Podgorski, I. I., Pantó, L., Papp, T., Harrach, B., & Benkö, M. (2016). Genome analysis of four old world monkey adenoviruses supports the proposed species classification of primate adenoviruses and reveals signs of possible homologous recombination.
Journal of General Virology, 97(7), 1604-1614. doi:10.1099/jgv.0.000465

Title: Genome analysis of four Old World monkey adenoviruses supports the proposed species classification of primate adenoviruses and reveals signs of possible homologous recombination

Running title: Analysis of SAdVs supports novel mastadenovirus species

Iva I. Podgorski*, Laura Pantó[a], Tibor Papp, Balázs Harrach, Mária Benkő

Institute for Veterinary Medical Research, Centre for Agricultural Research, Hungarian Academy of Sciences, Budapest, Hungary
[a]Present address: Laboratory of Genome Sciences, Graduate School of Information Science and Technology, Hokkaido University, Japan

*Author for correspondence: ivapodgorski@gmail.com, 00385 98 1930 944

Keywords: simian adenovirus, genome, evolution, species, U exon protein, recombination

Subject category: Animal - small DNA viruses

Word count: 4031

The GenBank accession numbers for the full genome sequences of simian adenoviruses are KP329561 (SAdV-8), KP329562 (SAdV-11), KP329564 (SAdV-16) and KP329565 (SAdV-19)

Abstract

Within the family *Adenoviridae* presently, *Simian mastadenovirus A* is the single species approved officially for monkey AdVs, whilst the establishment of six further species (*Simian mastadenovirus B* to *Simian mastadenovirus G*) has been proposed in the last few years. We examined the genetic content and phylogenetic relationships of four Old World monkey (OWM) AdV types (namely SAdV-8, -11, -16 and -19) which had been proposed to be classified into different AdV species: SAdV-11 to *Human mastadenovirus G*, and the other three viruses into three novel species. By full genome sequencing, we identified gene contents characteristic for the genus *Mastadenovirus*. Among the 36 ORFs, two genes of different lengths, predicted to code for the adenoviral cellular attachment protein (the fibre), were found. The E3 regions contained six genes, present in every OWM AdV, but lacked the E3 19K gene which has seemingly appeared only in the ape (hominid) AdV lineages during evolution. For the first time in SAdVs, the two other exons belonging to the gene of the so-called U exon protein were also predicted. Phylogenetic calculations, based on the fibre-1 and the major capsid protein, the hexon, implied that recombination events might have happened between different AdV species. Phylogeny inference, based on the viral DNA-dependent DNA polymerase and the penton base protein, further supported the species classification proposed earlier.

Introduction

Adenoviruses (AdVs) are dsDNA viruses widespread among humans and vertebrate animals, mostly non-pathogenic for their hosts although, in rare cases, they can cause infections with significant consequences (Benkő, 2015). Primate AdVs are members of the genus *Mastadenovirus*, and while we have relatively much knowledge about human and ape (chimpanzee, gorilla and bonobo) AdVs, the AdVs of the more ancient primate lineages such as Old World monkeys (OWM), New World monkeys and prosimians are hardly known. Although OWM AdVs were discovered more than 50 years ago (Hull *et al.*, 1956), and were found in monkeys of many different species (macaques, grivets, black and white colobuses, red colobuses, hamadryas baboons, yellow baboon), more than half of the described types have not been studied in detail. The interest in more ancient simian AdVs (SAdVs) rises with the awareness of the potential risk they may pose for humans in case of host switching (Benkő *et al.*, 2014). On the other hand, there is an increasing interest in gene delivery vectors derived from non-human

AdVs (Lopez-Gordo *et al.*, 2014), especially from SAdVs since they are the closest relatives to human AdVs (HAdVs), but evolutionally still far enough for not being influenced by pre-existing immunity in the human population.

In the following text, informal abbreviations will be used for the species names, e.g. *Human mastadenovirus A*, HAdV-A; *Simian mastadenovirus A*, SAdV-A. While all known HAdV types are grouped unambiguously into seven established species, HAdV-A to HAdV-G, there is only one species accepted officially for the classification of monkey AdVs, SAdV-A, containing OWM AdVs exclusively. The very first phylogenetic analysis of SAdVs (SAdV-1 to -25) was based on the very short sequences of the virus-associated RNA (VA-RNA) genes, studied in both OWM (SAdV-1 to -20) and chimpanzee (SAdV-21 to -25) AdVs (Kidd *et al.*, 1995). More than ten years ago, the first full sequence of an OWM AdV (SAdV-3) was published (Kovács *et al.*, 2004). The phylogenetic distance of SAdV-3 from all HAdVs, known at that time, warranted a novel species (established later as SAdV-A) which would contain OWM AdVs only. The next fully sequenced OWM AdV was SAdV-1, being the first SAdV recognised to have two fibre genes (Kovács *et al.*, 2005). Phylogenetically, SAdV-1 has been found to be closer to HAdVs with two fibre genes (in species HAdV-F) than to SAdV-3 (Kovács *et al.*, 2005). A few years ago, a new species, SAdV-B was proposed for OWM AdVs (Roy *et al.*, 2012). Additional OWM AdVs have been described in olive baboons (BaAdV-1, BaAdV-2/4 and BaAdV-3) along with a proposal for BaAdV-2/4 and -3 to form a new species, SAdV-C (Chiu *et al.*, 2013). Most recently, another SAdV type (strain 23336) from rhesus macaque has been proposed to form a new species, SAdV-D (Malouli *et al.*, 2014). Subsequently, we have proposed a species designation for all the 20 known serotypes of OWM AdVs. According to our proposal, serotype SAdV-13 would be the sole member of species SAdV-D (Pantó *et al.*, 2015), whereas, in the future, an additional novel monkey AdV species might be needed for the non-serotyped SAdV strain 23336 (Malouli *et al*., 2014). The establishment of three additional species (SAdV-E to SAdV-G) has been proposed in the same paper.

The main purpose of the present study was to provide further support, by full genomic sequence analyses, for the proposed new SAdV species. To this end, four OWM AdVs (SAdV-8, -11, -16 and -19) were sequenced and their genetic content and phylogenetic relationships scrutinised. Our results indeed warranted the formation of species SAdV-B, SAdV-C and SAdV-E. Furthermore, for the first time from SAdVs, all the three exons of the putative gene of the "U

exon protein" (UXP) were discerned. This might help the complete identification of this spliced gene in other AdVs in the future.

Results

General features. The DNA of four SAdV strains was fully sequenced and analysed. The main characteristics of the new genome sequences are summarised in Table 1, in comparison with the range of the examined values in known members of each SAdV species where the newly sequenced viruses are proposed to belong to. SAdV-16 is an exception as being the first representative of a newly proposed species, SAdV-E. Genomic assemblies revealed that all four genomes contain 36 putative coding regions characteristic for mastadenoviruses. These included two genes, of different lengths, predicted to code for the cellular attachment protein (the fibre) of AdVs. No homologue of the E3 19K gene, usually present in members of the ape (hominid) AdV lineages, was found in any of the studied genomes, and each of them contained only a single copy of VA-RNA gene.

The majority of the deduced protein sequences of SAdV-8, -11 and -19, were very similar to (sharing >95% identity with) their counterparts in other members of the AdV species (namely SAdV-B, HAdV-G and SAdV-C, respectively) where they are proposed to be classified. In the SAdV-8 genome, there is one exception to this, the UXP with 88% identity. For SAdV-11, lower identity values were seen in the E1A gene (88%), E3 region genes (in some cases as low as 40%), and UXP (85%). The deduced protein sequences of SAdV-16 did not exhibit particular similarity to any known AdVs, except the hexon, penton base and pVIII. In the SAdV-19 genome, there are several exceptions, including the hexon (89% identity), the proteins coded by genes in the E3 region (in some cases the identity is as low as 64%), UXP (66% identity), fibre-1 (only 27% identity to SAdV-C members, while 46% to HAdV-G members), fibre-2 (66% identity), and ORF6/7 (85% identity).

Phylogeny inference. Phylogeny reconstructions, performed with the aa sequences deduced from the DNA-dependent DNA polymerase (*pol*), penton base, fibre-1 and fibre-2 genes are presented in Fig. 1. SAdV-8 and -11 always clustered clearly with the SAdVs that had been proposed to form species SAdV-B (Roy *et al.*, 2012), and HAdV-G (Jones *et al.*, 2007), respectively. SAdV-16 formed an independent branch proposed to be accepted as a new species, SAdV-E (Pantó *et al.*, 2015). SAdV-19 appeared among the members of the previously proposed

species SAdV-C (Chiu *et al.*, 2013), except on tree based on fibre-1 (Fig. 1c). Interestingly, the phylogeny inference based on the hexon aa sequences (Fig. 2) implied divergent relationships among several AdVs. These contradictions could be explained by the results of recombination analysis of the SAdV-19 genome in comparison to that of members of species HAdV-G, HAdV-F and SAdV-C (Fig. 3a and b). The SimPlot and BootScan analyses indicated that recombination event(s) might have happened in the hexon gene. Nonetheless, our conclusion is that the results of phylogeny inference based on these proteins need to be handled with care.

U exon protein. The UXP sequences of the four SAdVs were compared to their counterparts in members of species HAdV-C in order to determine the position and splicing sites of all the three UXP exons (Tollefson *et al.*, 2007). In this study, the UXP sequences of three ape AdVs, sequenced by others earlier, namely chimpanzee (SAdV-34), gorilla (SAdV-43) and bonobo (SAdV-44) AdV (Roy *et al*., 2009), were identified in this study as well by comparison of the sequences to members of the species HAdV-C. Main characteristics of the UXPs are shown in Table 2. UXP sequences were aligned to compare the degree of conservation of the three exons in different AdVs (Fig. 4). Derived from the presence of splice donor and acceptor sites, the putative positions of the three UXP exons in the genomes are summarised in Fig. 5.

Discussion

Here we report the genomic characterisation of four OWM AdVs and discuss their taxonomical classification. One virus (SAdV-11) was found to belong to the previously established species HAdV-G, whereas each of the other three viruses seemed to represent a different species. Thus the proposals for the establishment of SAdV-B, SAdV-C and SAdV-E were supported. In most cases, the taxonomic classification could be decided unequivocally. SAdV-16 was an exception inasmuch as several additional aspects had to be considered for its species allocation.

The size of SAdV genomes, sequenced to date, range between 31,045 (SAdV-7; Roy *et al.*, 2011) and 36,838 bp (SAdV-20; Roy *et al.*, 2012). The four newly sequenced SAdV genomes fall within this range, and have 36 putative genes characteristic for members of the genus *Mastadenovirus*, including the presence of two fibre genes, a feature recognised in members of species HAdV-G and HAdV-F and in several monkey AdVs before (Fig. 2) (Alonso-Padilla *et al.*, 2015; Pantó *et al.*, 2015). Besides mastadenoviruses, the presence of two fibre genes have

also been found in many representatives of different aviadenovirus species (Kaján *et al.*, 2012; Kaján *et al.*, 2010; Marek *et al.*, 2014a; Marek *et al.*, 2014b; Zhao *et al.*, 2015) but only in two members of the genus *Atadenovirus* (Pénzes *et al.*, 2014; To *et al.*, 2014). The G+C content, an important AdV species demarcation criterion (Harrach, 2014) among SAdVs, varies between 47.8% (SAdV-20) and 65.7% (SAdV strain A1139; Roy *et al.*, 2012). However, within each of the species the differences in the base composition do not exceed 3% (Pantó *et al*., 2015). The G+C content of the four SAdVs presently studied conformed to this rule, and usually well corresponded to the narrow range of G+C content of the species they were proposed to belong to (Table 1).

The existence of a protein, coded by the U exon, was predicted more than 20 years ago (Davison *et al.*, 1993), but the entire gene with its three exons, has been described in members of the species HAdV-C only (Tollefson *et al.*, 2007). Deletion or truncation of the U exon results in impaired virus replication and causes aberrant localisation of the DBP in the nucleus of infected cells (Tollefson *et al.*, 2007). Amino acid sequence alignments of the predicted UXPs of SAdVs and HAdVs (Fig. 4) from species HAdV-C showed that the first and second exons are relatively well conserved, even in AdVs of different species, whilst the third exon is extremely variable both in sequence content and length (Table 2). Nonetheless, the genomic localisation of all the three exons in the studied SAdVs was comparable (Fig. 5). The third exon of the UXP gene overlaps with the coding region of DBP gene. Interestingly, in the individual AdV species, the position of the third exon relative to the DBP gene seems to be well conserved. More precisely, the splice acceptor site of the third exon was found to be 6, 9 or 15 nt upstream from the start codon of the DBP (Fig. 5). Thus the third exon was always in the next frame when compared to that of the DBP. The previously described and newly revealed putative UXP sequences might be of help in defining the complete gene sequences in other AdVs in the future. However, we can expect that defining all the three exons of the UXP in non-primate AdVs might be more challenging due to the increase of divergence in the sequences of evolutionally more distant AdV species.

All OWM AdVs, studied earlier, have been found to have one VA-RNA gene only. On the other hand, human and chimpanzee AdVs (in species HAdV-B to HAdV-E) have been described to possess two VA-RNA genes (Kidd *et al.*, 1995; Larsson *et al.*, 1986). In certain primate AdVs, the VA-RNA genes have not been studied yet, as the PCR used for their amplification has failed

either due to the high specificity of the primers, or because the genes had been missing indeed from the genomes of some viruses (Kidd *et al*, 1995). Nonetheless, the VA-RNA gene of SAdV-16 (strain SA7) was characterised almost 30 years ago, proving that this OWM AdV has only one such gene (Larsson *et al.*, 1986). Our SAdV-16 sequence confirmed the presence of this VA-RNA gene, albeit in one of the earlier studies its PCR amplification had failed probably for reasons described above (Kidd *et al.*, 1995). The sequence of the VA-RNA genes of SAdV-11 and SAdV-19 has been published earlier (Kidd *et al.*, 1995), and here we report their exact position in the genomes. However, our analysis revealed a longer (164 versus 104 nt) VA-RNA gene in SAdV-11. This size difference was identified as a 60-nt "deletion" between the 68-nt-long 5'and 36-nt-long 3'ends that were completely identical in our sequence and in the VA-RNA gene reported earlier (Kidd *et al*., 1995). The reason for the gene fragment missing from the previously reported sequence might be the formation of secondary structures interfering with the PCR amplification of the given locus. In general, VA RNAs of OWM AdVs seem to be shorter (93 to 104 nt) than those of HAdVs. Among the few, exceptions are SAdV-11 (164 nt), SAdV-13 (146 nt), as well as SAdV-16 and -19 (168 nt in both). All these viruses are proposed to be members of different species: HAdV-G, SAdV-D, SAdV-E and SAdV-C, respectively (Pantó *et al.*, 2015). Furthermore, we also proved for the first time the presence of a VA RNA in SAdV-8. This gene is longer (159 nt) than the VA RNAs of most OWM AdVs. SAdV-8 is a member of the proposed species SAdV-B, from which we do not have information about the VA-RNA gene of any other member. All OWM AdVs, studied to date, contain only one VA-RNA gene, confirming the results published in earlier studies (Kidd *et al.*, 1995; Larsson *et al.*, 1986).

Phylogenetic trees, based on the full *pol* (Fig. 1a), penton base (Fig. 1b) and fibre-2 (Fig.1d) aa sequences, also confirmed that SAdV-11 belongs to the previously established species HAdV-G, whereas the classification of SAdV-8, -16 and -19 requires the establishment of the previously proposed species SAdV-B, SAdV-C, and SAdV-E, respectively (Chiu *et al.*, 2013; Pantó *et al.*, 2015; Roy *et al.*, 2012). Interestingly however, the tree based on the fibre-1 aa sequence completely separates SAdV-19 from its proposed species SAdV-C, but also from all the other species we know (Fig. 1c). Comparison of this fibre with the available sequences revealed that it shares only 46% or less aa identity with the fibre sequence of members of species HAdV-G. With the recombination analyses, we were unable to find any AdV in the known species which could be supposed as the origin of this fibre gene (Fig. 3). Nonetheless the evolutionary tree

indicates that it might be the most ancient of all the known primate AdVs with two fibre genes (Fig. 1c).

The hexon-based tree shows different relationships among several simian and human AdVs (Fig. 2). This might be the result of some recombination events. SimPlot analysis of SAdV-19 clearly indicated the probability of a homologous recombination in the hexon gene, and the BootScan analysis suggested that the SAdV-19 hexon gene resulted from a recombination event between yet unknown members of species HAdV-F and HAdV-G (Fig. 3a). The hexon-based phylogenetic tree (Fig. 2) could not separate the two species from species SAdV-C. This also supports the probability of several recombination events among these AdVs. Homologous recombinations most often in the hexon and fibre genes have been described in many primate AdVs (Chiu *et al.*, 2013; Crawford-Miksza *et al.*, 1996; Dehghan *et al.*, 2013a; Dehghan *et al.*, 2013b; Walsh *et al.*, 2011). Consequently, a divergent topology of the other AdV lineages could also be observed on the hexon tree. For example, HAdV-4 and -16 are separated from other members of their species HAdV-E and HAdV-B, respectively. This is not surprising since it has been shown that the two viruses share very high overall nucleotide sequence identity in the hexon gene (Pring-Akerblom *et al.*, 1995). Furthermore, the hexon of chimpanzee AdV-63 (ChAd-63; proposed but officially not classified into species HAdV-C) is very similar to that of SAdV-36 (species HAdV-E), implying the possibility of an interspecies homologous recombination. On most trees SAdV-16 appeared far enough from the species SAdV-B to be considered as representative of a new species, SAdV-E (Fig. 1 and 2). However, the penton base and hexon trees show that SAdV-16 is close to, or falls within, species SAdV-B, respectively. A homologous recombination in the hexon gene of SAdV-16 is therefore very likely. Some other properties of SAdV-16, such as the host origin, the G+C content, and the results of the hemagglutination-inhibition tests (Rapoza, 1967) also support that it should be considered as a new species, distinct from SAdV-B.

A certain ambiguity arises with the classification of SAdV-19 as well. The ITRs of SAdV-19 (127 bp) are longer than those (87 bp) of the other BaAdVs from the species SAdV-C. SAdV-19 has an overall longer genome, and has been isolated from a different baboon species. Nonetheless, other features such as the G+C content (Table 1), as well as the phylogeny inference based on the *pol* and penton base are in favours of placing SAdV-19 into the species SAdV-C.

The results of the present study further support the need for establishing three new SAdV species: SAdV-B, SAdV-C and SAdV-E. The organisation of the genomes of all newly sequenced SAdVs was comparable, and very similar to that of the previously sequenced SAdVs. The gene of the UXP homologue was identified in all the four SAdV genomes based on comparison with the UXP gene of members of species HAdV-C. However, mRNA studies would be essential for the ultimate confirmation of these predictions. By further screening of primates, especially the more ancient New World monkeys and prosimians, numerous additional AdV types and lineages would likely be discovered in the future.

Materials and methods

Cells and virus stocks. Samples of the prototype SAdV strains originating from the American Type Culture Collection (SAdV-8, ATCC VR-1539, strain P-5, from crab-eating macaque; SAdV-11, ATCC VR-206, strain P-10, from rhesus macaque; SAdV-16, ATCC VR-941, strain C-8, from grivet; SAdV-19, ATCC VR-275, strain AA153, from yellow baboon) were used either directly in PCRs (SAdV-16 and -19), or for the inoculation of Vero E6 cells (SAdV-8 and -11) to propagate the virus for next generation sequencing (NGS). After a few passages, seven 175 cm^2 tissue culture flasks were used for virus production. The tissue culture supernatants and the cells (disrupted by three freezing and thawing cycles) were clarified with low-speed centrifugation. Then the virions were sedimented in a Beckman ultracentrifuge and the viral DNA was isolated with phenol-chloroform extraction method.

PCR methodology. The DNA of SAdV-16 and -19 was sequenced by PCR combined with traditional Sanger sequencing with consensus and specific primers designed as described previously (Kovács & Benko, 2009). To amplify the first fragments of the viral genome, a nested PCR with degenerate primers targeting the gene of the *pol* (Wellehan *et al.*, 2004) and that of the IVa2 protein (Pantó *et al.*, 2015; Vidovszky *et al.*, 2015) were used. Subsequently, standard PCR with degenerate primers targeting the hexon gene was used (Kiss *et al.*, 1996). Specific primers, based on partial sequences, were designed with the use of the Primer Designer program version 2.0. Dream Taq DNA polymerase (Fermentas) was found to be optimal for the PCR amplification of the shorter fragments. The PCRs were performed as described previously (Doszpoly *et al.*, 2013). For the amplification of longer (>1000bp) fragments, the Takara PrimeSTAR® Max DNA polymerase was used according to the manufacturer`s

recommendations. The PCR products were purified from agarose gels with the use of the MEGA quick-spin Total Fragment DNA Purification Kit (iNtRON Biotechnology, Kyungki-Do, Korea).

Sequencing and genome assembly. We used two different sequencing approaches. SAdV-8 and -11 were sequenced by NGS method. Their purified genomic DNA was sent to a commercial service (BGI in China) where paired-end sequence reads were generated using the Illumina HiSeq2500 system. The quality of the FASTQ sequences was enhanced by trimming off low-quality bases using the "Trim sequences" option of the CLC Genomics Workbench version 7.0.4. The quality-filtered sequence reads were puzzled into a number of contig sequences. The analysis was performed using the "*de novo* assembly" option of the CLC Genomics Workbench version 7.0.4. The remaining gaps were filled by PCR using specific primers and sequenced by traditional (Sanger) method. The sequences of the genome ends were successfully obtained from the NGS data. The genome of SAdV-16 and -19 was sequenced directly without prior large-scale propagation and purification of the virions. The genome fragments were obtained by PCR and sequenced with the PCR primers on both strands. For the larger fragments, genome walking strategy was applied. The sequences of the genome ends of SAdV-16 and -19 were also determined by PCR. Based on our former experience that every ITR in members of the species HAdV-F, HAdV-G, SAdV-A, SAdV-B and SAdV-C starts with the same octamer (CATCATCA), a primer was designed with a long 5` extension (5`-CACTCGGATTCCATCATCA-3`). This primer was used in pair with a specific, outward-oriented primer in each of the four cases, resulting in ~500 bp fragment. Thus, the ITR sequences could be obtained except the very conserved 8-nt motif. The conditions of the sequencing reactions and nucleotide sequence assembly have been described in detail previously (Pénzes *et al.*, 2014; Tarján *et al.*, 2014).

The genome sequences were annotated with the web-accessible annotation tool Artemis (Berriman & Rutherford, 2003; Marek *et al.*, 2013). Genomic sequence fragments were queried systematically against the non-redundant database of the National Centre for Biotechnology Information, using the BLASTX program online. The sequences of the genes, that are known to contain introns in other AdVs, were checked for the presence of putative splice donor and acceptor sites. Splice sites in the genomes were determined by manual search by comparison with the corresponding regions of the earlier described SAdVs and HAdVs. The UXP gene sequence and location in the genome were determined by comparison with the HAdV-5 UXP

sequence (Tollefson *et al.*, 2007). The VA-RNA gene sequence of SAdV-8 was determined by comparison with the available VA RNA sequences of primate AdVs, with special focus on the criteria for identification of VA-RNA genes and fully conserved nt positions determined in that study (Kidd *et al*., 1995). Similarity plots and bootscanning analyses were performed with Simplot 3.5.1 with window size 1000 bp, step size 50 bp (Lole *et al.*, 1999).

Phylogenetic calculations. Phylogenetic calculations were carried out essentially by a scheme described earlier (Pantó *et al.*, 2015). Analyses were based on full aa sequences deduced from the *pol*, penton base, hexon, fibre-1 and fibre-2 proteins of all primate AdVs sequenced to date. The tree shrew AdV (TSAdV) was also included. In the calculations based on fibre-1 and -2 sequences, the sequence of fowl adenovirus type 1 (FAdV-1) was also used. For the different proteins, the following models were applied: JTT+I+G for *pol*, WAG+I+G for penton base, LG+I+G for hexon, LG+G for fibre-1, and CpREV+G for fibre-2.

The GenBank accession numbers for the full genome sequences of studied simian adenoviruses are KP329561 (SAdV-8), KP329562 (SAdV-11), KP329564 (SAdV-16) and KP329565 (SAdV19). Following AdVs (GenBank acc. number) were used in phylogenetic analysis: HAdV-1 (AF534906), HAdV-2 (ADRCG), HAdV-3 (NC_011203), HAdV-4 (AY487947), HAdV-5 (AC_000008), HAdV-6 (FJ349096), HAdV-7 (AC_000018), HAdV-8 (AB448767), HAdV-9 (AJ854486), HAdV-10 (JN226746), HAdV-11 (AY163756), HAdV-12 (AC_000005), HAdV-13 (JN226747), HAdV-14 (AY803294), HAdV-15 (JN226748), HAdV-16 (AY601636), HAdV-17 (AC_000006), HAdV-18 (GU191019), HAdV-19 (EF121005), HAdV-20 (JN226749), HAdV-21 (AY601633), HAdV-22 (FJ404771), HAdV-23 (JN226750), HAdV-24 (JN226751), HAdV-25 (JN226752), HAdV-26 (EF153474), HAdV-27 (JN226753), HAdV-28 (FJ824826), HAdV-29 (JN226754), HAdV-30 (JN226755), HAdV-31 (AM749299), HAdV-32 (JN226756), HAdV-33 (JN226758), HAdV-34 (AY737797), HAdV-35 (AY128640), HAdV-36 (GQ384080), HAdV-37 (DQ900900), HAdV-38 (JN226759), HAdV-39 (JN226760), HAdV-40 (L19443), HAdV-41 (DQ315364), HAdV-42 (JN226761), HAdV-43 (JN226762), HAdV-44 (JN226763), HAdV-45 (JN226764), HAdV-46 (AY875648), HAdV-47 (JN226757), HAdV-48 (EF153473), HAdV-49 (DQ393829), HAdV-50 (AY737798), HAdV-51 (JN226765), HAdV-52 (DQ923122), SAdV-1 (AY771780), SAdV-3 (AY598782), SAdV-6 (JQ776547), SAdV-7 (DQ792570), SAdV-13 (KP329563), SAdV-18 (CQ982407), SAdV-20 (HQ605912), SAdV-21 (AC_000010), SAdV-22 (AY530876), SAdV-23 (AY530877), SAdV-24 (AY530878), SAdV-25 (AC_000011), SAdV-26

(HB426768), SAdV-27.1 (HC084988), SAdV-27.2 (FJ025928), SAdV-28.1 (HC084950), SAdV-28.2 (FJ025915), SAdV-29 (HC085020), SAdV-30 (HB426704), SAdV-31.1 (HC000816), SAdV-32 (HC085052), SAdV-33 (HC085083), SAdV-34 (HC000847), SAdV-35.1 (HC085115), SAdV-35.2 (FJ025910), SAdV-36 (HC191003), SAdV-37.1 (HB426639), SAdV-37.2 (FJ025919), SAdV-38 (FJ025919), SAdV-39 (HB426607), SAdV-40.1 (HC000785), SAdV-41.1 (HI964271), SAdV-42.1 (HC191035), SAdV-43 (FJ025900), SAdV-44 (FJ025899), SAdV-45 (FJ025901), SAdV-46 (FJ025930), SAdV-47 (FJ025929), SAdV-48 (JQ776547), SAdV-49 (HQ241819), SAdV-50 (HQ241820), SAdV-23336 (KM190146), RhAdV-51 (NC_025826), RhAdV-52 (NC_025827), RhAdV-53 (NC_025828), SAdV-ch1 (KF360047), ChAd3 (CS138463), ChAd6 (CS138464), Chseq13 (HH760489), SAdVch36 (CS479277), Chseq62 (HH760538), Chseq63 (HH760539), Chseq65 (HH760541), SAdV-A1139 (JN880448), SAdV-A1163 (JN880449), SAdV-A1173 (JN880450), SAdV-A1258 (JN880451), SAdV-A1285 (JN880452), SAdV-A1296 (JN880453), SAdV-A1312 (JN880454), SAdV-A1327 (JN880455), SAdV-A1335 (JN880456), BaAdV-1 (KC693021), BaAdV-2 (KC693022), BaAdV-3 (KC693023), titi monkey AdV (TMAdV; HQ913600), FAdV-1 (AC_000014), TSAdV (AC_000190).

Acknowledgements

The authors gratefully acknowledge the financial support provided by the Hungarian Scientific Research Fund grant OTKA K100163, as well as the EU grant ADVance FP7-290002, and Balassi Institute (Tempus Public Foundation).

Table 1. Genome characteristics of the simian adenovirus (SAdV) types studied in this work

Virus type	Proposed species	Genome length (bp)	ITRs length (bp)	ITRs length range in the proposed species (bp)	G+C content (%)	G+C range in the proposed species (%)	Acc. number
SAdV-8	*Simian mastadenovirus B*	35,685	188	166-220	60.3	60.1-62.9	**KP329561**
SAdV-11	*Human mastadenovirus G*	34,510	71	60-133	55.0	55.1-56.3	**KP329562**
SAdV-16	*Simian mastadenovirus E*	35,159	181	NA	57.9	NA	**KP329564**
SAdV-19	*Simian mastadenovirus C*	34,604	127	87	52.2	52.2-52.6	**KP329565**

Table 2. The size and position of the three putative exons of the U exon protein (UXP) identified previously in four human adenoviruses (HAdVs) and in seven SAdV types in this study (highlighted in bold)

Virus	1st exon position (the number of aa coded by each exon is in brackets)	2nd exon position	3rd exon position
SAdV-8	29531-29691 (54)	23643-23719 (25)	22766-23040 (91)
SAdV-11	28624-28781 (53)	22905-22981 (25)	21970-22388 (139)
SAdV-16	29161-29321 (54)	23442-23518 (25)	22590-22870 (93)
SAdV-19	28507-28673 (56)	22855-22931 (25)	21955-22349 (131)
HAdV-1	30927-31090 (55)	24725-24801 (25)	23704-24098 (131)
HAdV-2	30856-31019 (55)	24715-24791 (25)	23676-24088 (137)
HAdV-5	30868-31031 (55)	24668-24744 (25)	23629-24041 (137)
HAdV-6	30835-30998 (55)	24700-24776 (25)	23661-24073 (137)
SAdV-34	32643-32803 (54)	25085-25161 (25)	23701-24380 (226)
SAdV-43	32048-32211 (55)	24955-25031 (25)	23811-24274 (154)
SAdV-44	32686-32846 (54)	25127-25203 (25)	23761-24425 (221)

References

Alonso-Padilla, J., Papp, T., Kaján, G. L., Benkő, M., Havenga, M., Lemckert, A., Harrach, B. & Baker, A. H. (2015). Development of novel adenoviral vectors to overcome challenges observed with HAdV-5 based constructs. *Mol Ther*. doi: 10.1038/mt.2015.194

Benkő, M. (2015). Adenoviruses: Pathogenesis. In *Reference Module in Biomedical Sciences*. Elsevier. doi: 10.1016/B978-0-12-801238-3.02526-5

Benkő, M., Harrach, B. & Kremer, E. J. (2014). Do nonhuman primate or bat adenoviruses pose a risk for human health? *Future Microbiol* **9**, 269-272.

Berriman, M. & Rutherford, K. (2003). Viewing and annotating sequence data with Artemis. *Brief Bioinform* **4**, 124-132.

Chiu, C. Y., Yagi, S., Lu, X., Yu, G., Chen, E. C., Liu, M., Dick, E. J., Carey, K. D., Erdman, D. D., Leland, M. M. & Patterson, J. L. (2013). A novel adenovirus species associated with an acute respiratory outbreak in a baboon colony and evidence of coincident human infection. *MBio* **4**, e00084-13.

Crawford-Miksza, L. K. & Schnurr, D. P. (1996). Adenovirus serotype evolution is driven by illegitimate recombination in the hypervariable regions of the hexon protein. *Virology* **224**, 357-367.

Davison, A. J., Telford, E. A., Watson, M. S., McBride, K. & Mautner, V. (1993). The DNA sequence of adenovirus type 40. *J Mol Biol* **234**, 1308-1316.

Dehghan, S., Seto, J., Jones, M. S., Dyer, D. W., Chodosh, J. & Seto, D. (2013a). Simian adenovirus type 35 has a recombinant genome comprising human and simian adenovirus sequences, which predicts its potential emergence as a human respiratory pathogen. *Virology* **447**, 265-273.

Dehghan, S., Seto, J., Liu, E. B., Walsh, M. P., Dyer, D. W., Chodosh, J. & Seto, D. (2013b). Computational analysis of four human adenovirus type 4 genomes reveals molecular evolution through two interspecies recombination events. *Virology* **443**, 197-207.

Doszpoly, A., Wellehan, J., Childress, A., Tarjan, Z., Kovács, E., Harrach, B. & Benkő, M. (2013). Partial characterization of a new adenovirus lineage discovered in testudinoid turtles. *Infect Genet Evol* **17**, 106-112.

Harrach, B. (2014). Adenoviruses: General features. In *Reference Module in Biomedical Sciences*. Elsevier. doi:10.1016/B978-0-12-801238-3.02523-X

Hull, R. N., Minner, J. R. & Smith, J. W. (1956). New viral agents recovered from tissue cultures of monkey kidney cells. I. Origin and properties of cytopathogenic agents S.V.1, S.V.2, S.V.4, S.V.5, S.V.6, S.V.11, S.V.12 and S.V.15. *Am J Hyg* **63**, 204-215.

Jones, M. S., Harrach, B., Ganac, R. D., Gozum, M. M., Dela Cruz, W. P., Riedel, B., Pan, C., Delwart, E. L. & Schnurr, D. P. (2007). New adenovirus species found in a patient presenting with gastroenteritis. *J Virol* **81**, 5978-5984.

Kaján, G. L., Davison, A. J., Palya, V., Harrach, B. & Benkő, M. (2012). Genome sequence of a waterfowl aviadenovirus, goose adenovirus 4. *J Gen Virol* **93**, 2457-2465.

Kaján, G. L., Stefancsik, R., Ursu, K., Palya, V. & Benkő, M. (2010). The first complete genome sequence of a non-chicken aviadenovirus, proposed to be turkey adenovirus 1. *Virus Res* **153**, 226-233.

Kidd, A. H., Garwicz, D. & Oberg, M. (1995). Human and simian adenoviruses: phylogenetic inferences from analysis of VA RNA genes. *Virology* **207**, 32-45.

Kiss, I., Matiz, K., Allard, A., Wadell, G. & Benkő, M. (1996). Detection of homologous DNA sequences in animal adenoviruses by polymerase chain reaction. *Acta Vet Hung* **44**, 243-251.

Kovács, E. R. & Benkő, M. (2009). Confirmation of a novel siadenovirus species detected in raptors: partial sequence and phylogenetic analysis. *Virus Res* **140**, 64-70.

Kovács, G. M., Davison, A. J., Zakhartchouk, A. N. & Harrach, B. (2004). Analysis of the first complete genome sequence of an Old World monkey adenovirus reveals a lineage distinct from the six human adenovirus species. *J Gen Virol* **85**, 2799-2807.

Kovács, G. M., Harrach, B., Zakhartchouk, A. N. & Davison, A. J. (2005). Complete genome sequence of simian adenovirus 1: an Old World monkey adenovirus with two fiber genes. *J Gen Virol* **86**, 1681-1686.

Larsson, S., Svensson, C. & Akusjärvi, G. (1986). Characterization of a low-molecular-weight virus-associated (VA) RNA encoded by simian adenovirus type 7 which functionally can substitute for adenovirus type 5 VA RNAI. *J Virol* **60**, 635-644.

Lole, K. S., Bollinger, R. C., Paranjape, R. S., Gadkari, D., Kulkarni, S. S., Novak, N. G., Ingersoll, R., Sheppard, H. W. & Ray, S. C. (1999). Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J Virol* **73**, 152-160.

Lopez-Gordo, E., Podgorski, I. I., Downes, N. & Alemany, R. (2014). Circumventing antivector immunity: potential use of nonhuman adenoviral vectors. *Hum Gene Ther* **25**, 285-300.

Malouli, D., Howell, G. L., Legasse, A. W., Kahl, C., Axthelm, M. K., Hansen, S. G. & Früh, K. (2014). Full genome sequence analysis of a novel adenovirus of rhesus macaque origin indicates a new simian adenovirus type and species. *Virol Rep* **3-4**, 18-29.

Marek, A., Ballmann, M. Z., Kosiol, C., Harrach, B., Schlötterer, C. & Hess, M. (2014a). Whole-genome sequences of two turkey adenovirus types reveal the existence of two unknown lineages that merit the establishment of novel species within the genus *Aviadenovirus*. *J Gen Virol* **95**, 156-170.

Marek, A., Kaján, G. L., Kosiol, C., Harrach, B., Schlötterer, C. & Hess, M. (2014b). Complete genome sequences of pigeon adenovirus 1 and duck adenovirus 2 extend the number of species within the genus *Aviadenovirus*. *Virology* **462-463**, 107-114.

Marek, A., Kosiol, C., Harrach, B., Kaján, G. L., Schlötterer, C. & Hess, M. (2013). The first whole genome sequence of a *Fowl adenovirus B* strain enables interspecies comparisons within the genus *Aviadenovirus*. *Vet Microbiol* **166**, 250-256.

Pantó, L., Podgorski, I. I., Jánoska, M., Márkó, O. & Harrach, B. (2015). Taxonomy proposal for Old World monkey adenoviruses: characterisation of several non-human, non-ape primate adenovirus lineages. *Arch Virol* **160**, 3165-3177.

Pring-Akerblom, P., Trijssenaar, F. E. & Adrian, T. (1995). Sequence characterization and comparison of human adenovirus subgenus B and E hexons. *Virology* **212**, 232-236.

Pénzes, J. J., Menéndez-Conejero, R., Condezo, G. N., Ball, I., Papp, T., Doszpoly, A., Paradela, A., Pérez-Berná, A. J., López-Sanz, M., Nguyen, T. H., van Raaij, M. J., Marschang, R. E., Harrach, B., Benkő, M. & San Martín, C. (2014). Molecular characterization of a lizard adenovirus reveals the first atadenovirus with two fiber genes and the first adenovirus with either one short or three long fibers per penton. *J Virol* **88**, 11304-11314.

Rapoza, N. P. (1967). A classification of simian adenoviruses based on hemagglutination. *Am J Epidemiol* **86**, 736-745.

Roy, S., Clawson, D. S., Adam, V. S., Medina, A. & Wilson, J. M. (2011). Construction of gene transfer vectors based on simian adenovirus 7. *J Gen Virol* **92**, 1749-1753.

Roy, S., Sandhu, A., Medina, A., Clawson, D. S. & Wilson, J. M. (2012). Adenoviruses in fecal samples from asymptomatic rhesus macaques, United States. *Emerg Infect Dis* **18**, 1081-1088.

Roy, S., Vandenberghe, L.H., Kryazhimskiy, S., Grant, R., Calcedo, R., Yuan, X., Keough, M., Sandhu, A., Wang, Q., Medina-Jaszek, C.A., Plotkin, J.B. & Wilson, J.M. (2009). Isolation and characterization of adenoviruses persistently shed from the gastrointestinal tract of non-human primates. *PLoS Pathog* **5**, e1000503.

Tarján, Z. L., Pénzes, J. J., Tóth, R. P. & Benkő, M. (2014). First detection of circovirus-like sequences in amphibians and novel putative circoviruses in fishes. *Acta Vet Hung* **62**, 134-144.

To, K. K., Tse, H., Chan, W. M., Choi, G. K., Zhang, A. J., Sridhar, S., Wong, S. C., Chan, J. F., Chan, A. S., Woo, P. C., Lau, S. K., Lo, J. Y., Chan, K. H., Cheng, V. C. & Yuen, K. Y. (2014). A novel psittacine adenovirus identified during an outbreak of avian chlamydiosis and human psittacosis: zoonosis associated with virus-bacterium coinfection in birds. *PLoS Negl Trop Dis* **8**, e3318.

Tollefson, A. E., Ying, B., Doronin, K., Sidor, P. D. & Wold, W. S. (2007). Identification of a new human adenovirus protein encoded by a novel late l-strand transcription unit. *J Virol* **81**, 12918-12926.

Vidovszky, M., Kohl, C., Boldogh, S., Görföl, T., Wibbelt, G., Kurth, A. & Harrach, B. (2015). Random sampling of the Central European bat fauna reveals the existence of numerous hitherto unknown adenoviruses. *Acta Vet Hung* **63**, 508-525.

Walsh, M. P., Seto, J., Liu, E. B., Dehghan, S., Hudson, N. R., Lukashev, A. N., Ivanova, O., Chodosh, J., Dyer, D. W., Jones, M. S. & Seto, D. (2011). Computational analysis of two species C human adenoviruses provides evidence of a novel virus. *J Clin Microbiol* **49**, 3482-3490.

Wellehan, J. F., Johnson, A. J., Harrach, B., Benkő, M., Pessier, A. P., Johnson, C. M., Garner, M. M., Childress, A. & Jacobson, E. R. (2004). Detection and analysis of six lizard adenoviruses by consensus primer PCR provides further evidence of a reptilian origin for the atadenoviruses. *J Virol* **78**, 13366-13369.

Zhao, J., Zhong, Q., Zhao, Y., Hu, Y. X. & Zhang, G. Z. (2015). Pathogenicity and complete genome characterization of fowl adenoviruses isolated from chickens associated with Inclusion body hepatitis and hydropericardium syndrome in China. *PLoS One* **10**, e0133073.

Figure legends

Fig. 1. Phylogenetic tree based on the full aa sequence of the (a) DNA-dependent DNA polymerase, (b) the penton base, (c) the fibre-1, and (d) fibre-2. Black arrow indicates the node which separates the group of AdVs with two fibre genes (except SAdV-18). Virus associated RNA (VA RNA) and E3 19K labels mark the nodes after which either one or two copies of these genes appeared during the evolution. Abbreviations: HAdV - human AdV; BaAdV - baboon AdV; TSAdV - tree shrew AdV; TMAdV - titi monkey AdV; FAdV - fowl AdV.

Fig. 2. Phylogenetic tree based on full hexon aa sequences. Black arrow indicates the node which separates the group of AdVs with two fibre genes (except SAdV-18). Virus associated RNA (VA RNA) and E3 19K labels mark the nodes after which either one or two copies of these genes appeared during the evolution. Abbreviations: see at Fig. 1.

Fig. 3. (a) SimPlot and (b) BootScan analyses of *Simian mastadenovirus C*, *Human mastadenovirus G* and *Human mastadenovirus F* members relative to SAdV-19. The annotated

genome of SAdV-19 is between the two graphs to allow easier observation of the genomic locus where the putative recombination might have taken place. Black arrows indicate the assumed possible recombination spots in the genome, i.e. in the hexon and fibre gene regions (genes highlighted in red in the annotated genome).

Fig. 4. Alignment of the aa sequences of the so-called U exon protein (UXP) identified in four HAdVs previously and in seven SAdV types in this study.

Fig. 5. Genomic position and splicing pattern of the UXP gene of SAdVs, identified in this study, in comparison to those in HAdV-5. The reading frame of the DNA-binding protein (DBP) gene, overlapping the third exon, is also shown.