Genomic characterization of human adenovirus type 4 strains isolated worldwide since 1953 identifies two separable phylogroups evolving at different rates from their most recent common ancestor

Gabriel Gonzalez[1†*], Camden R. Bair[2†], Daryl M. Lamson[3], Hidemi Watanabe[4], Laura Panto[4], Michael J. Carr[5,6], and Adriana E. Kajon[2*]

1. Research Center for Zoonosis Control, Hokkaido University, Sapporo, Japan
2. Infectious Disease Program, Lovelace Respiratory Research Institute, New Mexico, USA
3. Wadsworth Center, New York State Department of Health, New York, USA
4. Graduate School of Information Science and Technology, Hokkaido University, Japan
5. Global Institution for Collaborative Research and Education (GI-CoRE), Hokkaido University, Japan
6. National Virus Reference Laboratory, School of Medicine, University College Dublin, Ireland

[†]These authors contributed equally to this work

*Corresponding authors:

Gabriel Gonzalez, Research Center for Zoonosis Control, Hokkaido University, Sapporo, Japan; Tel: +81 11 706 9504; Fax: +81 11 706 9491

email: gagonzalez@czc.hokudai.ac.jp

Adriana Kajon, Infectious Disease Program, Lovelace Respiratory Research Institute, Albuquerque, New Mexico, USA; Tel: +1(505) 348-9159, Fax: +1(505) 348-8567, email: akajon@lrri.org

Word Count of the Abstract: 143

Word Count of the Main Text: 5208

Number of Tables and Figures: 2 Tables and 6 Figures

Number of Supplementary Tables and Figures: 1 Table and 3 Figures

Key words: adenovirus type 4, human mastadenovirus E, genetic diversity, evolution

Abbreviations: HAdV, human adenovirus; HAdV-E, *Human mastadenovirus E*; HAdV-E4, Human adenovirus type 4; SAdV, simian adenovirus; NHP, non-human primates; RFLP, restriction fragment length polymorphism; NGS, next-generation sequencing; WGS, whole genome sequences; CDS, coding sequences; ORF, open reading frames; tMRCA, time to the most recent common ancestor; PG I, phylogroup I; PG II, phylogroup II; %G+C, percentage of genomic guanine-cytosine; 95% HPD, 95% highest posterior density range; ITR, inverted terminal repeats; E3, early region 3; E1A, early region 1A; E1B, early region 1B; VA RNA, virus-associated RNA; L4, late region 4;

Declaration of interests: none

Abstract

Species *Human mastadenovirus E* (HAdV-E) comprises several simian types and a single human type: HAdV-E4, a respiratory and ocular pathogen. RFLP analysis for the characterization of intratypic genetic variability has previously distinguished two HAdV-E4 clusters: prototype (p)-like and a-like. Our analysis of whole genome sequences confirmed two distinct lineages, which we refer to as phylogroups (PGs). PGs I and II comprise the p- and a-like genomes, respectively, and differ significantly in their G+C content (57.7% ± 0.013 vs 56.3% ± 0.015). Sequence differences distinguishing the two clades map to several regions of the genome including E3 and ITR. Bayesian analyses showed that the two phylogroups diverged approximately 602 years before the present. A relatively faster evolutionary rate was identified for PG II. Our data provide a rationale for the incorporation of phylogroup identity to HAdV-E4 strain designation to reflect the identified unique genetic characteristics that distinguish PGs I and II.

1. Introduction

The more than 89 currently recognized human adenovirus (HAdV) genotypes (1, 2) are categorized into seven species designated *Human mastadenovirus A* to *G* (HAdV-A to HAdV-G) based on their genetic characteristics (3). The number of constituent types varies among the seven species from a single one in species HAdV-E and -G, to more than 50 in species HAdV-D (2, 4-6).

Human adenovirus type 4 (HAdV-E4) is the only type in species *Human mastadenovirus E* (HAdV-E) thus far isolated from humans. Species HAdV-E also comprises several simian adenovirus types isolated from non-human primates (NHP), SAdV-21 through -26, SAdV-30, SAdV-36 through -39, and CHAdV Y25, suggesting that the emergence of HAdV-E4 as a human pathogen was the result of a zoonotic event or of an interspecies recombination process involving adenoviruses of two or more taxonomic species (7, 8). Among all SAdVs in species HAdV-E, SAdV-26 is the most closely related to HAdV-E4 (7). In humans, HAdV-E4 infection is associated with acute respiratory disease of variable severity affecting both military recruits in basic training and civilians in various settings (9-15) and with conjunctivitis of variable clinical manifestations, including epidemic keratoconjunctivitis, pharyngoconjunctival fever and hemorrhagic conjunctivitis (16-20).

Extensive intratypic genetic variability manifested by the occurrence of multiple genomic variants discriminable by restriction fragment length polymorphism (RFLP) analysis of the viral genome has been reported for HAdV-E4 since the late 1980s (14, 21, 22). By determining the percentage of comigrating restriction fragments, two major clusters of genetic homology were recognized among described genomic variants of HAdV-E4: a genomic cluster comprised of prototype (p)-like strains closely related to the prototype strain RI-67, and a second genomic cluster comprised of a-like strains (21).

The advent of next-generation sequencing (NGS) technologies has greatly facilitated the detailed characterization of complete HAdV genomes and their comparison. In 2005, the genome of the prototype strain of HAdV-E4, RI-67 was sequenced and annotated (23). Later, the first genomic comparisons between HAdV-E4 strains were reported together with the initial observations of possible interspecies recombination events underlying the evolution of this unique HAdV type (7).

In the present study, we obtained whole genome sequences (WGS) for a collection of 15 new HAdV-E4 strains isolated from cases of respiratory and ocular disease in

the United States and Japan, to assemble a large sample representing the spectrum of genetic diversity identified for this HAdV type. Using a global dataset of WGS from strains isolated between 1953 and 2015, we have conducted a comprehensive computational analysis of their evolutionary relationships and rates of divergence over time.

2. Material and Methods

2.1 Viral strains and next generation whole genome sequencing

Whole genome sequences were determined by NGS for 15 new HAdV-E4 strains isolated from cases of acute respiratory or ocular disease. The new sequences were generated at Hokkaido University using an Ion Torrent platform similar to that described in other studies (2), and at the Wadsworth Center, New York State Department of Health using an Illumina MiSeq platform, as previously described (13).

Additional WGS were gathered from the NCBI GenBank (Table 1). The genome type (p- or a-like) of each strain was determined (or verified) *in silico* with CLC Genomics Workbench software (v10, QIAGEN, Aarhus, Denmark) for enzymes BamHI, SmaI, SspI and XhoI. In addition, to re-examine the putative recombinant origins of HAdV-E4, WGS of simian adenoviruses (SAdVs) classified within species HAdV-E, SAdV-23 (AY530877), -24 (AY530878), -25 (AF394196 and FJ025918), -26 (FJ025923), -30 (FJ025920), -36 (FJ025917), -37 (FJ025921 and FJ025919), -38 (FJ025922), -39 (FJ025924) and chimpanzee adenovirus Y25 in HAdV-E (CHAdV-E25) (JN254802), as well as WGS for HAdV-B3 (DQ086466), -B7 (KF268134), -B11 (AF532578), -B14 (FJ822614), -B16 (JN860680), -B34 (AY737797), -B35 (AY271307), -B55 (FJ643676), -B66 (JN860676), -B68 (JN860678) and -B79 (LC177352) were included in the analysis. WGS and individual coding sequences (CDS) of 36 open reading frames (ORF) were multiple-sequence aligned with MAFFT (24). The nucleotide content was assessed as the percentage of guanine and cytosine (%G+C) calculated with CLC Genomics Workbench software. The %G+C for the first, second and third codon positions was estimated for each HAdV-4 sequence. The mean %G+C in species HAdV-B, -C and -D was estimated using at least one sequence for each genotype for each of these three species. The corresponding GenBank accession numbers are those reported in cited references 4-6.

2.2 Sequence alignment and analysis

Phylogenetic trees were inferred with MrBayes v3.2.7 (25) using the general time-reversible substitution model with heterogeneity among sites, modeled under a gamma distribution and allowing for a proportion of invariable sites (GTR+Γ+I) as substitution model, chosen as the model with the highest corrected Akaike information criterion (AICc) calculated with jModelTest 2 v0.1.10 (26) for the multiple sequence alignments. Trees were inferred with chain lengths of 10^6 states to assure convergence. Multiple sequence alignments comparisons were performed with Clustal X (27).

2.3 Similarity analysis, topological testing and G+C content

Similarity analyses between groups of sequences were performed by a sliding window approach with window size 500 bp and step size 250 bp. In each window the average evolutionary distance with Kimura model between groups was calculated as the mean of the distances between sequences in both groups. In addition, the averaged %G+C difference among HAdV-4 sequences was estimated for each window.

Topological testing of the cluster of sequences was performed by comparing the Bayes factor for the likelihood of one model considering the clustering and the model with the null hypothesis without that clustering (28). The topological model with the highest Bayes factor and a factor > 5 difference to other models was considered as the model with the highest support. For each window, the evolutionary models were tested with jModelTest 2 to assure the GTR+I+G model was among the best fitting models.

The percent similarities among nucleotide and protein sequences were estimated by using the Sequence Demarcation Tool software v1.2 (29).

Mean percent sequence similarities between two genotypes in the same species in HAdV-B 1, -B 2, -C and -D, were estimated by pairwise comparisons of sequences for all recognized genotypes in each of these three species using the Sequence Demarcation Tool software v1.2. The GenBank accession numbers for the analyzed sequences are those reported in cited references 4-6.

Simplot, a sequence similarity plotting tool (30), was used to identify conserved and divergent regions along the genomes of the examined HAdV-E4 strains. WGS were aligned with MAFFT v7.388 and default parameters using the Geneious 11.1.4 software platform (Biomatters, New Zealand). A similarity plot was generated in

Simplot V3.5.1 using a 200-nucleotide sliding window, a 20-nucleotide step size, GapStrip: On, Kimura distance model, and Ts/Tv=2.0.

2.4 Bayesian estimation of the time to the most recent common ancestors

To test whether the HAdV-E4 dataset provided enough data to analyze the temporal signal, the clock-likeness was checked by performing a linear regression between the parameters 'root-to-tip divergence' and 'sampling date' with TempEst (31). Time to the most recent common ancestor (tMRCA) was estimated by independent Bayesian Markov Chain Monte Carlo (MCMC) coalescent analyses by BEAST v2.4.6 (32) with chain lengths of 5×10^7 to ensure effective sample size (ESS) > 300 in all parameters of the models. Analyses were performed separately for WGS in the two groups of strains identified as p- and a-like genomes. Additionally, the tMRCA for both groups was estimated by analyzing a combination of CDS alignments excluding those suspected to contain effects of recombination events. Strict and relaxed exponential clock models were considered for the datasets in combination with coalescent constant, exponential and Bayesian skyline models for the populations (33, 34). The marginal likelihood of the combination of models and data was estimated in BEAST and with the Path-Sampler application in the BEAST package. Additionally, the distribution of the mutation rate for clades was calculated by extracting the parameter of each tree sampled every 5×10^4 states in the BEAST chain using TreeStat v1.2 (http://tree.bio.ed.ac.uk/software/treestat/). The values extracted from trees sampled along the BEAST chain were used to model the distribution of the parameter.

2.5 Statistical analyses

Statistical assessments were performed in R v3.5 (35). The statistical significance of %G+C differences among groups of sequences was assessed with phylogenetic independent contrasts (PIC) to correct for the shared ancestry among sequences before analyzing the correlation with the assigned phylogroup. In addition, parameters such as %G+C and percent sequence identity are reported as mean values, respectively, followed by the standard deviation.

3. Results and Discussion

3.1. HAdV-E4 genomic variants cluster into two separable phylogroups

WGS of HAdV-E4 strains isolated in the United States and Japan (n=15) were combined with prior publicly available sequences (n=32) to compile and align a total of 47 genomic sequences of 45 HAdV-E4 strains representing a diversity of genomic variants (Fig. 1), geographical locations and year of specimen collection (Table 1). The original genome typing data (p- or a-like) were confirmed by *in silico* RFLP analysis using recognition sequences for the restriction endonucleases BamHI, SmaI, SspI and XhoI, with 7, 19-20, 4 and 9 cleavage sites for p-like and 7-8, 13-15, 5-6 and 8-10 cleavage sites for a-like strains (Table 1 and Supplementary Fig. 1). The phylogenetic tree of WGS, including those of SAdVs in HAdV-E and HAdV genotypes classified within species HAdV-B (Fig. 1), showed two major clades of HAdV-E4 strains consistent with the original genogrouping described by Li and Wadell based on the analysis of percentage of comigrating restriction fragments (14, 21, 22). HAdV-E was rooted by a cluster containing all considered SAdV genomes. This phylogenetic position supported the previously formulated hypothesis of a zoonotic origin for HAdV-E4 (7, 8). Based on the highly supported phylogenetic distinction, these clades are hereafter referred to as phylogroup I (PG I) and phylogroup II (PG II) for HAdV-E4 p- and a-like strains, respectively.

The results of our phylogenetic analysis prompted us to examine further the genetic divergence. In addition to distinct digestion profiles with various restriction endonucleases (Supplementary Fig. 1), the two phylogroups of genomic variants also differed in their mean %G+C. Genomes in PG I showed a mean %G+C of 57.7% ± 0.013 while genomes in PG II showed a significantly lower mean %G+C of 56.3% ± 0.015 ($P < 2.2 \times 10^{-16}$, after correction applying PIC) (Table 1). The difference in %G+C between both phylogroups is also reflected in the nucleotide content for 1st, 2nd and 3rd codon positions, where the average %G+C content for each position in PG I is 58.8% ± 0.03, 44.5% ± 0.01 and 73.1% ± 0.02, respectively. The corresponding values in PG II are 58.3% ± 0.02, 44.1% ± 0.01 and 70.6% ± 0.05, respectively. Although more pronounced in the 3rd codon position, the average %G+C for the three codon positions was significantly higher in PG I than in PG II ($P < 2.2 \times 10^{-16}$, after correction applying PIC). Also, the distribution along the genome of such a difference in %G+C content was assessed with a sliding window approach (Fig. 2B). Such assessment reflected an uneven but widespread difference in the G+C content. The evolutionary significance of the differences in mean %G+C between the phylogroups is expected to be tested as WGS for more HAdV-E4 strains become available. Nevertheless, the

striking difference highlights the absence of intermediary PG strains, which may be attributable to a founder effect, a fitness cost for recombinants and/or insufficient sampling. It is noteworthy that the %G+C among types in other species is: 51.20% ± 0.09 in HAdV-B 1, 49.20% ± 0.83 in HAdV-B 2, 55.25% ± 0.06 in HAdV-C and 56.89% ± 0.54 in HAdV-D.

3.2. Evolutionary divergence between both phylogroups

The branching of two distinct lineages is attributable to the accumulation of mutations along the genome and/or to recombination events occurring over time. The average inter-phylogroup evolutionary distance was 0.0413 ± 0.0002 mutations/site. To assess the distribution of this divergence, we compared the average evolutionary distance among sequences within both phylogroups to other clusters of sequences in a sliding window approach with 500 bp window and 250 bp step size (Figs. 2C and 2D). These clusters included subspecies HAdV-B 1 and HAdV-B 2 sequences, specifically type HAdV-B16 previously suggested to be related to HAdV-E (7), and sequences of SAdVs in species HAdV-E. Furthermore, the topological hypotheses of PG I clustering with PG II or with SAdVs in species HAdV-E, and PG II with SAdVs in species HAdV-E or with subspecies HAdV-B 2, were tested by a Bayesian approach (Fig. 2E). The topological testing for 60% of the genomic windows (93/154) showed high support for the clustering of PG I with PG II (Supplementary Table 1), as suggested by the complete genome phylogeny (Fig. 1). On the other hand, 40% of the genomic windows (61/154) supported the clustering of one of the phylogroups with SAdVs in HAdV-E more strongly than with the other phylogroup, suggesting possible recombination events with NHP adenoviruses, as has been proposed (7).

The 60% of windows showing support for the topological clustering of both phylogroups, prompted us to estimate the time to the most recent common ancestor (tMRCA) for these windows following a Bayesian approach with BEAST and calibrating the tree with the isolation year of each strain due to the lack of other temporal data. Similar approaches have been followed for other viruses as fossil data are not available (36, 37).The temporal structure was tested to assure the divergence time could be estimated in datasets including a) SAdV + HAdV-E, b) HAdV-E, c) PG I and d) PG II (Supplementary Figure 2); the results supported strong time structure in HAdV-E (R^2 = 0.68, $P < 10^{-12}$), PG I (R^2 = 0.86, $P < 10^{-6}$) and PG II (R^2 = 0.59, $P < 10^{-7}$). The dataset including SAdV sequences was less conformant with a linear

regression (R^2 = 0.11, $P < 10^{-2}$), possibly as a consequence of uncharacterized recombination events in SAdV, and hampered the inclusion of this group for the divergence analysis. Nevertheless, these estimates are expected to be further refined as new genomic sequences become available. The Bayesian analysis under different combinations of molecular clocks and population models showed that the tMRCA of PG I is longer than that for PG II (Table 2). The models involving relaxed molecular clocks, which allow for different molecular clock rates along the branches modeled under exponential or log-normal distributions, were well supported. These models showed that both phylogroups diverged approximately 600 years before the present (ybp), established as 2015, the most recent calibration point (strains #12, 44 and 45, Table 1), or around year 1400 in the absolute time scale (Fig. 3). The relaxed molecular clocks also showed a slightly higher median clock rate for PG II. The comparison of the mutation rate distributions per phylogroup extracted from the sampled trees in the Bayesian MCMC supported the mutation rate of PG II (median rate: 3.69×10^{-5} mutations/site/year) was 13% significantly higher than the mutation rate of PG I (median rate: 3.24×10^{-5} mutations/site/year) ($P < 2 \times 10^{-115}$, Student's t-test). Furthermore, a similar trend was shown in the slopes of the linear regression for PG I and II (Supplementary Figure 2C-D). These mutation rates were comparable to previous estimates of mutation rates of 7.20×10^{-5} and 3.46×10^{-5} for HAdV-B and HAdV-C, respectively (36). Notably, these figures were two orders of magnitude greater than those expected for other double-stranded DNA viral DNA polymerases (38), and approximately four orders of magnitude greater than the mutation rate of primate hosts, thus providing a strong argument against the hypothesis of host-parasite co-speciation as the divergence between *Homo sapiens* and *Pan troglodytes*, which is estimated as 6.4 MYA (CI: 5.1 – 11.8 MYA) (derived from 79 studies in http://timetree.org/), would require an average adenoviral genome mutation rate of approximately 10^{-8} mutations/year/site.

The results of our analyses suggested that the currently circulating strains in PG I are descendants of an ancestral strain circulating ~91 ybp (~1924 in the absolute time scale) in the 95% highest posterior density range (95% HPD) [67, 144 ybp] while strains in PG II are descendants of an ancestral strain circulating ~54 ybp (~1961 in the absolute time scale) in the 95% HPD [40, 84 ybp].

Two independent descriptions of the 1965 Chinese strain BC129 as an "a-like" genomic variant (21, 22) date the detection of PG II to 13 years before 1978, the year

of detection of the oldest strain in the examined collection, V0014 (Table 1), lending support to the Bayesian estimations and suggesting that both PG I and PG II have been circulating for similar periods of time. As the genomic sequence for strain BC129 is not available we could not include 1965 as a calibration point in the analysis.

The significantly higher mutation rate in PG II under the different analysis models (Table 2) suggested: (i) a higher number of mutations accumulated during replication in PG II than in PG I, or (ii) a higher number of infections by PG II that increased the overall frequency of mutations despite a relatively similar mutation rate in each replicative cycle from both PGs. The number of samples in PG II and its frequent isolation across the world supported the second hypothesis and we hypothesize that this may be attributable to a higher viral fitness for PG II.

3.3. Detailed analysis of sequence diversity between phylogroups I and II

On average, PG I genomes were found to be ~94.5% identical to genomes in PG II. Interestingly, this level of genetic relatedness is comparable to that between any two of the currently recognized HAdV genotypes within a given species: 94% ± 5 for HAdV-B, 96% ± 1 for HAdV-C, and 94% ± 1 for HAdV-D. Many of these are also distinguishable as unique serotypes in neutralization assays.

We conducted an additional sequence identity analysis using Simplot to identify conserved and divergent regions along the genomes of PG I and PG II strains. A representative simplified plot including only 3 genomic variants from each phylogroup is shown in Figure 4.

The most striking differences between the genomes in PG I and PG II map to the inverted terminal repeats (ITRs) and the early region 3 (E3), with mutations with the potential to result in phenotypic differences relevant to pathogenesis are found in multiple genomic loci, including E1A, E1B, VARNA, L3 and L4, as described in detail below.

3.3.1 Inverted terminal repeat

Our analysis of 12 WGS in PG I and 33 WGS in PG II confirmed previous reports of differences in the ITR sequences between the two lineages of HAdV-E4 genomic variants including previously reported differences in length (7, 39). PG I genomes had an average ITR length of 113.8 bp ± 3.4 and an intragroup mean percent identity of 98.6% ± 1.7, whereas the average length for PG II genomes was 206.5 ± 2.7 and their

mean intragroup percent identity was 99.6% ± 0.3. As shown in Fig. 2D and in Supplementary Fig. 3, in this region of the genome PG I and the SAdVs in species HAdV-E cluster together while PG II clusters more closely with members of subspecies HAdV-B 2.

Downstream from the origin of DNA replication, the ITRs of most HAdV genomes encode binding motifs for the host cellular transcription factors NFI and NFIII, which are required for efficient genome replication (40-44). As reported previously, the canonical NFI binding motif encoded in the genomes of most HAdVs is notably absent in the genomes of HAdV-E4 PG I strains (7, 23, 45). The HAdV-E4 PG I ITR only encodes the NFIII binding motif while PG II ITRs carry the same NFIII motif and an NFI recognition sequence similar to that found in the genomes of members of species HAdV-B (7, 45).

In our analysis, we found the NFI and NFIII binding motifs for all examined PG II strains to be identical. A variant NFIII binding motif 5'-TATGTAAATAA-3' was identified in the genomes of PG I strains #9-11.

The terminal 8 bp section of the ITR among the examined HAdV-E4 strains was generally conserved (5'-CATCATCA-3'). PG I strain RI-67 (ATCC VR-4) had a divergent sequence (5'-CTATCTAT-3') as reported previously (7, 8, 23). A different sequence, 5'-CATCATCA-3', was reported by Hang and colleagues (46) for RI-67 ATCC VR-1572 (GenBank accession KX384949,) suggesting variation in RI-67 stocks among different repositories. Interestingly, we identified novel variant terminal sequences in PG I and PG II strains: 5'-ATAATATA-3' in strain #34; 5'-AATAATAT-3' in strains #3, 8, and 44; 5'-CAATAATA-3' in strains #12, 36, 37, 39, and 41-43; and 5'-GCATCATC-3' for strain #14. In addition, a 194 bp insertion by duplication of the neighboring genomic loci was found adjacent to the right hand ITR in strain NHRC22650 (#29 in Table 1).

Using vectors constructed from the HAdV-B35 background, Wunderlich and colleagues showed that the terminal ITR sequence can affect viral replication (47). The elucidation of the functional significance of the variation identified in this and other studies for the ITR region among PG I and PG II strains will require experimental evaluation using engineered mutant viruses.

3.3.2 Early region 1

Early region 1A (E1A)

The HAdV genome is predicted to encode two predominant E1A polypeptides resulting from alternative splicing. E1A is an important multifunctional protein that induces transition of the host cell into the S phase of the cell cycle (48), and is a potent transactivator of HAdV early gene expression (49, 50). Four conserved regions (CR1 to CR4) are found in the large E1A protein while the small E1A protein only includes CR1, CR2, and CR4 (51). The genomes of PG I encode a 28 kDa (257 aa) polypeptide and a 24.6 kDa (226 aa) polypeptide. The genomes of PG II strains encode slightly shorter polypeptides of 27 kDa (246 aa) and 23.5 kDa (215 aa), respectively. The predicted polypeptide sequences encoded by all examined genomes in PG I are identical, while PG II sequences had an average intragroup percent identity of 99.8% ± 0.3. Collectively, all examined strains had an inter-group percent identity of 96.1% ± 6.3.

The most striking difference between the E1A polypeptides encoded by the two phylogroups is an 11 amino acid deletion (aa 82-94) between the conserved regions CR1 and CR2 found in all PG II genomes examined in this study. The deletion includes a leucine at position 91 and a threonine at position 93 which Avvakumov and colleagues described as highly conserved residues in the E1A proteins encoded by members of species HAdV-B, -D, and -E (51). While no functional role has been assigned to these amino acids, this region has been shown to be a flexible linker between CR1 and CR2. This linker is important in the formation of stable ternary complexes between E1A, RB and CBP/p300 (52). The minimum length of the linker required for functionality is not known. However, removal of the linker in HAdV-C5 was shown to result in failure to induce colony formation in infected BRK cells (53). Importantly, compared to other published E1A sequences (51), the linker in the E1A polypeptide encoded by PG II strains ranks among the shortest linker. Interestingly, the 11 amino acid deletion described above was not present in any of the examined genomes of simian members of species HAdV-E.

Early region 1B (E1B)

The E1B transcriptional unit encodes two polypeptides, E1B 19K and E1B 55K, which are translated from two distinct initiation codons in different reading frames (54). Both proteins serve important functions in blocking p53-dependent induction of apoptosis through different mechanisms. In addition, E1B 55K in conjunction with E4 ORF6 has been shown to aid in the transport of viral mRNAs late in infection (55).

The E1B 19K protein plays a critical role in suppressing apoptosis induced by E1A and is regarded as the Bcl-2 homolog encoded by adenoviruses (reviewed in (56)). The predicted polypeptide sequence for E1B 19K is conserved among members of PG I as well as among members of PG II with a 97.6% ± 2.6 inter-phylogroup sequence identity. There are three non-synonymous mutations at positions 43, 100, and 125, using the sequence of RI-67 as a reference. Additionally, there is a 30 nucleotide (10 amino acid) in-frame insertion among PG II members located in the shared coding region for E1B 19K and E1B 55K. This insertion is also present in the coding sequence for E1B 19K in the genomes of several simian members of species HAdV-E (SAdV-23 to -26 and ChAdVY25), although only 5 of the 10 amino acids are conserved within PG II sequences.

The E1B 55K protein performs several functions critical for viral replication (57). The predicted sequences for E1B 55K encoded by all examined PG I strains were identical. The predicted sequences for E1B 55K encoded by PG II strains had an average sequence identity of 98.2% ± 1.9. All of them are characterized by a 10 amino acid insertion at their N-terminus resulting from the 30-nucleotide insertion described above.

3.3.3 Virus-associated RNAs

The genomes of all examined HAdV-E4 strains encode two virus-associated (VA) RNAs, designated VA RNA_I and VA RNA_{II}. The VA RNAs are non-coding RNAs, transcribed by RNA polymerase III that fold into highly structured RNAs resembling microRNA precursors. VA RNAs function as suppressors of RNAi by interfering with the activity of the endoribonuclease Dicer (58). While the function of VA RNA_I has been well characterized as a competitive substrate that binds the interferon-inducible double-stranded RNA-dependent protein kinase (PKR) (59), the role of VA RNA_{II} in the virus life cycle is still poorly understood. Consistent with the original observations reported by Kidd and colleagues (60), all PG II genomes examined in this study exhibit a 65 bp deletion in VA RNA_{II} starting at position 10593 (relative to the prototype strain RI-67) that partially ablates the promoter element A and results in the complete loss of promoter element B with a predicted lack of expression of VA RNA_{II}. The genomes of a subset of PG I strains (#8, 9, 10, and 11) have an additional 20 bp deletion starting at nucleotide position 10640, immediately downstream of promoter element B.

3.3.4 Late region 3

Our analysis focused exclusively on the hypervariable regions 1-7 of the hexon gene (HVR 1-7) which encode the serotype-specific residues displayed in loops 1 and 2 of the hexon capsid protein that projects from the surface of the virion (61). HVR-7 in loop 2 was recently shown to contain a conformational neutralization epitope (62).

We identified non-synonymous point mutations in HVR 1 and HVRs 3-7 that distinguish the genomes of PG I and II. The results of our sequence data analysis for this region are summarized in Figure 5.

Using a colorimetric neutralization assay and rabbit anti-HAdV-E4 strain RI-67 hyper-immune sera Crawford-Miksza and colleagues showed that strain Z-G with sequence characteristics of a-like genomes in PG II (#18 in our sample), exhibited a four-fold reduction in neutralization titer compared to that of the prototype strain RI-67 (63). A reduced cross-reactivity of HAdV-16 antisera with the Z-G strain compared to the prototype RI-67 was also observed, identifying this a-like strain as an antigenic variant. Taken together, these prior findings and our data showing conservation of HVR 1-7 sequence features among strains of PG II suggest that this clade may have also drifted antigenically.

3.3.5. Late region 4 100K

Late region 4 (L4) 100K is an abundantly expressed polypeptide necessary for efficient translation of late viral mRNAs (64, 65), trimerization and nuclear localization of the hexon polypeptide (66), and also to protect infected cells from granzyme B-dependent cell death by cytotoxic lymphocytes (67). The predicted L4-100K amino acid sequences encoded by PG I and PG II strains had an average percent identity of 98.4% ± 5.4. PG I strain #12 (13) exhibited a single glutamine insertion (nucleotides CAG) at amino acid position 20. Interestingly, several differences were identified between PG I and PG II sequences in the glycine-arginine rich (GAR) C-terminal region of L4-100K which is a site of arginine methylation (reviewed by (68, 69)). The GAR region contains an RGG domain (also referred to as RGG box) with three RGG peptide motifs that are conserved among various types in species HAdV-B, -C and -F as well as in PG I strain RI-67 (70). Mutations of these arginines in HAdV-C5 L4-100K interfere with protein interactions with late viral transcripts, possibly disrupt the role of L4-100K in hexon trimerization, and prevent shuttling of L4-100K to the nucleus, ultimately resulting in decreased viral replication (70, 71). Our analysis identified a 5 amino acid deletion (glycine-glycine-glycine-arginine-serine) in all PG II strains

between amino acid positions 755 and 761 (relative to RI-67) that disrupts the third RGG motif in the RGG domain. In addition, a single glycine insertion at position 743 (relative to RI-67) is encoded in the genomes of all examined strains of PG II creating an additional RGG tripeptide. The genomes of strains #18, 23, 24, 26, 28, 33, 35, 37, 38, and 40-43 also exhibit a glycine to glutamic acid replacement at position 717 that disrupts the consensus for an RGG tripeptide.

Further work is needed to elucidate whether the sequence differences detected at the C-terminus of L4-100K between PG I and PG II strains affect viral replication.

3.3.6 Early region 3

The early region 3 (E3) transcriptional unit comprises gene repertoires that vary considerably among HAdV species HAdV-A through -G. The conserved E3 genes encode non-structural modulators of host responses to infection (72-80). The variable repertoires of species-specific E3 genes are located between the highly conserved E3-gp19K and RIDα, and encode non-structural type I membrane glycoproteins expressed at early and late times post infection (81-83). In the same location as HAdV-C E3-11.6K encoding the adenovirus death protein (ADP) (84), the E3 region of members of species HAdV-E encodes 2 or 3 CR1 genes (85, 86) of unknown function. As shown schematically in Figure 6A, while the SAdVs in HAdV-E genomes encode 3 CR1 genes designated CR1β, CR1γ and CR1δ (7, 8, 23), CR1γ is absent in the genomes of both HAdV-E4 phylogroups. Interestingly, as originally reported (8), a vestigial E3 CR1γ sequence lacking an initiating ATG and splice acceptor site is present in the genomes of strains of PG I due to a 326 bp deletion relative to the SAdV-26 genome. We identified in PG II genomes a unique 318 bp deletion in the 5' region of the vestigial E3 CR1γ sequence. Additionally, while the splice acceptor sequence present in SAdV-26 is retained in these genomes, a mutation (ATG to ATA) ablates the start codon. Although unlikely to be expressed, a short 165 bp ORF evolutionarily unrelated to CR1γ annotated as E3-6.3K (77, 87) is present in this region in the genomes of PG I strains. As a result of a deletion introducing an early stop codon, the E3-6.3K ORF is significantly truncated in the genomes of PG II strains (data not shown).

Marked amino acid sequence differences resulting from point mutations as well as from insertions or deletions (indels) in the N-terminal ectodomains of E3 CR1β and E3 CR1δ (Figure 6B) distinguish the genomes of PG I and PG II and highlight the

divergence of these two ORFs from those encoded by SAdV-26. These genes have very low sequence similarity to any other genes in the NCBI database, thus making sequence-based prediction of biological activity and function challenging. Using an extracellular protein array, Martinez-Martin and colleagues recently showed the interaction of the ectodomain of E3-24.8K/CR1β encoded by PG I strain RI-67 with the inhibitory receptor LILRB1, suggesting a possible immunomodulatory function for this protein (88). No interactions were detected for E3-29.7K/CR1δ.

4. Conclusions and insights towards a more meaningful strain designation for HAdV-E4 based on genomics data

Our data from the computational analyses of 45 WGS strains of HAdV-E4 representing the spectrum of intratypic genetic variability described to date, indicate that the two phylogroups of HAdV-E4 have been circulating and evolving independently from a common ancestor, presumably a simian adenovirus as was suggested by the close genetic relationship to SAdVs in HAdV-E (89). The genomic differences between PG I and PG II identified in this study are strongly indicative of a genetic basis for probable differences in pathogenesis and fitness between the two separable evolutionary lineages. Data from molecular epidemiology studies of both respiratory and ocular disease associated with HAdV-E4 infection (13-15, 90, 91) show that both lineages have been in circulation over the last three decades with a noticeable predominance of PG II strains among examined clinical isolates. This supports the hypothesis of a potential selective advantage and/or of an increased virulence for this clade.

The shift in the last decade towards molecular diagnosis of viral infections and the growing capabilities for molecular typing of virus strains from original clinical specimens create an opportunity for the development of assays that could discriminate PGs I and II, thus overcoming the challenges posed by the costly and labor-intensive genome typing by *in silico* or gel-based RFLP. The International Committee on Taxonomy of Viruses (ICTV) provides no guidelines for the classification or designation of viruses beyond the species level. The use of a designation that could reflect both the unique genetic and associated phenotypic characteristics of any given HAdV-E4 strain would be extremely informative for epidemiological and functional studies of HAdV-E4 infection and associated disease.

We propose the use of the term phylogroups I and II in the designation of HAdV-E4 strains when molecular typing data are available. The basic designation HAdV-E4 PG I or HAdV-E4 PG II will better reflect the distinct genomic characteristics identified in the present work and those reported in other studies.

The implications of the identified genetic differences between phylogroups for viral pathogenesis and fitness, and the value of phylogrouping when typing clinical isolates of HAdV-E4 merits further investigation. Moving forward, the following *in vitro* and *in vivo* phenotypes should be considered for comparison between phylogroups: serological reactivities, replication and viral progeny release kinetics, plaque size, proinflammatory responses induced by infection in cell culture, and pulmonary inflammation in rodent models of HAdV respiratory infection.

Acknowledgements

The authors wish to thank Professors Koki Aoki and Nobuyoshi Kitaichi for their support and encouragement of this collaborative effort. C.R.B. is supported by the University of New Mexico Infectious Diseases and Inflammation NIH Training Grant T32-AI007538.

REFERENCES

1. Dhingra A., Hage E., Ganzenmueller T., Bottcher S., Hofmann J., Hamprecht K., Obermeier P., Rath B., Hausmann F., Dobner T., Heim A. 2019. Molecular Evolution of Human Adenovirus (HAdV) Species C. Sci Rep 9:1039. http://doi.org/c7cw
2. Kajan G.L., Lipiec A., Bartha D., Allard A., Arnberg N. 2018. A multigene typing system for human adenoviruses reveals a new genotype in a collection of Swedish clinical isolates. PLoS One 13:e0209038. http://doi.org/c6sh
3. Benko M., Harrach B., Both G., Russell W., Adair B., Adam E., De Jong J., Hess M., Johnson M., Kajon A. 2005. Adenoviridae. Virus Taxonomy Eighth Report of the International Committee on the Taxonomy of Viruses (Fauquet, CM, Mayo, MA, Maniloff, J, Desselberger U and Ball, LA eds), San Diego, CA: Elsevier Academic press:213-228.
4. 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 49:3482-90. https://doi.org/10.1128/JCM.00156-11
5. Hashimoto S., Gonzalez G., Harada S., Oosako H., Hanaoka N., Hinokuma R., Fujimoto T. 2018. Recombinant type human mastadenovirus D85 associated with epidemic keratoconjunctivitis since 2015 in Japan. J Med Virol 90:881-9. https://doi.org/10.1002/jmv.25041
6. Yoshitomi H., Sera N., Gonzalez G., Hanaoka N., Fujimoto T. 2017. First isolation of a new type of human adenovirus (genotype 79), species Human mastadenovirus B (B2) from sewage water in Japan. J Med Virol 89:1192-1200. https://doi.org/10.1002/jmv.24749
7. Dehghan S., Seto J., Liu E.B., Walsh M.P., Dyer D.W., Chodosh J., Seto D. 2013. Computational analysis of four human adenovirus type 4 genomes reveals molecular evolution through two interspecies recombination events. Virology 443:197-207. http://doi.org/f46pzg
8. Jacobs S.C., Davison A.J., Carr S., Bennett A.M., Phillpotts R., Wilkinson G.W. 2004. Characterization and manipulation of the human adenovirus 4 genome. J Gen Virol 85:3361-6. https://doi.org/10.1099/vir.0.80386-0

9. Mölsa M., Hemmila H., Ronkko E., Virkki M., Nikkari S., Ziegler T. 2016. Molecular characterization of adenoviruses among finnish military conscripts. J Med Virol 88:571-577. https://doi.org/10.1002/jmv.24364
10. Kandel R., Srinivasan A., D'Agata E.M.C., Lu X.Y., Erdman D., Jhung M. 2010. Outbreak of Adenovirus Type 4 Infection in a Long-Term Care Facility for the Elderly. Infect Control Hosp Epidemiol 31:755-757. https://doi.org/10.1086/653612
11. Narra R., Bono P., Zoccoli A., Orlandi A., Piconi S., Grasselli G., Crotti S., Girello A., Piralla A., Baldanti F., Lunghi G. 2016. Acute respiratory distress syndrome in adenovirus type 4 pneumonia: A case report. J Clin Virol 81:78-81. https://doi.org/10.1016/j.jcv.2016.06.005
12. Kalimuddin S., Chan Y.F.Z., Wu I.Q., Tan Q.L., Murthee K.G., Tan B.H., Oon L.L.E., Yang Y., Lin R.T.P., Joseph U., Sessions O.M., Smith G.J.D., Ooi E.E., Low J.G.H. 2017. A Report of Adult Human Adenovirus Infections in a Tertiary Hospital. Open Forum Infect Dis 4. http://doi.org/c6sj
13. Kajon A.E., Lamson D.M., Bair C.R., Lu X.Y., Landry M.L., Menegus M., Erdman D.D., St George K. 2018. Adenovirus Type 4 Respiratory Infections among Civilian Adults, Northeastern United States, 2011-2015. Emerg Infect Dis 24:201-209. https://doi.org/10.3201/eid2402.171407
14. Kajon A.E., Moseley J.M., Metzgar D., Huong H.S., Wadleigh A., Ryan M.A.K., Russell K.L. 2007. Molecular epidemiology of adenovirus type 4 infections in US military recruits in the postvaccination era (1997-2003). J Infect Dis 196:67-75. https://doi.org/10.1086/518442
15. Rogers A.E., Lu X., Killerby M., Campbell E., Gallus L., Kamau E., Froh I.B., Nowak G., Erdman D.D., Sakthivel S.K., Gerber S.I., Schneider E., Watson J.T., Johnson L.A. 2019. Outbreak of Acute Respiratory Illness Associated with Adenovirus Type 4 at the U.S. Naval Academy, 2016. MSMR 26:21-27.
16. Aoki K., Kato M., Ohtsuka H., Ishii K., Nakazono N., Sawada H. 1982. Clinical and aetiological study of adenoviral conjunctivitis, with special reference to adenovirus types 4 and 19 infections. Br J Ophthalmol 66:776-80. http://dx.doi.org/10.1136/bjo.66.12.776
17. Muzzi A., Rocchi G., Lumbroso B., Tosato G., Barbieri F. 1975. Acute haemorrhagic conjunctivitis during an epidemic outbreak of adenovirus-type-4 injection. Lancet 2:822-3. http://doi.org/b4s3tn

18. Schepetiuk S.K., Norton R., Kok T., Irving L.G. 1993. Outbreak of adenovirus type 4 conjunctivitis in South Australia. J Med Virol 41:316-8. http://doi.org/cfbsjq
19. Tullo A.B., Higgins P.G. 1980. An Outbreak of Adenovirus Type-4 Conjunctivitis. Br J Ophthalmol 64:489-493. https://doi.org/10.1136/bjo.64.7.489
20. Tsuzuki-Wang L., Aoki K., Isobe K., Shiao S., Toba K., Kobayashi N., Noguchi Y., Ohno S. 1997. Genome analysis of adenovirus type 4 strains isolated from acute conjunctivitis in Japan. Jpn J Ophthalmol 41:308-311. http://doi.org/cpkcq6
21. Li Q.G., Wadell G. 1988. The degree of genetic variability among adenovirus type 4 strains isolated from man and chimpanzee. Arch Virol 101:65-77. https://doi.org/10.1007/BF01314652
22. Adrian T. 1992. Genome Type Analysis of Adenovirus Type-4. Intervirology 34:180-183. https://doi.org/10.1159/000150280
23. Purkayastha A., Ditty S.E., Su J., McGraw J., Hadfield T.L., Tibbetts C., Seto D. 2005. Genomic and bioinformatics analysis of HAdV-4, a human adenovirus causing acute respiratory disease: implications for gene therapy and vaccine vector development. J Virol 79:2559-72. http://doi.org/cszgtp
24. Katoh K., Standley D.M. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol 30:772-780. http://doi.org/f4qwp9
25. Ronquist F., Teslenko M., van der Mark P., Ayres D.L., Darling A., Hohna S., Larget B., Liu L., Suchard M.A., Huelsenbeck J.P. 2012. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Syst Biol 61:539-542. http://doi.org/f2zx9m
26. Darriba D., Taboada G.L., Doallo R., Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9:772-772. https://doi.org/10.1038/nmeth.2109
27. Larkin M.A., Blackshields G., Brown N.P., Chenna R., McGettigan P.A., McWilliam H., Valentin F., Wallace I.M., Wilm A., Lopez R., Thompson J.D., Gibson T.J., Higgins D.G. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23:2947-8. http://doi.org/fj5d5p
28. Sasaki M., Gonzalez G., Wada Y., Setiyono A., Handharyani E., Rahmadani I., Taha S., Adiani S., Latief M., Kholilullah Z.A. 2016. Divergent bufavirus

652 harboured in megabats represents a new lineage of parvoviruses. Sci Rep
653 6:24257. http://doi.org/f8njsv
654 29. Muhire B.M., Varsani A., Martin D.P. 2014. SDT: a virus classification tool
655 based on pairwise sequence alignment and identity calculation. PLoS One
656 9:e108277. https://doi.org/10.1371/journal.pone.0108277
657 30. Lole K.S., Bollinger R.C., Paranjape R.S., Gadkari D., Kulkarni S.S., Novak
658 N.G., Ingersoll R., Sheppard H.W., Ray S.C. 1999. Full-length human
659 immunodeficiency virus type 1 genomes from subtype C-infected
660 seroconverters in India, with evidence of intersubtype recombination. J Virol
661 73:152-160.
662 31. Rambaut A., Lam T.T., Carvalho L.M., Pybus O.G. 2016. Exploring the
663 temporal structure of heterochronous sequences using TempEst (formerly
664 Path-O-Gen). Virus Evol 2. http://doi.org/gfv5wt
665 32. Bouckaert R., Heled J., Kuhnert D., Vaughan T., Wu C.H., Xie D., Suchard M.A.,
666 Rambaut A., Drummond A.J. 2014. BEAST 2: a software platform for Bayesian
667 evolutionary analysis. PLoS Comput Biol 10:e1003537. http://doi.org/gfv5wx
668 33. Drummond A.J., Ho S.Y., Phillips M.J., Rambaut A. 2006. Relaxed
669 phylogenetics and dating with confidence. PLoS Biol 4:e88.
670 http://doi.org/dfpmbm
671 34. Drummond A.J., Rambaut A., Shapiro B., Pybus O.G. 2005. Bayesian
672 coalescent inference of past population dynamics from molecular sequences.
673 Mol Biol Evol 22:1185-92. http://doi.org/cx4hw9
674 35. Team R.C. 2013. R: A language and environment for statistical computing.
675 36. Firth C., Kitchen A., Shapiro B., Suchard M.A., Holmes E.C., Rambaut A. 2010.
676 Using time-structured data to estimate evolutionary rates of double-stranded
677 DNA viruses. Mol Biol Evol 27:2038-51. http://doi.org/dcqptt
678 37. Zhang Y., Vrancken B., Feng Y., Dellicour S., Yang Q., Yang W., Zhang Y.,
679 Dong L., Pybus O.G., Zhang H., Tian H. 2017. Cross-border spread, lineage
680 displacement and evolutionary rate estimation of rabies virus in Yunnan
681 Province, China. Virol J 14:102. http://doi.org/gbhb7h
682 38. Duffy S., Shackelton L.A., Holmes E.C. 2008. Rates of evolutionary change in
683 viruses: patterns and determinants. Nat Rev Genet 9:267-76.
684 https://doi.org/10.1038/nrg2323

685 39. Houng H.S., Clavio S., Graham K., Kuschner R., Sun W., Russell K.L., Binn
686 L.N. 2006. Emergence of a new human adenovirus type 4 (Ad4) genotype:
687 identification of a novel inverted terminal repeated (ITR) sequence from majority
688 of Ad4 isolates from US military recruits. J Clin Virol 35:381-7.
689 https://doi.org/10.1016/j.jcv.2005.11.008
690 40. Mul Y.M., Verrijzer C.P., van der Vliet P.C. 1990. Transcription factors NFI and
691 NFIII/oct-1 function independently, employing different mechanisms to enhance
692 adenovirus DNA replication. J Virol 64:5510-8.
693 41. Rosenfeld P.J., Oneill E.A., Wides R.J., Kelly T.J. 1987. Sequence-Specific
694 Interactions between Cellular DNA-Binding Proteins and the Adenovirus Origin
695 of DNA-Replication. Mol Cell Biol 7:875-886. http://doi.org/c7cx
696 42. Hatfield L., Hearing P. 1993. The Nfiii/Oct-1 Binding-Site Stimulates Adenovirus
697 DNA-Replication Invivo and Is Functionally Redundant with Adjacent
698 Sequences. J Virol 67:3931-3939.
699 43. Pruijn G.J.M., Vandriel W., Vandervliet P.C. 1986. Nuclear Factor-Iii, a Novel
700 Sequence-Specific DNA-Binding Protein from Hela-Cells Stimulating
701 Adenovirus DNA-Replication. Nature 322:656-659.
702 https://doi.org/10.1038/322656a0
703 44. Hay R.T. 1985. Origin of adenovirus DNA replication. Role of the nuclear factor
704 I binding site in vivo. J Mol Biol 186:129-36.
705 45. Zhang J., Kang J.E., Dehghan S., Sridhar S., Lau S.K.P., Ou J.X., Woo P.C.Y.,
706 Zhang Q.W., Seto D. 2019. A Survey of Recent Adenoviral Respiratory
707 Pathogens in Hong Kong Reveals Emergent and Recombinant Human
708 Adenovirus Type 4 (HAdV-E4) Circulating in Civilian Populations. Viruses-
709 Basel 11. https://doi.org/10.3390/v11020129
710 46. Hang J., Vento T.J., Norby E.A., Jarman R.G., Keiser P.B., Kuschner R.A., Binn
711 L.N. 2017. Adenovirus type 4 respiratory infections with a concurrent outbreak
712 of coxsackievirus A21 among United States Army Basic Trainees, a
713 retrospective viral etiology study using next-generation sequencing. J Med Virol
714 89:1387-1394. https://doi.org/10.1002/jmv.24792
715 47. Wunderlich K., van der Helm E., Spek D., Vermeulen M., Gecgel A., Pau M.G.,
716 Vellinga J., Custers J. 2014. An alternative to the adenovirus inverted terminal
717 repeat sequence increases the viral genome replication rate and provides a

selective advantage in vitro. J Gen Virol 95:1574-84. https://doi.org/10.1099/vir.0.064840-0

48. Berk A.J. 2005. Recent lessons in gene expression, cell cycle control, and cell biology from adenovirus. Oncogene 24:7673-85. https://doi.org/10.1038/sj.onc.1209040

49. Winberg G., Shenk T. 1984. Dissection of Overlapping Functions within the Adenovirus Type-5 E1a Gene. Embo Journal 3:1907-1912. http://doi.org/c7rk

50. Montell C., Courtois G., Eng C., Berk A. 1984. Complete Transformation by Adenovirus-2 Requires Both E1a-Proteins. Cell 36:951-961. http://doi.org/b2r5nt

51. Avvakumov N., Kajon A.E., Hoeben R.C., Mymryk J.S. 2004. Comprehensive sequence analysis of the E1A proteins of human and simian adenoviruses. Virology 329:477-492. https://doi.org/10.1016/j.virol.2004.08.007

52. Ferreon J.C., Martinez-Yamout M.A., Dyson H.J., Wright P.E. 2009. Structural basis for subversion of cellular control mechanisms by the adenoviral E1A oncoprotein. Proc Natl Acad Sci U S A 106:13260-5. http://doi.org/bh5ttx

53. Wang H.G.H., Moran E., Yaciuk P. 1995. E1a Promotes Association between P300 and Prb in Multimeric Complexes Required for Normal Biological-Activity. Journal of Virology 69:7917-7924.

54. Bos J.L., Polder L.J., Bernards R., Schrier P.I., van den Elsen P.J., van der Eb A.J., van Ormondt H. 1981. The 2.2 kb E1b mRNA of human Ad12 and Ad5 codes for two tumor antigens starting at different AUG triplets. Cell 27:121-31.

55. Bridge E., Ketner G. 1990. Interaction of adenoviral E4 and E1b products in late gene expression. Virology 174:345-53.

56. Cuconati A., White E. 2002. Viral homologs of BCL-2: role of apoptosis in the regulation of virus infection. Genes Dev 16:2465-78. https://doi.org/10.1101/gad.1012702

57. Blackford A.N., Grand R.J.A. 2009. Adenovirus E1B 55-Kilodalton Protein: Multiple Roles in Viral Infection and Cell Transformation. J Virol 83:4000-4012. https://doi.org/10.1128/Jvi.02417-08

58. Andersson M.G., Haasnoot P.C., Xu N., Berenjian S., Berkhout B., Akusjarvi G. 2005. Suppression of RNA interference by adenovirus virus-associated RNA. J Virol 79:9556-65. http://doi.org/cbz4pg

751 59. Vachon V.K., Conn G.L. 2016. Adenovirus VA RNA: An essential pro-viral non-
752 coding RNA. Virus Res 212:39-52. http://doi.org/f8b3st
753 60. Kidd A.H., Garwicz D., Oberg M. 1995. Human and simian adenoviruses:
754 phylogenetic inferences from analysis of VA RNA genes. Virology 207:32-45.
755 https://doi.org/10.1006/viro.1995.1049
756 61. Crawford-Miksza L.K., Schnurr D.P. 1996. Adenovirus serotype evolution is
757 driven by illegitimate recombination in the hypervariable regions of the hexon
758 protein. Virology 224:357-367. https://doi.org/10.1006/viro.1996.0543
759 62. Tian X., Qiu H., Zhou Z., Wang S., Fan Y., Li X., Chu R., Li H., Zhou R., Wang
760 H. 2018. Identification of a Critical and Conformational Neutralizing Epitope in
761 Human Adenovirus Type 4 Hexon. J Virol 92.
762 https://doi.org/10.1128/JVI.01643-17
763 63. Crawford-Miksza L.K., Nang R.N., Schnurr D.P. 1999. Strain variation in
764 adenovirus serotypes 4 and 7a causing acute respiratory disease. J Clin
765 Microbiol 37:1107-12.
766 64. Hayes B.W., Telling G.C., Myat M.M., Williams J.F., Flint S.J. 1990. The
767 Adenovirus L4 100-Kilodalton Protein Is Necessary for Efficient Translation of
768 Viral Late Messenger-Rna Species. J Virol 64:2732-2742.
769 65. Xi Q., Cuesta R., Schneider R.J. 2004. Tethering of eIF4G to adenoviral
770 mRNAs by viral 100k protein drives ribosome shunting. Genes Dev 18:1997-
771 2009. https://doi.org/10.1101/gad.1212504
772 66. Hong S.S., Szolajska E., Schoehn G., Franqueville L., Myhre S., Lindholm L.,
773 Ruigrok R.W., Boulanger P., Chroboczek J. 2005. The 100K-chaperone protein
774 from adenovirus serotype 2 (Subgroup C) assists in trimerization and nuclear
775 localization of hexons from subgroups C and B adenoviruses. J Mol Biol
776 352:125-38. http://doi.org/b7g36h
777 67. Andrade F., Bull H.G., Thornberry N.A., Ketner G.W., Casciola-Rosen L.A.,
778 Rosen A. 2001. Adenovirus L4-100K assembly protein is a granzyme B
779 substrate that potently inhibits granzyme B-mediated cell death. Immunity
780 14:751-761. http://doi.org/chqfcq
781 68. Thandapani P., O'Connor T.R., Bailey T.L., Richard S. 2013. Defining the
782 RGG/RG motif. Mol Cell 50:613-23. http://doi.org/f43gns
783 69. Blanc R.S., Richard S. 2017. Arginine Methylation: The Coming of Age. Mol
784 Cell 65:8-24. http://doi.org/c6sg

70. Iacovides D.C., O'Shea C.C., Oses-Prieto J., Burlingame A., McCormick F. 2007. Critical role for arginine methylation in adenovirus-infected cells. J Virol 81:13209-17. https://doi.org/10.1128/JVI.01415-06
71. Koyuncu O.O., Dobner T. 2009. Arginine Methylation of Human Adenovirus Type 5 L4 100-Kilodalton Protein Is Required for Efficient Virus Production. J Virol 83:4778-4790. https://doi.org/10.1128/Jvi.02493-08
72. Bhat B.M., Brady H.A., Pursley M.H., Wold W.S. 1986. Deletion mutants that alter differential RNA processing in the E3 complex transcription unit of adenovirus. J Mol Biol 190:543-57.
73. Gooding L.R., Wold W.S. 1990. Molecular mechanisms by which adenoviruses counteract antiviral immune defenses. Crit Rev Immunol 10:53-71.
74. Wold W.S., Hermiston T.W., Tollefson A.E. 1994. Adenovirus proteins that subvert host defenses. Trends Microbiol 2:437-43. http://doi.org/bzjnrd
75. Wold W.S., Tollefson A.E., Hermiston T.W. 1995. E3 transcription unit of adenovirus. Curr Top Microbiol Immunol 199 (Pt 1):237-74. http://doi.org/csp9rh
76. Burgert H.G., Blusch J.H. 2000. Immunomodulatory functions encoded by the E3 transcription unit of adenoviruses. Virus Genes 21:13-25. http://doi.org/dskb94
77. Burgert H.G., Ruzsics Z., Obermeier S., Hilgendorf A., Windheim M., Elsing A. 2002. Subversion of host defense mechanisms by adenoviruses. Curr Top Microbiol Immunol 269:273-318. http://doi.org/fqr6tg
78. Lichtenstein D.L., Toth K., Doronin K., Tollefson A.E., Wold W.S. 2004. Functions and mechanisms of action of the adenovirus E3 proteins. Int Rev Immunol 23:75-111. http://doi.org/cmg65j
79. Lichtenstein D.L., Doronin K., Toth K., Kuppuswamy M., Wold W.S.M., Tollefson A.E. 2004. Adenovirus E3-6.7K protein is required in conjunction with the E3-RID protein complex for the internalization and degradation of TRAIL receptor 2. J Virol 78:12297-12307. http://doi.org/dk5swx
80. Windheim M., Hilgendorf A., Burgert H.G. 2004. Immune evasion by adenovirus E3 proteins: exploitation of intracellular trafficking pathways. Curr Top Microbiol Immunol 273:29-85. http://doi.org/c6sp

81. Hawkins L.K., Wold W.S. 1995. The E3-20.5K membrane protein of subgroup B human adenoviruses contains O-linked and complex N-linked oligosaccharides. Virology 210:335-44. http://doi.org/fh5xhs
82. Frietze K.M., Campos S.K., Kajon A.E. 2010. Open Reading Frame E3-10.9K of Subspecies B1 Human Adenoviruses Encodes a Family of Late Orthologous Proteins That Vary in Their Predicted Structural Features and Subcellular Localization. J Virol 84:11310-11322. https://doi.org/10.1128/Jvi.00512-10
83. Robinson C.M., Rajaiya J., Zhou X.H., Singh G., Dyer D.W., Chodosh J. 2011. The E3 CR1-gamma gene in human adenoviruses associated with epidemic keratoconjunctivitis. Virus Res 160:120-127. http://doi.org/c5cp5q
84. Tollefson A.E., Scaria A., Hermiston T., Ryerse J.S., Wold L.J., Wold W.S.M. 1996. The adenovirus death protein (E3-11.6K) is required at very late stages of infection for efficient cell lysis and release of adenovirus from infected cells. J Virol 70:2296-2306.
85. Davison A.J., Akter P., Cunningham C., Dolan A., Addison C., Dargan D.J., Hassan-Walker A.F., Emery V.C., Griffiths P.D., Wilkinson G.W. 2003. Homology between the human cytomegalovirus RL11 gene family and human adenovirus E3 genes. J Gen Virol 84:657-63. https://doi.org/10.1099/vir.0.18856-0
86. Davison A.J., Benko M., Harrach B. 2003. Genetic content and evolution of adenoviruses. J Gen Virol 84:2895-2908. https://doi.org/10.1099/vir.0.19497-0
87. Li Y., Wold W.S. 2000. Identification and characterization of a 30K protein (Ad4E3-30K) encoded by the E3 region of human adenovirus type 4. Virology 273:127-38. https://doi.org/10.1006/viro.2000.0384
88. Martinez-Martin N., Ramani S.R., Hackney J.A., Tom I., Wranik B.J., Chan M., Wu J., Paluch M.T., Takeda K., Hass P.E., Clark H., Gonzalez L.C. 2016. The extracellular interactome of the human adenovirus family reveals diverse strategies for immunomodulation. Nat Commun 7:11473. https://doi.org/10.1038/ncomms11473
89. 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. http://doi.org/b2sbv9

90. Cooper R.J., Bailey A.S., Killough R., Richmond S.J. 1993. Genome analysis of adenovirus 4 isolated over a six year period. J Med Virol 39:62-6. http://doi.org/b8hckj

91. Ren C.S., Nakazono N., Ishida S., Fujii S., Yoshii T., Yamazaki S., Ishii K., Fujinaga K. 1985. Genome Type Analysis of Adenovirus Type-4 Isolates, Recently Obtained from Clinically Different Syndromes in Some Areas in Japan. Jpn J Med Sci Biol 38:195-199. http://doi.org/c6sk

Table 1. Origin and genomic characteristics of HAdV-E4 strains included in the study

				Isolation					Genomic data		Restriction Fragment Length Polymorphisms (RFLP) * (# of sites)			
No.	**Strain ID**	**Phylogroup**	**Genome type**	**Year**	**Place**	**Specimen**	**Accession number****	**WGS source****	**Genome Length (bp)**	**G+C (%)**	**BamHI**	**SmaI**	**SspI**	**XhoI**
1	RI-67			1953	MO,USA	respiratory	AY594253	GenBank	35990	57.7	7	20	4	9
	prototype strain						KX384949	GenBank	35990	57.7	7	20	4	9
2	CL 68578			1963	NC,USA	respiratory	AY487947	GenBank	35994	57.7	7	20	4	9
	vaccine strain						AY594254	GenBank	35994	57.7	7	19	4	9
3	RU2533			1966	USA	respiratory	**MF002043**	**NYSDOH**	35975	57.7	7	20	4	9
4	RDU2954			1966	NJ,USA	respiratory	KX384948	GenBank	35991	57.7	7	20	4	9
5	RU4445			1968	EGP	respiratory	KX384947	GenBank	35991	57.7	7	20	4	9
6	RU7872	**I**	**p-like**	1971	MN,USA	respiratory	KX384950	GenBank	35983	57.7	7	20	4	9
7	V1003			1981	NY,USA	respiratory	KX384957	GenBank	35929	57.7	7	20	4	9
8	V2029E			1986	GA,USA	respiratory	KX384946	GenBank	35904	57.7	7	20	4	9
9	NHRC90255			2000	NJ,USA	respiratory	**AP014852**	**Hokkaido**	35914	57.7	7	20	4	9
10	NHRC90870			2004	NJ,USA	respiratory	**AP014853**	**Hokkaido**	35914	57.7	7	20	4	9
11	NHRC90339			2011	NJ,USA	respiratory	EF371058	GenBank	35914	57.7	7	20	4	9
12	NYS 15-4054			2015	NY,USA	respiratory	KY996447	GenBank	35968	57.7	7	20	4	9
13	V0014			1978	FRA	respiratory	KX384956	GenBank	35960	56.4	8	15	5	10
14	J1007			1981	JPN	respiratory	**KY996452**	**NYSDOH**	35962	56.3	8	15	5	10
15	NA			1984	JPN	ocular	**AB679754**	**Hokkaido**	35960	56.3	8	15	5	10
16	V1933			1985	NM,USA	respiratory	KX384955	GenBank	35961	56.3	8	15	5	8
17	NA			1991	JPN	ocular	**AB679755**	**Hokkaido**	35961	56.3	8	15	5	10
18	ZG 95-873			1995	CA,USA	respiratory	KX384951	GenBank	35967	56.3	8	14	5	10
19	078Jax			1997	SC,USA	respiratory	KX384953	GenBank	35963	56.3	8	15	5	10
20	186Jax			1998	SC,USA	respiratory	KX384952	GenBank	35963	56.3	8	15	5	10
21	10Jax			2001	SC,USA	respiratory	KX384954	GenBank	35962	56.3	8	15	5	10
22	NA			2001	JPN	ocular	**AB679756**	**Hokkaido**	35963	56.3	7	14	5	10
23	NHRC11023			2001	IL,USA	respiratory	**AP014849**	**Hokkaido**	35973	56.3	8	14	5	9
24	NHRC50654			2001	TX,USA	respiratory	**AP014850**	**Hokkaido**	35964	56.3	8	14	5	9
25	T158			2002	SC,USA	respiratory	KX384945	GenBank	35965	56.3	8	15	5	10
26	NHRC3			2002	TX,USA	respiratory	AY599837	GenBank	35965	56.3	8	14	5	10
27	NHRC42606	**II**	**a-like**	2003	SC,USA	respiratory	AY599835	GenBank	35965	56.3	8	15	5	10
28	NHRC70935			2004	SC,USA	respiratory	**AP014844**	**Hokkaido**	35967	56.3	8	14	5	10
29	NHRC22650			2006	CA,USA	respiratory	**AP014841**	**Hokkaido**	36155	56.3	8	15	5	10
30	GZ01			2008	CHN	respiratory	KF006344	GenBank	35960	56.3	8	14	5	10
31	NHRC23703			2008	CA,USA	respiratory	**AP014842**	**Hokkaido**	35959	56.3	8	15	5	10
32	NHRC92165			2009	NJ,USA	respiratory	**AP014845**	**Hokkaido**	35964	56.3	8	15	5	10
33	WPAFB7			2009	CA,USA	respiratory	**AP014847**	**Hokkaido**	35961	56.3	8	13	6	10
34	TB071911			2011	CT,USA	respiratory	KY996453	GenBank	35952	56.3	8	15	5	10
35	NHRC36401			2011	MO,USA	respiratory	**AP014851**	**Hokkaido**	35960	56.3	8	13	6	10
36	NYS 12-12752			2012	NY,USA	respiratory	KY996450	GenBank	35955	56.3	8	15	5	10
37	NYS 12-27440			2012	NY,USA	respiratory	KY996451	GenBank	35948	56.3	8	14	6	10
38	NYS 13-5497			2013	NY,USA	respiratory	KY996449	GenBank	35960	56.3	8	13	6	10
39	NYS 14-4876			2014	NY,USA	respiratory	KY996448	GenBank	35934	56.3	8	15	5	10
40	NYS 14-38662			2014	NY,USA	respiratory	KY996443	GenBank	35960	56.3	8	14	6	10
41	NYS 14-38813			2014	NY,USA	respiratory	KY996444	GenBank	35948	56.3	8	14	6	10
42	NYS 14-33430			2014	NY,USA	respiratory	KY996445	GenBank	35948	56.3	8	14	6	10
43	NYS 14-9111			2014	NY,USA	respiratory	KY996442	GenBank	35948	56.3	8	14	6	10
44	NYS 15-3477			2015	NY,USA	respiratory	KY996446	GenBank	35949	56.3	8	14	5	9
45	NYS 15-1428			2015	NY,USA	respiratory	MF002042	GenBank	35960	56.3	8	14	6	10

* Restriction sites predicted by CLC Genomics Workbench v10.1.1 ** Sequences obtained deposited in this study are in bold font

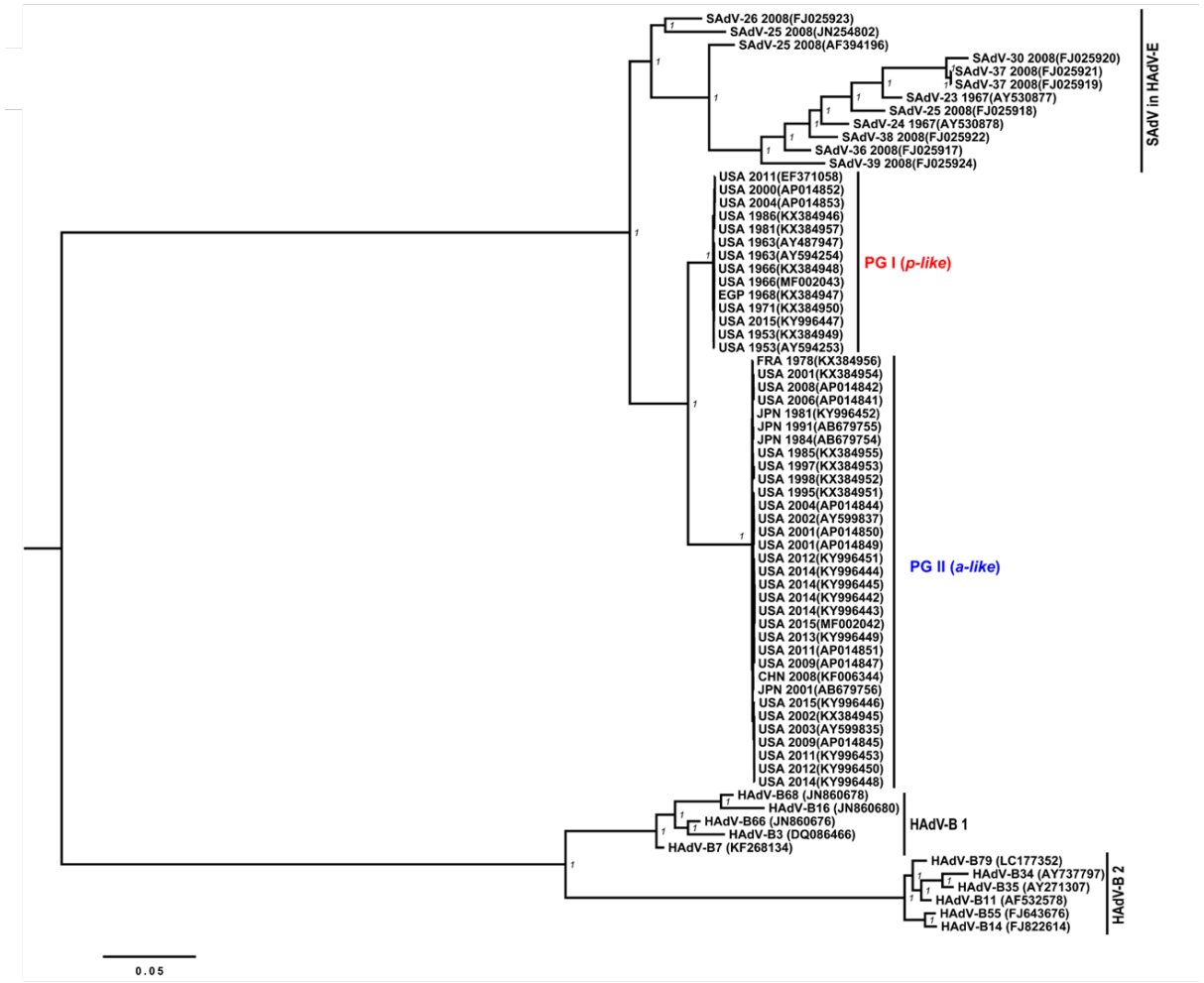


Figure 1. HAdV-E4 comprises two distinguishable phylogroups. Phylogenetic tree of whole genome sequences of HAdVs of species B, E and non-human primate adenovirus of species HAdV-E (SAdVs). Bayesian posterior probability support is shown next to the branches.

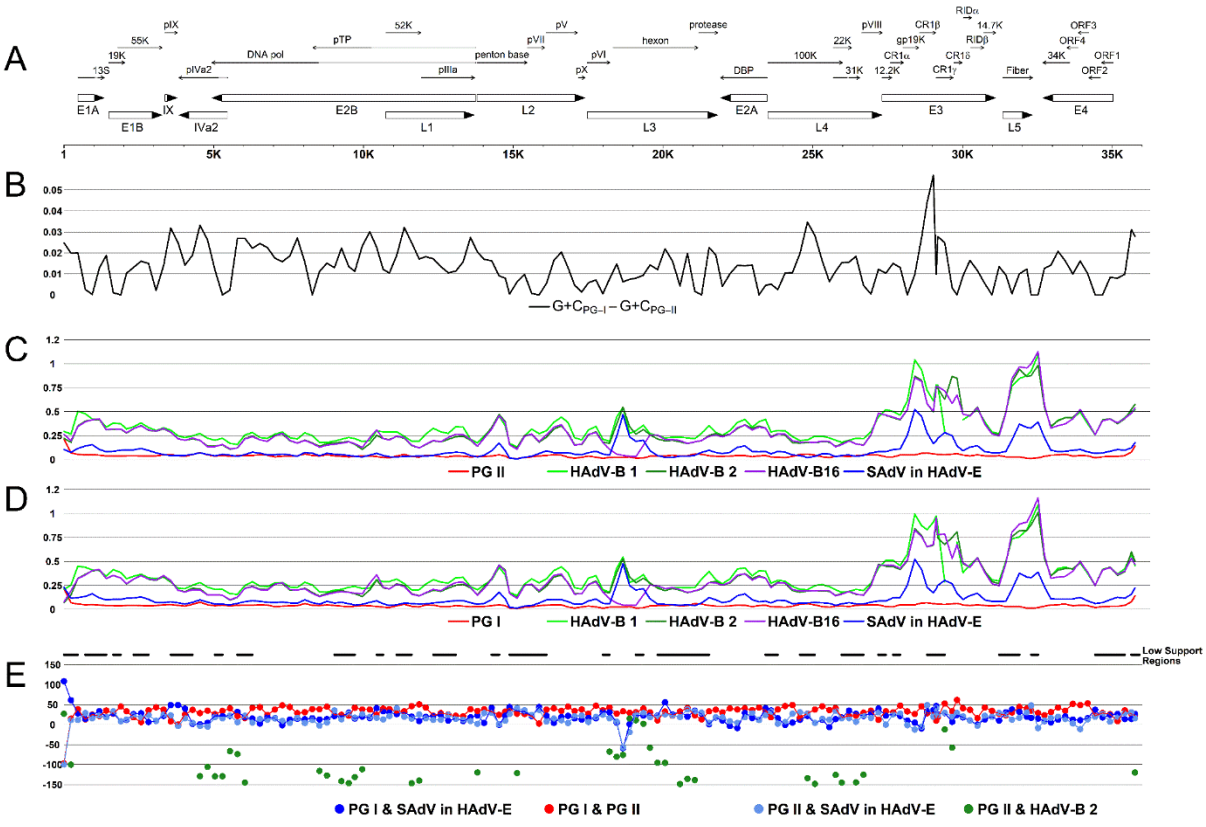
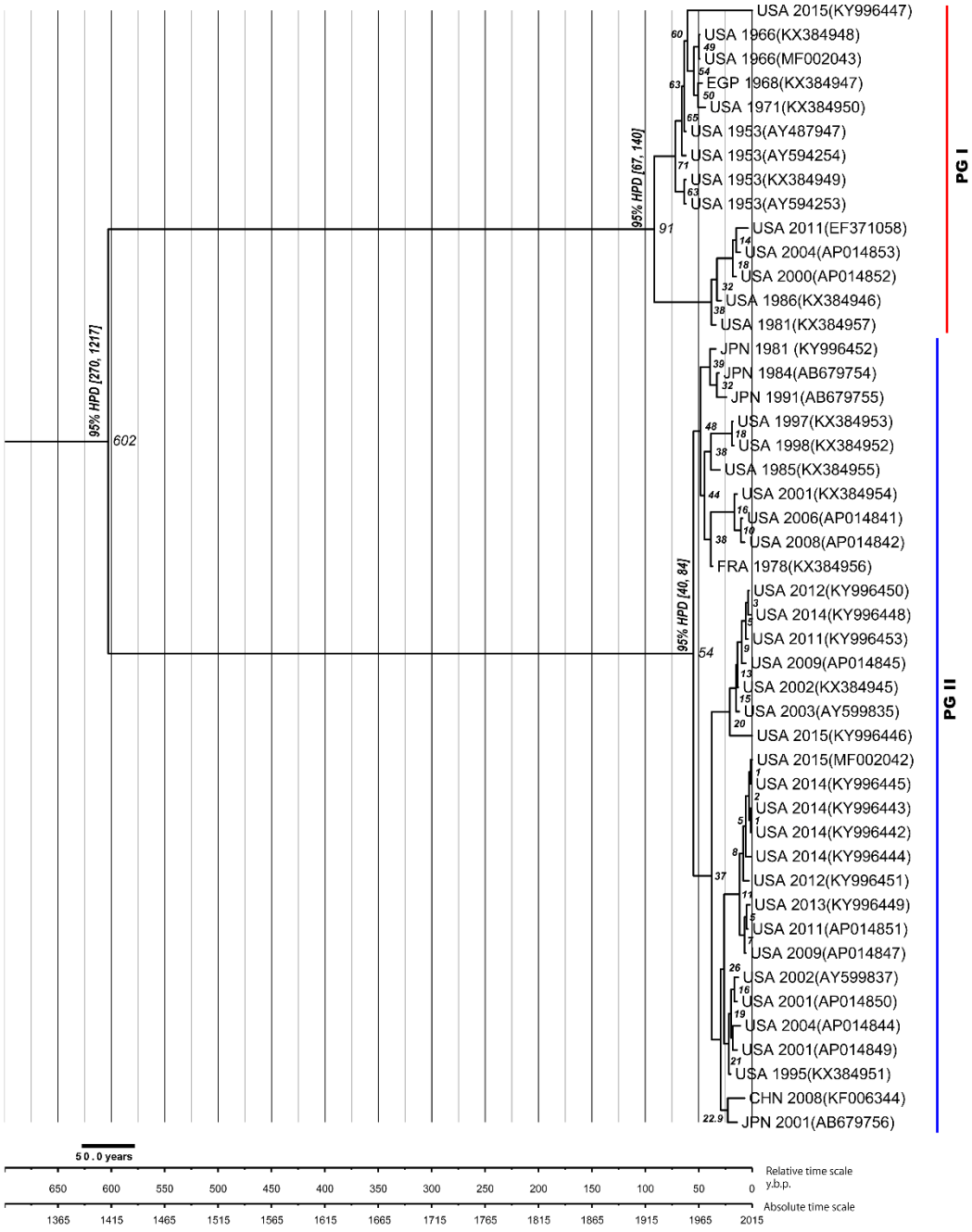


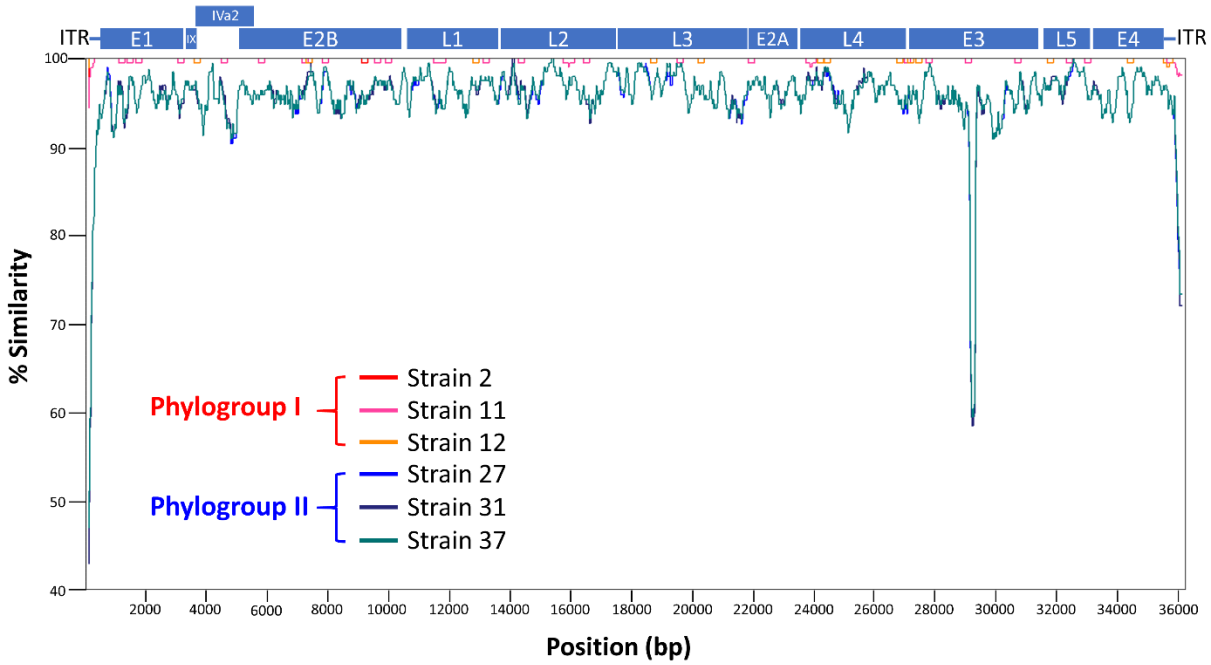
Figure 2. Genomic differences along the two phylogroups. The horizontal axes represent the genomic positions in HAdV-E relative to the prototype strain RI-67 USA, 1953 (KX384949). (A) Genomic annotation of HAdV-E. (B) Sliding window analysis of the average %G+C difference between PG I and PG II across the genome. The vertical and horizontal axes show the average percentage difference between both phylogroups and the genome position, respectively. (C and D) Sliding window analyses of evolutionary distance between members of PG I and PG II to sequences of other clusters, respectively, vertical axes show the average evolutionary distance in the respective window to sequences in subspecies HAdV-B 1 (except HAdV-B16), subspecies HAdV-B 2, HAdV-B16 (JN860680) and SAdVs in species HAdV-E (see Fig. 1). (E) Sliding window analysis comparing the support for PG I and PG II cluster versus clusters of PGI and PGII with other types and species. The vertical axis shows the Bayes factor between sequences in clusters color-coded as per the key below with higher values showing higher support for the topological clustering of the groups as shown in the bottom of the panel. Regions with low topological support for the clustering of PG I and PG II are highlighted by black lines on the top of the panel.



886

887 **Figure 3. Bayesian estimation of the time to the most recent common ancestor**
888 **for HAdV-E4 strains in PGI and PGII.** The phylogenetic tree is annotated in the
889 branches with years before the present. The 95% highest posterior density (HPD)
890 ranges for tMRCAs of all sequences are shown for both phylogroups between brackets.
891 The relative time and absolute time scales are shown in the bottom. The median
892 relative divergence time for other branches is shown next to the branches.

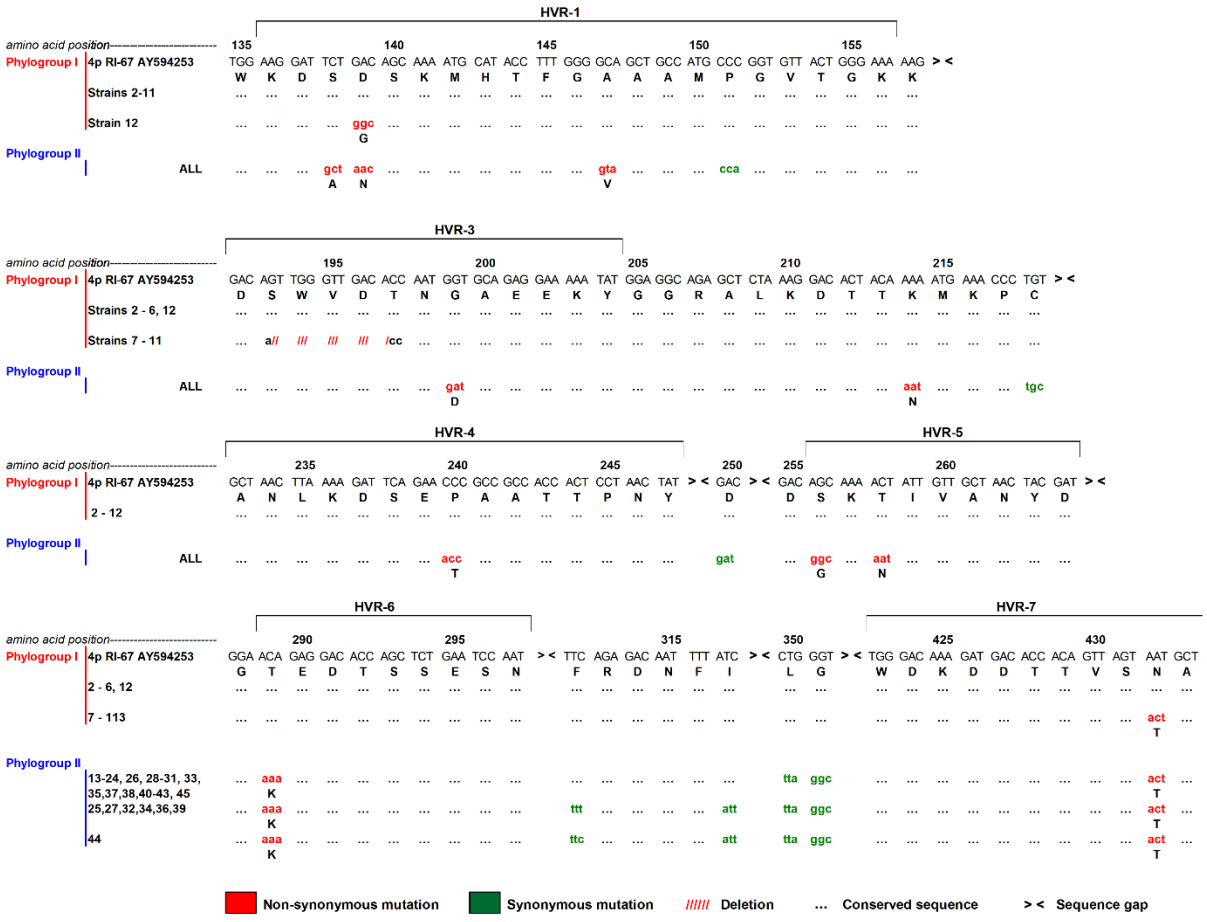
893



894

895 **Figure 4. Analysis of genomic sequences of representative strains of**
896 **phylogroups I and II for regions of divergence and similarity.** A similarity plot was
897 generated in Simplot using the whole genome sequence (WGS) for strain 1 (RI-67) as
898 the query and the WGS for strains 2, 11, 12, 27, 31 and 37 as the references. The plot
899 represents the percent similarity in a 200 nucleotides sliding window and 20-nucleotide
900 step size with gapped sites removed.

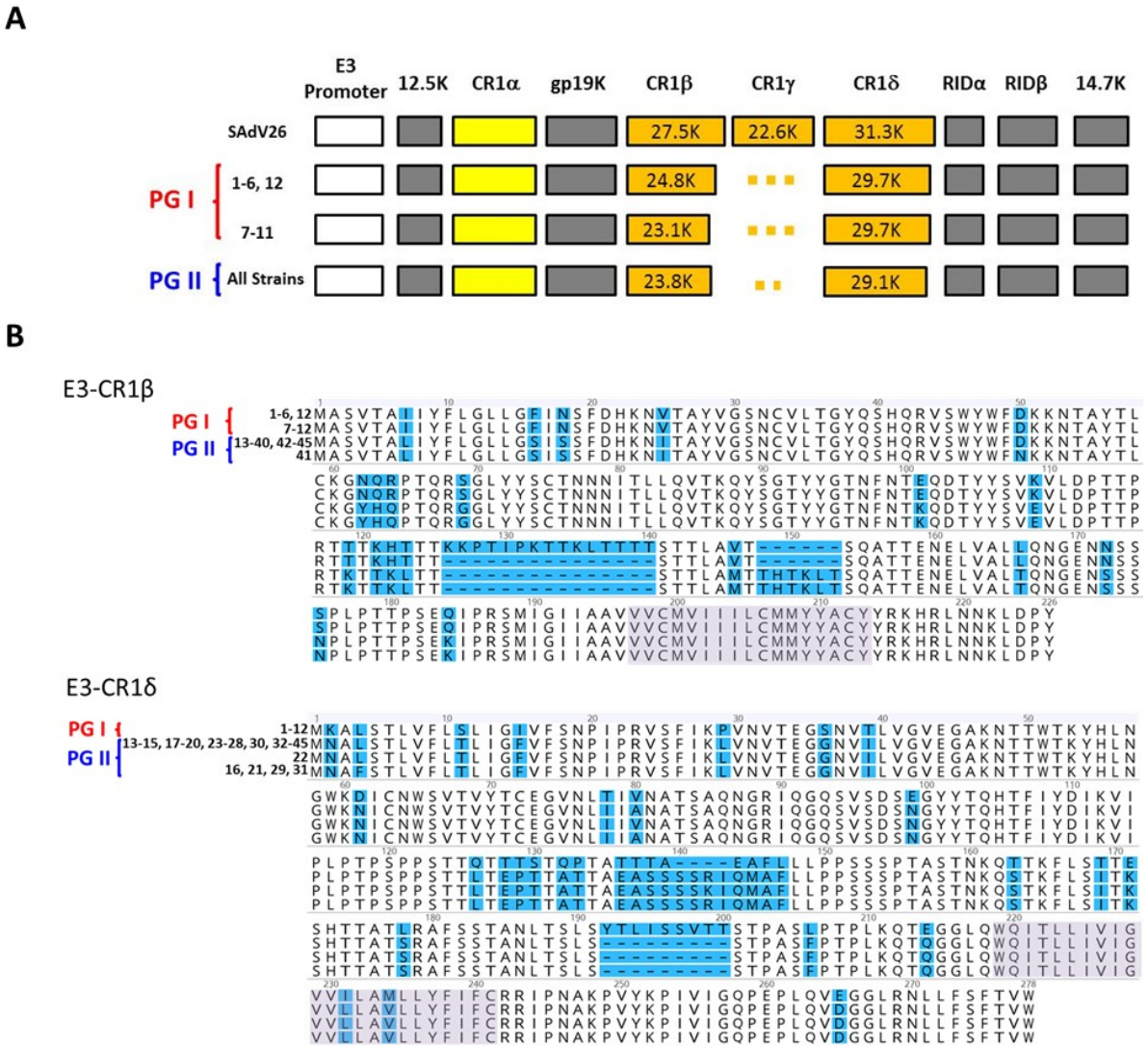
901



902

903 **Figure 5. Amino acid differences identified in the hypervariable regions (HVR)**
904 **1 and HVRs 3-7 of the hexon polypeptide among examined strains of**
905 **phylogroups I and II**. No differences were identified for HVR 2 so the corresponding
906 section of the sequences is not shown

907



908

909 **Figure 6. A. Schematic comparing the genetic content of the E3 region for SAdV-**
910 **26 and strains of HAdV-E4 in phylogroups I and II. B. Alignment of amino acid**
911 **sequences of E3 CR1β and E3-CR1δ.** Amino acid differences between phylogroups
912 are highlighted in blue. The transmembrane domain of the predicted type I membrane
913 proteins is highlighted in light purple.

914

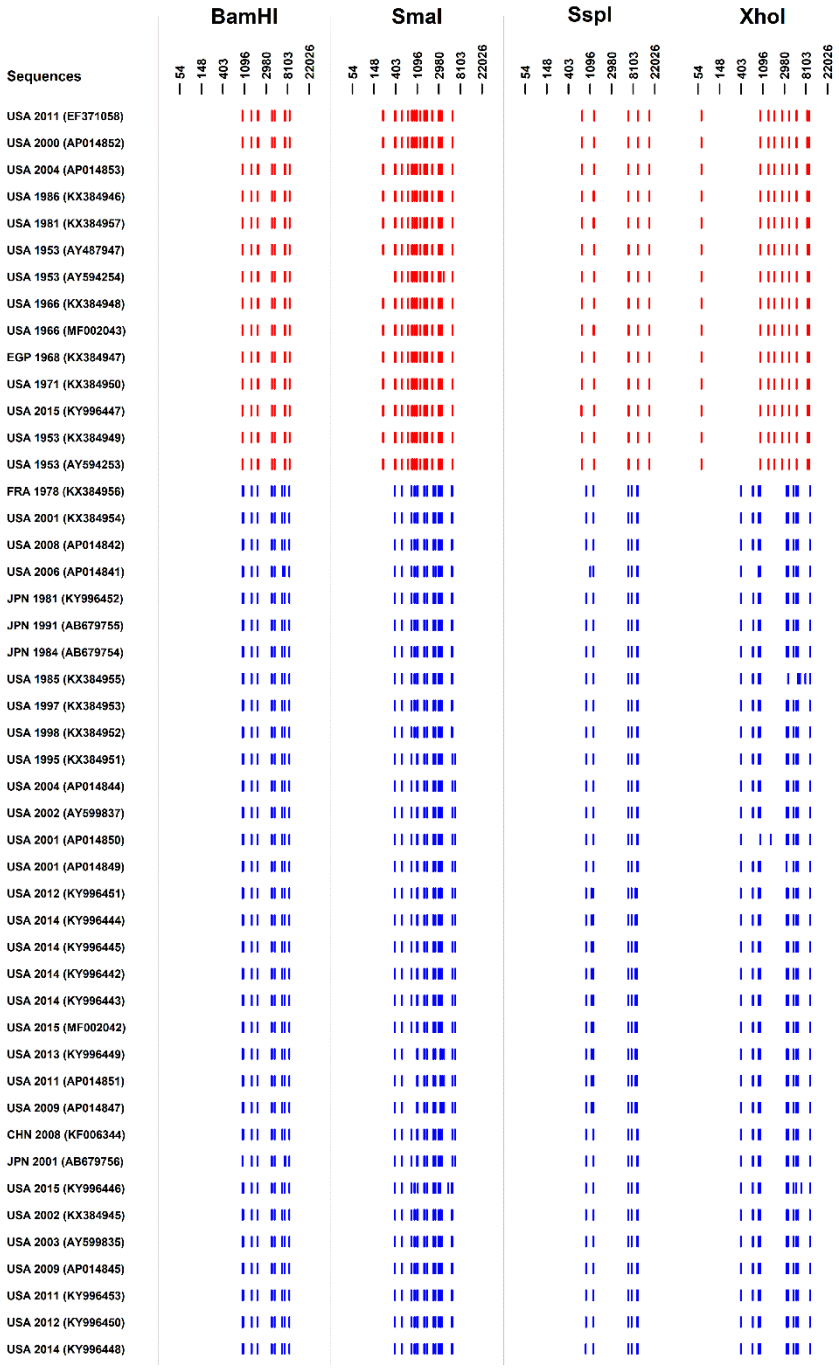
915

Supplementary Table 1: Multiple sequence positions of the start and end sites of windows supporting the clustering of PG I and PG II. The unaligned positions are shown for AY594253 and KX384956 as reference.

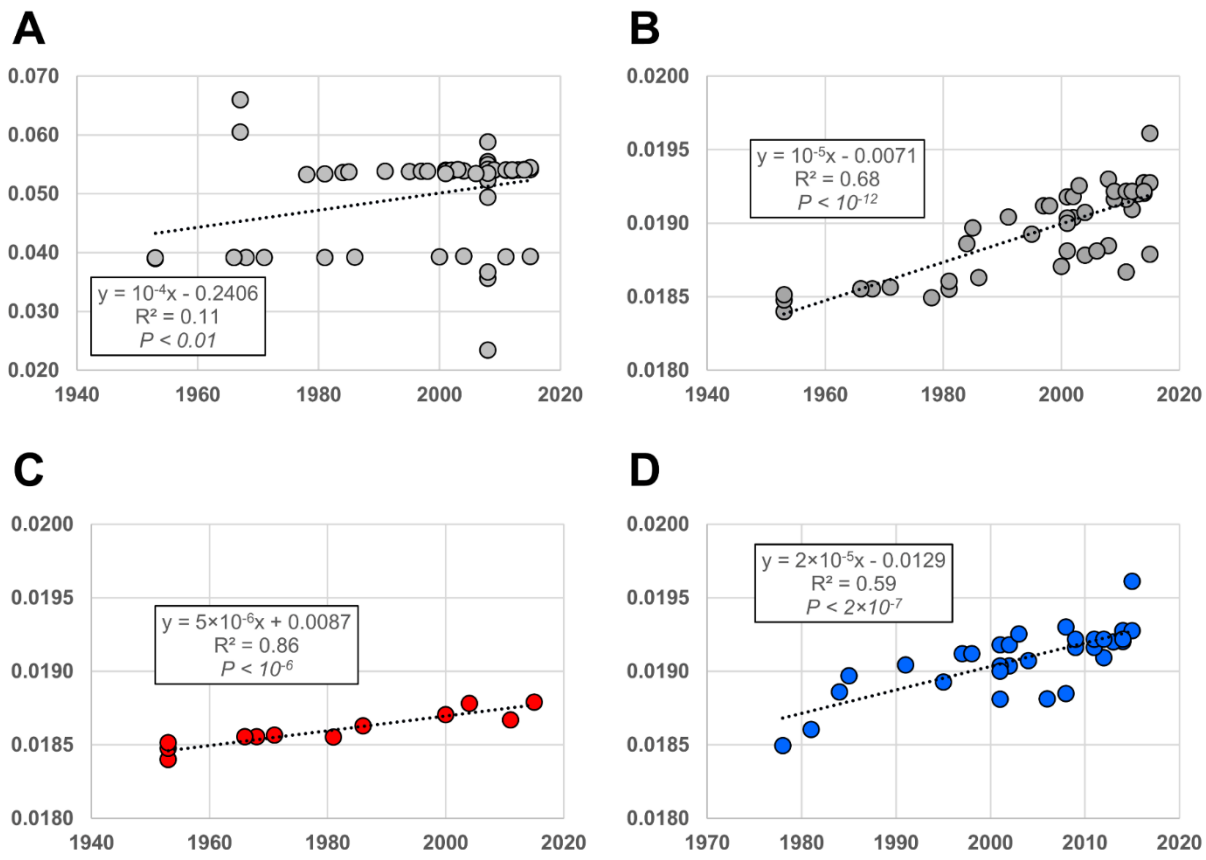
	MSA		PG I [USA 1953 (AY594253)]		PG II [FRA 1978 (KX384956)]	
No.	Start	End	Start	End	Start	End
1	501	1001	466	945	466	912
2	1501	2001	1414	1898	1381	1863
3	2001	2501	1898	2326	1863	2327
4	2251	2751	2091	2576	2089	2577
5	3001	3501	2826	3326	2827	3327
6	3251	3751	3076	3565	3077	3565
7	3501	4001	3326	3812	3327	3812
8	4501	5001	4293	4793	4295	4795
9	4751	5251	4543	5043	4545	5045
10	5001	5501	4793	5293	4795	5295
11	5501	6001	5293	5793	5295	5795
12	5751	6251	5543	6043	5545	6045
13	6501	7001	6293	6793	6295	6795
14	6751	7251	6543	7043	6545	7045
15	7001	7501	6793	7293	6795	7295
16	7251	7751	7043	7543	7045	7545
17	7501	8001	7293	7793	7295	7795
18	7751	8251	7543	8043	7545	8045
19	8001	8501	7793	8293	7795	8295
20	8251	8751	8043	8531	8045	8533
21	8501	9001	8293	8781	8295	8783
22	8751	9251	8531	9031	8533	9033
23	9001	9501	8781	9266	8783	9268
24	10001	10501	9711	10211	9713	10213
25	10251	10751	9961	10441	9963	10443
26	10501	11001	10211	10660	10213	10596
27	11001	11501	10660	11110	10596	11045
28	11251	11751	10866	11360	10801	11295
29	12001	12501	11610	12081	11545	12016
30	12251	12751	11860	12331	11795	12266
31	12501	13001	12081	12581	12016	12516
32	13501	14001	13081	13572	13016	13513
33	13751	14251	13328	13804	13263	13746
34	14001	14501	13572	14029	13513	13971
35	14251	14751	13804	14279	13746	14221
36	14501	15001	14029	14529	13971	14471
37	15001	15501	14529	14879	14471	14821
38	15251	15751	14733	15129	14675	15071
39	16751	17251	16112	16606	16054	16548
40	17001	17501	16359	16829	16301	16771
41	17251	17751	16606	17049	16548	16991
42	17501	18001	16829	17287	16771	17229
43	17751	18251	17049	17511	16991	17453
44	18001	18501	17287	17761	17229	17699
45	18251	18751	17511	17992	17453	17919
46	18501	19001	17761	18215	17699	18142
47	19001	19501	18215	18666	18142	18593
48	19251	19751	18455	18886	18382	18813

49	19501	20001	18666	19095	18593	19022
50	19751	20251	18886	19345	18813	19272
51	20251	20751	19345	19818	19272	19745
52	20501	21001	19568	20068	19495	19995
53	22501	23001	21539	21984	21466	21914
54	22751	23251	21762	22234	21692	22164
55	23001	23501	21984	22484	21914	22414
56	23251	23751	22234	22734	22164	22664
57	23501	24001	22484	22978	22414	22908
58	23751	24251	22734	23207	22664	23131
59	24001	24501	22978	23422	22908	23346
60	24251	24751	23207	23606	23131	23530
61	25001	25501	23832	24323	23756	24247
62	25251	25751	24079	24573	24003	24497
63	25501	26001	24323	24823	24247	24747
64	26251	26751	25073	25555	24997	25479
65	26501	27001	25317	25763	25241	25675
66	26751	27251	25555	25969	25479	25869
67	27001	27501	25763	26210	25675	26113
68	28001	28501	26693	27193	26595	27095
69	28251	28751	26943	27440	26845	27342
70	28751	29251	27440	27921	27342	27823
71	29251	29751	27921	28405	27823	28307
72	29501	30001	28158	28632	28060	28533
73	29751	30251	28405	28819	28307	28700
74	30001	30501	28632	29025	28533	28899
75	31251	31751	29407	29815	29261	29658
76	31501	32001	29655	30015	29510	29868
77	31751	32251	29815	30250	29658	30090
78	32001	32501	30015	30499	29868	30339
79	32251	32751	30250	30743	30090	30583
80	32501	33001	30499	30984	30339	30824
81	32751	33251	30743	31222	30583	31062
82	33001	33501	30984	31449	30824	31263
83	34001	34501	31899	32293	31713	32107
84	34251	34751	32146	32521	31960	32335
85	34751	35251	32521	32951	32335	32765
86	35001	35501	32738	33190	32552	33004
87	35251	35751	32951	33431	32765	33245
88	35501	36001	33190	33681	33004	33495
89	35751	36251	33431	33931	33245	33745
90	36001	36501	33681	34178	33495	33992
91	36251	36751	33931	34428	33745	34242
92	36501	37001	34178	34677	33992	34491
93	37751	38251	35420	35756	35234	35571

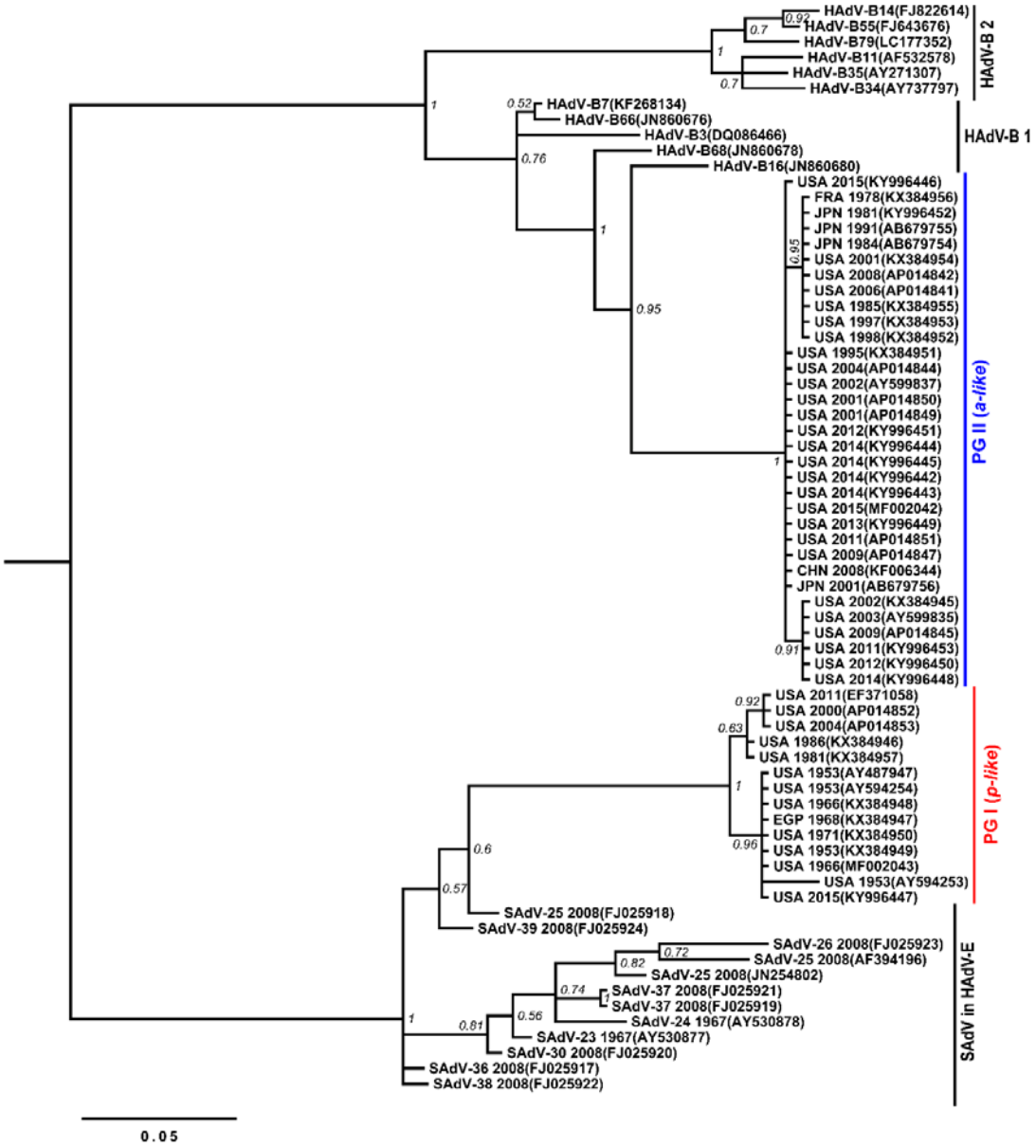
919



Supplementary Figure 1: *In silico* RFLP analysis for all examined genomes of PG I and II with endonucleases BamHI, SmaI, SspI and XhoI (see also Table 1).



Supplementary Figure 2: Time-structure assessment by tip-to-root regression for HAdV-E datasets. The vertical and horizontal axes show the distance from the tip to the root and the year of sampling, respectively. For each dataset, the dotted line corresponding to the linear regression, the equation of the regression, the R^2 and P are displayed in each panel. The panels display the estimated regression for the datasets including A) HAdV-E and SAdV, B) HAdV-E, C) PG I and D) PG II.



933

934 **Supplementary Figure 3:** Phylogenetic tree of the section between sites 1 and 230
935 encompassing the ITR for PG I and PG II. The phylogenetic tree also includes
936 sequences of HAdVs of species B, E and non-human primate adenovirus of species
937 HAdV-E (SAdVs). Bayesian posterior probability support is shown next to the
938 branches.

939