relative abundance of CpG has an important effect on the base composition of AdV genomes. The imperfect correlation indicates that alternative evolutionary factors could also be involved, such as structural constraints, viral genes and regulatory elements, and host-related driving forces.

The mutational bias increasing the mutation rate of methylated CpG toward TpG (and CpA in the complementary strand) has been well recognized to affect the CpG representation in the genomes of vertebrates and some vertebrate viruses [10, 12, 31]. The mechanism of methylation of AdV genomes is not well understood. In actively replicating genomes of HAdV-2 and 12, only few or no methylated bases have been identified, whereas viral genomes integrated into host genomes are known to be hypermethylated [32]. Our observation of the nearly perfect correlation between Δf_{CpG} and Δf_{TpG+CpA} in the family Adenoviridae indicated that CpG representation in AdV genomes depends primarily on the level of methylation.

Toll-like receptor 9 (TLR9) is known to recognize unmethylated CpG motifs and trigger innate immune responses [33]. In previous reports, a low G+C content in AdV genomes has been suggested to result from rapid adaptation to new hosts after interspecies transmission [2, 28]. However, our analysis indicated that the low G+C content in the genomes of WIV12, 13, 17 and 18 was not a result of recent interspecies transmissions. The clear phylogenetic distance between WIV12 and 13 indicated a long evolutionary history of the two strains in the present host, and the close relationship between WIV17/18 and other AdVs infecting fruit bats also suggested virus-host co-evolution. Notably, no decrease in the G+C content and O/E_{CpG} ratio was observed in the primate AdVs that potentially evolved through recent host-jumping events. In fact, TLR9 and CpG motifs seem unlikely to be predominant factors in the induction of innate immune responses during AdV infection because (a) HAdV type 12 (46.5% G+C, O/E_{CpG}=0.812) was more active in inducing innate immune responses than types 2 (55.2% G+C, O/E_{CpG}=0.887) and 5 (55.2% G+C, O/E_{CpG}=0.881) [34]; (b) TLR9 was not required in non-plasmacytid dendritic cells and in vivo for IFN induction by AdV infection [35, 36]; and (c) AdVs import their genome into the nucleus directly through capsid docking to the nuclear pore [37], which may allow the viruses to spatially avoid the recognition by TLR9. TLR9 is localized in endosomal compartments of immune cells [33, 38]. Therefore, TLR9 seems not to be a major factor affecting the base composition of AdV genomes.

Interestingly, correlations between the G+C content and O/E_{CpG} as well as the genome length were observed in the family Adenoviridae. One of the potential explanations is that AT-rich genes are more likely to have short lengths as the stop codon has a bias toward A and T [39]. As viral genomes with longer sizes tend to encode more genes [40], these correlations also suggest that the increased CpG and G+C content may facilitate mechanisms such as gene acquisition and duplication, which lead to the enlargement of AdV genomes. Alternatively, if the acquired genes are resistant to host cell methylation capabilities, they may facilitate the increases in CpG and G+C contents. If we accept the hypothesis that AdVs originated from a short genome [3], the ancestral strain was likely to have exhibited a low G+C content and suppressed CpG level. Given the great genetic diversity of bat AdVs, bats are likely to be important natural reservoirs for AdVs and play an important role in AdV evolution.

**METHODS**

**Sample collection and bat species identification**

As described previously [16], single faeces of *M. schreibersii* and *R. leschenaultii* were collected in Mojiang, Yunnan Province, China, in 2012. The bat species were confirmed by PCR amplification and sequencing of the complete cytochrome *b* gene [41].

---

**Fig. 6.** Correlation analysis of the adenoviral genome size, G+C content and CpG representation. The genome size was plotted versus the G+C content (a) and O/E_{CpG} ratio (b). The correlations were measured using Pearson’s correlation coefficient. Members of *Adenovirus*, *Aviadenovirus*, *Mastadenovirus* and *Siadenovirus* are indicated as At, Avi, Mast and Sia, respectively.
Cell culture and virus isolation

Virus isolation was performed using MsIn and RlKi cells originated from M. schreibersii and R. leschenaultii, respectively [19]. Bat faeces were used to inoculate cells after homogenization. The inocula were sub-cultured for five passages. Viral particles were purified through a 40% sucrose cushion and centrifuged at 110,000 g using an SW28 rotor (Beckman Coulter). Purified viruses were observed under electron microscopy and analysed by a PCR assay [16].

Viral DNA preparation and full-length genome sequencing

DNA was extracted from purified viral particles using phenol–chloroform [19] and sequenced using an FLX genome sequencer (454; Life Sciences). The resulting 454 raw reads were assembled using the de novo assembler in Geneious 5.5.9 [42]. The terminal sequences of each genome were confirmed using 3’-RACE PCR (primers are available on request) as described previously [19].

Bioinformatics analysis

General analyses of genome sequences were performed as described previously [16], including finding ORFs, prediction of splicing nature, comparison of protein identity and phylogenetic analysis. The global genome pairwise comparison of splicing nature, comparison of protein identity and GC-content evolution was carried out using the mVISTA LAGAN tool (http://genome.lbl.gov/vista/index.shtml) [43].

Statistical analysis

Data were analysed with Pearson’s correlation coefficient using the R program software (www.r-project.org/), which calculates $r^2$ and $P$-values. The results were considered significant at $P<0.05$. The strength of the correlations was divided into five orders of magnitude: no or very weak correlation ($r^2=0.0–0.19$), weak ($r^2=0.2–0.39$), moderate ($r^2=0.4–0.59$), strong ($r^2=0.6–0.79$) and very strong ($r^2=0.8–1$) correlation.

The odds ratio of dinucleotide $XpY$ ($O/E_{XpY}$), measured by comparing the average ratio of observed to expected frequencies ($f$), is a common assessment of dinucleotide bias (in other words, $O/E_{XpY} = f_{XpY}/f_{XfY}$). When observed and expected frequencies are equal, this results in an odds ratio of 1; otherwise, the relative abundance of $XpY$ tends to be over- or underrepresented. In vertebrates, $O/E_{CPG}$ is ~0.4, indicating a loss of ~60% of the expected frequency of Cpg [44].

In dsDNA, $f_A = f_T = (f_A + f_T)/2$; $f_C = f_G = (f_C + f_G)/2$; and $f_{TPG} = f_{CPA} = (f_{TPG} + f_{CPA})/2 = f_{TPG+CPA}$. Therefore, $O/E_{TPG} = O/E_{CPA} = f_{TPG}/f_{TPG} = O/E_{TPG+CPA} = 2(f_{TPG} + f_{CPA})/(f_{G} + f_{C})(f_T + f_A)$. All other dinucleotides comply with these rules.

The average deviation frequency of $XpY$ dinucleotides ($\Delta f_{XpY}$) between the observed and expected frequencies is calculated as $f_{XpY} - f_{XfY}$. Similarly, $\Delta f_{TPG+CPA} = (\Delta f_{TPG} + \Delta f_{CPA})/2$.

A total of 383 distinct sequences were used in this analysis, which included all of the full-length genome sequences of AdVs available in the GenBank database. Sequences of low quality were excluded. The GenBank accession numbers used in this analysis are presented in Table S3.

Funding information

This work was jointly funded by the National Natural Science Foundation of China (81290341) (to Z.-L. S.), China Mega-Project for Infectious Disease (2014ZX10004001-003) (to Z.-L. S.) and Scientific and Technological Basis Special Project (2013FY113500) (to Z.-L. S. and Y.-Z. Z.) from the Ministry of Science and Technology of the People’s Republic of China, and Science and Technology Planning Project of Guangdong (2015A020209093) and Funds for Environment Construction and Capacity Building of GDAS’ Research Platform (2016DASPT-0215) (to L.-B. Z.).

Conflicts of interest

The authors declare no conflict of interest.

Ethical statement

No animals were used in the present study.

References


