Genetic content and evolution of adenoviruses

Andrew J. Davison,1 Mária Benkö2 and Balázs Harrach2

1MRC Virology Unit, Institute of Virology, Church Street, Glasgow G11 5JR, UK
2Veterinary Medical Research Institute, Hungarian Academy of Sciences, H-1581 Budapest, Hungary

This review provides an update of the genetic content, phylogeny and evolution of the family Adenoviridae. An appraisal of the condition of adenovirus genomics highlights the need to ensure that public sequence information is interpreted accurately. To this end, all complete genome sequences available have been reannotated. Adenoviruses fall into four recognized genera, plus possibly a fifth, which have apparently evolved with their vertebrate hosts, but have also engaged in a number of interspecies transmission events. Genes inherited by all modern adenoviruses from their common ancestor are located centrally in the genome and are involved in replication and packaging of viral DNA and formation and structure of the virion. Additional niche-specific genes have accumulated in each lineage, mostly near the genome termini. Capture and duplication of genes in the setting of a ‘leader–exon structure’, which results from widespread use of splicing, appear to have been central to adenovirus evolution. The antiquity of the pre-vertebrate lineages that ultimately gave rise to the Adenoviridae is illustrated by morphological similarities between adenoviruses and bacteriophages, and by use of a protein-primed DNA replication strategy by adenoviruses, certain bacteria and bacteriophages, and linear plasmids of fungi and plants.

INTRODUCTION

The purpose of this review is to provide a comprehensive update of the genetic content, phylogeny and evolution of the family Adenoviridae, whose members infect hosts throughout the vertebrates (Russell & Benkö, 1999). This area has frequently taken a back seat to studies of the interactions of selected human adenovirus proteins with cellular processes and, more recently, to the use of adenoviruses as vectors. Our aim is to redress this imbalance by bringing together published information and by offering fresh insights into the genomics of the family as a whole. It is not our purpose to deal in broad scope with other areas, such as the expression, functions and interactions of adenovirus gene products and the ways in which host defences are manipulated during the infectious cycle. These have been reviewed recently by Russell (2000) and Shenk (2001) for human adenoviruses.

Phylogeny, classification and genetic organization

Members of the family Adenoviridae are non-enveloped, icosahedral viruses that replicate in the nucleus. Their linear, double-stranded DNA molecules are 26–45 kbp in size and rank as medium-sized among the DNA viruses. The genomes are characterized by an inverted terminal repeat (ITR) ranging in size from 36 to over 200 bp, and the 5’ ends are linked to a terminal protein (TP). Phylogenetic relationships among a large number of adenoviruses infecting vertebrates from fish to humans are shown in Fig. 1. The major clades (groups of viruses sharing a common ancestor) correspond to the four accepted genera plus a fifth genus that is likely to be added in due course (Benkö et al., 2002; Harrach & Benkö, 1998). Two genera (Mastadenovirus and Aviadenovirus) originate from mammals or birds, respectively, and the other two genera (Atadenovirus and Siadenovirus) have a broader range of hosts. Atadenoviruses were named as a consequence of the bias of their genomes towards high A+T content (Benkö & Harrach, 1998; Both, 2002a), and infect various ruminant, avian and reptilian hosts, as well as a marsupial. The two known siadenoviruses were isolated from birds and a frog (Davison & Harrach, 2002). The only confirmed fish adenovirus falls into the fifth clade. Within each genus, viruses are grouped into species, which are named from the host and supplemented with letters of the alphabet, as listed in Table 1 (Benkö et al., 2000). Host origin is only one of several criteria used to demarcate the species taxon. Thus, for example, known chimpanzee adenoviruses are classified into human adenovirus species. In the case of human adenoviruses, the present species correspond to groups or subgenera defined previously (Bailey & Mautner, 1994; Wadell, 1984). Questions of classification into species are not yet resolved for many non-human adenoviruses.
Fig. 1. Distance tree summarizing the phylogeny of adenovirus hexon genes. Members of the various genera are indicated in different colours, and viruses that belong to the same species are grouped by light-green ovals. Abbreviations of virus names are indicated at the ends of the branches, with species names listed to the right (recognized species in italics): B, bovine; C, canine; D, duck; E, equine; F, fowl; Fr, frog; H, human; M, murine; O, ovine; P, porcine; Po, possum; Sn, snake; T, turkey; and TS, tree shrew. The distance matrix was calculated from amino acid sequences for hexon available in GenBank (in some cases combined from different partial entries) and from our unpublished sequences. The PROTDIST (Dayhoff PAM 001 matrix) and FITCH (global rearrangements) programs of the PHYLIP package, version 3.6, were used. The tree was rooted by specifying sturgeon adenovirus as the outgroup and displayed using TREEVIEW (Page, 1996). Bootstrap values of less than 80/100 (not including branches within species) are indicated by small, filled circles.
Table 1. Available sequences for complete adenovirus genomes

<table>
<thead>
<tr>
<th>GENUS</th>
<th>Species</th>
<th>Serotype</th>
<th>Abbreviation</th>
<th>Accession</th>
<th>Size (bp)</th>
<th>Reference</th>
<th>TPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>MASTADENOVIUS</td>
<td>Human adenovirus A</td>
<td></td>
<td>HAdV-A</td>
<td>X73487</td>
<td>34125</td>
<td>Sprengel et al. (1994)</td>
<td>BK000405</td>
</tr>
<tr>
<td></td>
<td>Human adenovirus B</td>
<td></td>
<td>HAdV-B</td>
<td>Y163756†</td>
<td>35749</td>
<td>Stone et al. (2003)</td>
<td>BK001453</td>
</tr>
<tr>
<td></td>
<td>Human adenovirus C</td>
<td></td>
<td>HAdV-C</td>
<td>J01917</td>
<td>35937</td>
<td>Roberts et al. (1986)</td>
<td>BK000407</td>
</tr>
<tr>
<td></td>
<td>Human adenovirus D</td>
<td></td>
<td>HAdV-D</td>
<td>AF106105</td>
<td>35100</td>
<td>Chillon et al. (1999)</td>
<td>BK000406</td>
</tr>
<tr>
<td></td>
<td>Human adenovirus F</td>
<td></td>
<td>HAdV-F</td>
<td>L19443</td>
<td>34214</td>
<td>Davison et al. (1993)</td>
<td>(\dag)</td>
</tr>
<tr>
<td></td>
<td>Ovine adenovirus A</td>
<td></td>
<td>OAdV-B</td>
<td>AF252854</td>
<td>33034</td>
<td>Unpublished (2000)</td>
<td>BK000400</td>
</tr>
<tr>
<td></td>
<td>Bovine adenovirus 2</td>
<td></td>
<td>BAdV-2</td>
<td>AF252854</td>
<td>33034</td>
<td>Unpublished (2000)</td>
<td>BK000400</td>
</tr>
<tr>
<td></td>
<td>Porcine adenovirus C</td>
<td></td>
<td>PAdV-C</td>
<td>AF289282</td>
<td>32621</td>
<td>Nagy et al. (2001)</td>
<td>BK000411</td>
</tr>
<tr>
<td></td>
<td>Porcine adenovirus 5</td>
<td></td>
<td>PAdV-5</td>
<td>AF289282</td>
<td>32621</td>
<td>Nagy et al. (2001)</td>
<td>BK000411</td>
</tr>
<tr>
<td></td>
<td>Bovine adenovirus B</td>
<td></td>
<td>BAdV-B</td>
<td>AF030154</td>
<td>34446</td>
<td>Reddy et al. (1998b)</td>
<td>BK000401</td>
</tr>
<tr>
<td></td>
<td>Bovine adenovirus 3</td>
<td></td>
<td>BAdV-3</td>
<td>Y07760†</td>
<td>30536</td>
<td>Morrison et al. (1997)</td>
<td>BK000402</td>
</tr>
<tr>
<td></td>
<td>Porcine adenovirus A</td>
<td></td>
<td>PAdV-A</td>
<td>AF0833123§</td>
<td>34094</td>
<td>Reddy et al. (1998a)</td>
<td>BK000410</td>
</tr>
<tr>
<td></td>
<td>Murine adenovirus A</td>
<td></td>
<td>MAAdV-A</td>
<td>NC_000942</td>
<td>30944</td>
<td>Meissner et al. (1997)</td>
<td>BK000415</td>
</tr>
<tr>
<td></td>
<td>Tree shrew adenovirus</td>
<td></td>
<td>TSAdV</td>
<td>AF256784</td>
<td>33501</td>
<td>Schönfeld et al. (2003)</td>
<td>BK000455</td>
</tr>
<tr>
<td></td>
<td>Tree shrew adenovirus 1</td>
<td></td>
<td>TSAdV-1</td>
<td>AF256784</td>
<td>33501</td>
<td>Schönfeld et al. (2003)</td>
<td>BK000455</td>
</tr>
<tr>
<td>ATADENOVIUS</td>
<td>Duck adenovirus A</td>
<td></td>
<td>DAdV-A</td>
<td>Y09598</td>
<td>33213</td>
<td>Hess et al. (1997)</td>
<td>BK000404</td>
</tr>
<tr>
<td></td>
<td>Duck adenovirus 1</td>
<td></td>
<td>DAdV-1</td>
<td>Y09598</td>
<td>33213</td>
<td>Hess et al. (1997)</td>
<td>BK000404</td>
</tr>
<tr>
<td></td>
<td>Ovine adenovirus D</td>
<td></td>
<td>OAdV-D</td>
<td>U40839</td>
<td>29674</td>
<td>Vrati et al. (1996)</td>
<td>(\dag)</td>
</tr>
<tr>
<td></td>
<td>Ovine adenovirus 7</td>
<td></td>
<td>OAdV-7</td>
<td>U40839</td>
<td>29674</td>
<td>Vrati et al. (1996)</td>
<td>(\dag)</td>
</tr>
<tr>
<td></td>
<td>Bovine adenovirus D</td>
<td></td>
<td>BAdV-D</td>
<td>AF036092</td>
<td>31300</td>
<td>Unpublished (1997)</td>
<td>(\dag)</td>
</tr>
<tr>
<td>AVIADENOVIUS</td>
<td>Fowl adenovirus A</td>
<td></td>
<td>FAAdV-A</td>
<td>U46933</td>
<td>43804</td>
<td>Chiocca et al. (1996)</td>
<td>BK001452</td>
</tr>
<tr>
<td></td>
<td>Fowl adenovirus 1</td>
<td></td>
<td>FAAdV-1</td>
<td>U46933</td>
<td>43804</td>
<td>Chiocca et al. (1996)</td>
<td>BK001452</td>
</tr>
<tr>
<td></td>
<td>Fowl adenovirus D</td>
<td></td>
<td>FAAdV-3</td>
<td>AF083975</td>
<td>45083</td>
<td>Ojkic &amp; Nagy (2000)</td>
<td>BK001451</td>
</tr>
<tr>
<td></td>
<td>Fowl adenovirus 9</td>
<td></td>
<td>FAAdV-9</td>
<td>AF083975</td>
<td>45083</td>
<td>Ojkic &amp; Nagy (2000)</td>
<td>BK001451</td>
</tr>
<tr>
<td>SIADENOVIUS</td>
<td>Turkey adenovirus A</td>
<td></td>
<td>TAAdV-A</td>
<td>AF074946</td>
<td>26283</td>
<td>Pitcovski et al. (1998)</td>
<td>BK000454</td>
</tr>
<tr>
<td></td>
<td>Turkey adenovirus 3</td>
<td></td>
<td>TAAdV-3</td>
<td>AF074946</td>
<td>26283</td>
<td>Pitcovski et al. (1998)</td>
<td>BK000454</td>
</tr>
<tr>
<td></td>
<td>Frog adenovirus</td>
<td></td>
<td>FrAdV</td>
<td>AF224336</td>
<td>26183</td>
<td>Davison et al. (2000)</td>
<td>(\dag)</td>
</tr>
</tbody>
</table>

*Another sequence is available for this strain (AF332578; Mei et al., 2003).
†Another sequence was derived subsequently for this serotype by the same group (AF394196; Farina et al., 2001). It is not clear whether this is from the same strain, but in any case it is substantially inferior in quality, containing 59 ambiguous nucleotides and several frameshifts.
‡Another sequence is available for a strain of this serotype (U55001; unpublished, 1996). This differs from the sequence given in the Table at 53 locations, as well as in a deletion of part of the E3 region. None of these differences indicates frameshift errors.
§A duplicated accession for a passaged derivative of this strain is available (AB026117 and AJ237815; unpublished, 1999). This sequence differs from that given in the Table at 33 locations. Eight of these differences indicate frameshift errors in one or other sequence.
¶TPAs were not submitted since the authors' original accessions have been updated directly.
Gene arrangements for representative members of the adenovirus genera recognized are illustrated in Fig. 2. The upper panel shows the layout for a human adenovirus species (specifically, a chimpanzee virus). In the discussion below, we use the same term (e.g. pX) to apply to gene or protein, as indicated by the context, with proteins that are produced as precursors for cleavage by the viral protease prefaced by a ‘p’. In outline, genomes of human adenovirus species consist of a central block of rightward-oriented late genes from 52K to fiber, interrupted on the same strand by...

Fig. 2. Gene layout in representatives of the four adenovirus genera recognized. Each genome is represented by a central blue horizontal line marked at 5 kbp intervals. Protein-encoding regions are shown as arrows. Cyan denotes genus-common genes, other colours indicate related genus-specific genes (E1B 55K and E4 34K are related between the Mastadenovirus and Atadenovirus genera), and genus-specific genes that lack relatives are white. E1 and E4 are marked in all genera and E3 in the Mastadenovirus and Siadenovirus genera, and these regions are shaded violet. E2A and E2B genes are present in all genera, and these regions are indicated for the Mastadenovirus genus. Untranslated leader exons in the Mastadenovirus genus are shown for late, E2, E3 and E4 genes, with early transcripts in red, late transcripts in blue and 5'–terminal exons broader than internal exons. The VA RNA genes in the Mastadenovirus and Aviadenovirus genera are also included.
block of early genes in the E3 region and on the opposing strand by E2 genes in the form of DNA-binding protein (DBP) in the E2A region and pre-terminal protein (pTP) and DNA polymerase (pol) in the E2B region. The right-terminal region is occupied by E4 genes and the left-terminal region by early E1A and E1B genes plus two intermediate genes (IX and IVa2).

Splicing

Splicing was first discovered in human adenoviruses as the mechanism by which several late genes are expressed from a shared non-translated tripartite leader (Berget et al., 1977; Chow et al., 1977). It has since become clear that splicing is involved in expression of most human adenovirus genes (Akusjärvi et al., 1986). Thus, all members of the array of late genes (which are largely involved in virion formation and structure), from 52K rightwards to pVIII, and also fiber, are spliced from the tripartite leader (Fig. 2). Similarly, early genes and one intermediate gene are spliced. These include: E1A; DBP, which is spliced from a leader consisting of two untranslated exons controlled by the E2 promoter (expression also occurs from a separate late promoter; Fig. 2); pTP and pol, which are spliced from a leader consisting of two untranslated and one translated exons controlled by the E2 promoter (Fig. 2); IVa2, an intermediate gene that is spliced from a short coding leader within pol; and the E3 and E4 groups of genes, each of which is initiated from a single promoter so that the proximal gene is not spliced whereas others are expressed by splicing from the relevant untranslating leader (Fig. 2). Transcription of E1A, E1B, E3 and E4, and to a lesser extent other genes, has been perceived historically as more complex, potentially giving rise via alternative splicing to a large catalogue of RNAs and proteins, some of which have detectable functions in in vitro assays (Akusjärvi et al., 1986; Chow et al., 1979; Tigges & Raskas, 1984; Virtanen et al., 1984; Wold & Gooding, 1991). However, transcript mapping studies have largely been carried out on a single human adenovirus species (HAdV-C, represented by HAdV-2 and HAdV-5), and many of the alternative splice sites are not conserved even among the human adenovirus species. Indeed, an overview of mastadenovirus genomes now forces us to question whether these evolutionarily ephemeral genes have meaningful roles during the natural life cycle. On this basis, we have omitted from Fig. 2 any coding regions specified by alternative splicing.

Splicing is not universal in adenoviruses: in the human viruses, E1B 19K, E1B 55K and IX are single exon genes. Nonetheless, a pattern in which a common leader is spliced to any one of several downstream protein-encoding exons is evident in the late, E2, E3 and E4 regions. This simple layout, termed here the ‘leader–exon structure’, facilitates the evolution of new genes by capture or duplication, merely requiring the presence of a splice acceptor site upstream from a newly inserted protein-encoding region. This strategy has apparently functioned throughout adenovirus evolution and is considered in further detail below. As far as can be deduced from sequence comparisons, the general splicing patterns of the late and E2 genes is common to all adenoviruses. Thus, late genes are spliced from an untranslated leader, whether tripartite, as in mastadenoviruses and atadenoviruses (Khatri & Both, 1998), or bipartite, as in av adenoviruses (Sheppard et al., 1998). Also, in all adenoviruses, the initiation codon for pTP is located in a short exon between pIIIA and III (Fig. 2). This initiation codon is also utilized for expression of pol in the mastadenoviruses, but apparently not in the other genera, where the short exon is probably still spliced to the main coding exon, but pol translation is initiated in the main exon. These differences reflect minor evolutionary tinkering; in this case, the change of register in splice sites to allow the leader in the leader–exon structure to switch between being non-coding and coding. Whichever situation pertains, the structure lends itself well to evolution of new genes. Differences between the genera are also evident in splicing of DBP. This gene is probably spliced from upstream non-coding exons in all adenoviruses, but in av adenoviruses and siadenoviruses it has accrued an additional coding exon upstream from the main exon.

Genome sequences

We conducted a thorough analysis of complete adenovirus genome sequences deposited in GenBank (http://www.ncbi.nlm.nih.gov/), utilizing current knowledge of adenovirus gene content. The sequences are derived from 23 serotypes encompassed by the four genera recognized, and are listed in Table 1. Our conclusions on the status of the entries are summarized in Table 2. Shortcomings in interpretation and annotation are obvious in several respects, and often belie a much higher quality of analysis in the primary publications. Firstly, annotations may be incomplete and in some cases totally absent. Secondly, the coordinates of some protein-encoding regions may be incorrect owing to neglect of comparative data or splicing. Thirdly, a sizeable subset of genes may exhibit frameshifts, most, if not all, due probably to sequencing errors. Indeed, several errors in the HAdV-5 sequence were demonstrated experimentally by Dix & Leppard (1992) and have been confirmed by us. Errors outside protein-encoding regions may also be apparent, for example, in differences between the supposedly identical ITRs of TSAdV-1 and also of DAdV-1. Fourthly, updated analysis may reveal previously unrecognized genes, some with unequivocal cellular counterparts. New examples from this category are incorporated into the discussion below.

These findings imply that access to reliable adenovirus genome data is compromised. As a result, the amino acid sequences of many adenovirus proteins are unavailable, and some of those that are available are incorrect. This situation may be mitigated by separate and more extensive annotation of fragments of certain genomes. However, the extent of this amelioration is partial, uneven and laborious to ascertain for individual viruses. For example, although GenBank contains entries for only eight HAdV-5 proteins derived from the
Table 2. Accuracy and annotation status of complete adenovirus genome accessions

Cells lacking black borders indicate regions where orthology is difficult to establish and is therefore not implied. This is because the genes therein are highly variable in number and sequence and in some cases belong to families of related genes. There is little or no similarity between certain primate and non-primate mastadenovirus genes in E3 and E4, and the nomenclature shown applies only to the former. The nomenclature of E3 CR1 genes, which are present only in primate mastadenoviruses, is based on order in the genome.
genome sequence, entries for the majority of proteins are accessible via annotations of genome fragments. In contrast, amino acid sequences are accessible only for the six genes that are annotated in the complete TSAdV-1 genome, as entries for fragments provide no additional information. This evaluation of the dismal condition of adenovirus genomics also applies to entries in the GenBank Reference Sequence library (RefSeq, http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/viruses.html), which is intended as a means of avoiding fossilization of sequence interpretations. These entries are presently very similar to the originals, though some have been updated sparingly and with varying success by the gene prediction program GENEMARK (http://opal.biology.gatech.edu/GeneMark/) (Borodovsky & McIninch, 1993). Accession numbers for our annotations of adenovirus genomes are listed in Table 1, and are available under the GenBank Third Party Annotation scheme (TPA, http://www.ncbi.nlm.nih.gov/Genbank/tpa.html). Although questions remain concerning expression of certain regions, we consider that these entries are the best available at present. We conclude this section by noting that the lack of a readily updated data resource is hindering systematic advancement of comparative adenovirology.

The nucleotide composition properties of adenovirus genomes have attracted interest, with a focus on the atadenoviruses (Benkő & Harrach, 1998; Farkas et al., 2002). Wide variation in nucleotide composition is characteristic of many groups of organisms, including adenoviruses, but the driving evolutionary forces remain elusive. Davison et al. (2000) highlighted the fact that certain adenovirus genomes (mastadenoviruses CAdV-1 and CAdV-2, atadenovirus OAdV-7 and siadenoviruses TAdV-3 and FrAdV-1) are depleted throughout in the CG dinucleotide. In cellular genomes, this evolutionary phenomenon is thought to result from methylation of the cytosine residue in the CG dinucleotide, followed by spontaneous deamination to TG and fixation by DNA replication. CG depletion is also a feature of certain herpesviruses and has been attributed to methylation of latent genomes resident in dividing cell populations (Honess et al., 1989). This parallel yields provocative indicators to unexplored aspects of adenovirus biology.

**Genus-common genes**

The principal means of assigning genes across the *Adenoviridae* is comparative analysis, which involves identifying conserved protein-encoding regions. Broadly, we class genes with homologues in all genera as 'genus-common genes' and all others as 'genus-specific genes'.

There are 16 clearly defined genus-common genes. We presume that these were inherited from a common ancestral adenovirus, in which they were all expressed by splicing. Their primary functions are in DNA replication (pol, pTP and DBP), DNA encapsidation (52K and IVa2) and formation and structure of the virion (pIIia, III, pVII, pX, pVI, hexon, protease, 100K, 33K, pVIII and fiber). Two additional protein-encoding regions may be added. One, 22K, originates from lack of splicing in 33K, and thus the N-terminal sequence of the protein is identical to that of 33K but the C-terminal sequence is encoded by the 33K intron. However, the putative 22K-encoding region of atadenoviruses does not extend through the intron, and thus its inclusion for this genus is tentative. The second addition is the U exon, which extends from an initiation codon to a splice donor site and is regulated by a minor late promoter (Chow et al., 1979; Davison et al., 1993, 2000). The downstream exons spliced to this exon have not yet been identified. They may correspond to IVa2, pol, pTP or DBP and yield differentially spliced forms of one or more of these genes encoding N-terminally extended proteins. The U exon appears to be a genus-common feature that has been lost in certain mastadenoviruses (PaDv-5 and MaDv-1).

Candidate splice sites for the majority of genus-common genes may be predicted from DNA sequences. However, particular uncertainty should be expressed about two. One gene, 33K, is poorly conserved among genera and identification of the splice sites is tentative in some cases. Also, the first exon is not visible in DaDv-1. The second problem concerns IVa2, which is involved in transcriptional activation of the late promoter and in capsid assembly and DNA packaging (Zhang & Imperiale, 2003). In mastadenoviruses, conservation of the region upstream from the first ATG codon supports splicing from an upstream protein-encoding exon. This is in accord with data that demonstrate splicing of the main coding region from a short leader close upstream, such that the five N-terminal residues of IVa2 correspond to residues within pol (Baker & Ziff, 1981; Reddy et al., 1998a). However, the splice sites are not conserved throughout the mastadenoviruses. In atadenoviruses, mapping data for OAdV-7 indicate that IVa2 is not spliced (Khatri & Both, 1998), but this is at variance with conservation of the region upstream from the initiation codon, and again points to the existence of an upstream coding exon. We conclude that the transcriptional pattern of IVa2, which is likely to have been inherited by all genera, is worth further investigation.

Several adenovirus proteins (genus-common pTP, pIIia, pVII, pX, pVI and pVIII, and genus-specific p32K of atadenoviruses) are cleaved by the protease in steps that are essential for the synthesis of infectious virus (Weber, 1995). Reported consensus cleavage sites may be summarized as (M/L/I)XG’X and (M/L/I/N/Q)X(A/G)X’G (Anderson, 1990; Farkas et al., 2002; Ruzindana-Umumyana et al., 2002; Vrati et al., 1996; Webster et al., 1989). Since the protease is a genus-common gene, proteolytic maturation is evidently an ancient feature of the adenovirus replicative cycle. As an example, Fig. 3 shows putative cleavage sites for all of the sequenced viruses in pX, which, along with pVII and V, generates the core proteins of the virion. The primary translation product consists of an extended, basic N-terminal region linked to a short, basic C-terminal region via a highly conserved hydrophobic domain. This protein is cleaved by the protease between the basic N-terminal and
hydrophobic domains (and in some viruses at a second site nearer the N terminus) to give rise to a core protein (X or m) that is closely associated with the virion DNA. An incidental, but provocative, observation is that unprocessed pX in all adenoviruses has the primary sequence features of a class II membrane protein.

Genus-specific genes

Most genus-specific genes are located near the ends of the genome. These regions are termed E1 and E4 for human adenoviruses, and, despite the general lack of genetic similarity between genera in these regions, we utilize this nomenclature for all adenoviruses. In mastadenoviruses and siadenoviruses, genus-specific genes are also located in the E3 region, between pVIII and the U exon, and, in addition, the mastadenoviruses contain a single genus-specific gene (V) between pVII and pX. This general location pattern for genus-specific genes is not restricted to adenoviruses. It is also a feature of other linear, double-stranded DNA genomes, such as the Herpesviridae (McGeoch & Davison, 1999) and Poxviridae (Upton et al., 2003), which consist of centrally located genus-common genes with most genus-specific genes located terminally. In these families, many genus- and virus-specific genes are involved in interactions with the host, presumably to promote survival in relevant biological niches, and a number have been captured from the host. It is interesting to note that eukaryotes, as represented by yeast, exhibit rapid evolution in the telomeric regions of chromosomes (Kellis et al., 2003).

Gene capture has played an important role throughout adenovirus evolution. Thus, genus-common genes, such as pol, IVa2 (which contains an ATP-binding domain) and protease, possibly resulted from very ancient capture events.

Table 3 lists adenovirus genes with homologues in cellular or other viral genomes that have been captured more recently, after the genera diverged. Genes with cellular counterparts are taken to represent imports into adenoviruses, but those with counterparts only in other viruses

Fig. 3. Alignment of the complete predicted primary amino acid sequences of pX from the four adenovirus genera recognized. The overall alignment is centred on the conserved putative transmembrane domain, shown in blue. Predicted or confirmed protease cleavage sites are indicated by red vertical lines. Sites conform to (M/L/I/V/F)XGGX and (M/L/I/V/N/Q)X(A/G)XG, and are slightly less specific at the first residue than those described in the literature [(M/L/I)XGGX and (M/L/I/N/Q)X(A/G)XG].

2902

A. J. Davison, M. Benkő and B. Harrach

Journal of General Virology

84
could have been transferred from virus to virus in either direction. Whether discernable as captured or not, genus-specific genes range from two (E1B 55K and E4 34K) whose development pre-dated divergence of mastadenoviruses and atadenoviruses, through genes that characterize all viruses in a genus (e.g. E1A in mastadenoviruses and p32K in atadenoviruses) to genes that are specific to a subset of viruses, or even a single virus, within a genus (e.g. E3 12.5K in mastadenoviruses). This complexity in definition is not unexpected, since genes could be gained or lost at any stage. The processes of gene duplication, divergence and functional partition have clearly been at work among genus-specific genes. This is an evolutionary mechanism used widely by larger DNA viruses for generating new genes. Several examples in adenoviruses are listed in Table 3, and include CR1 genes in human mastadenoviruses, the dUTPase gene in mastadenoviruses, 34K in mastadenoviruses and atadenoviruses, and RH genes in aviadeno-viruses. The involvement of gene duplication in the leader–exon setting of the E3 and E4 regions is discussed below and highlights the utility of this evolutionary strategy. Relatively recent duplications are also apparent in genus-common genes. Examples include fiber in FAdV-1 and HAdV-F and DBP in TSAdV-1. Indeed, PAdV-3 and FrAdV-1 exhibit very recent perfect duplications of sequences near the right end of the genome, which, although possibly a result of passage in cell culture and therefore not of functional or evolutionary significance, nonetheless attest to the ease with which duplications occur. Given the leader–exon structure of late genes, it is tempting to speculate that some of these evolved by ancient duplication events but that any evidence in the form of amino acid sequence similarity has been obliterated over time.

Evolution of E4 in the mastadenoviruses appears particularly complex. In this region, precise genetic relationships are readily catalogued between human adenovirus species, but in other mastadenoviruses the relationships are difficult to assess, although the basic structure and expression pattern appear to be conserved. As a result, genetic relationships between non-primate mastadenovirus E4 genes are not specified in Table 2. In five of the human adenovirus species, E4 contains six leftward-oriented genes (ORF1, ORF2, ORF3, ORF4, 34K and ORF6/7), transcripts from which are regulated by a promoter near the right end of the genome. ORF6/7 mRNA results from further splicing between the 5' end of 34K and the region immediately downstream of 34K, and thus the ORF6/7 and 34K proteins share N-terminal sequences. HAdV-F is unusual in lacking a counterpart to ORF1, presumably as a result of deletion. ORF1 appears to be derived from a captured dUTPase gene, but its descendents in human adenoviruses have not retained the active site residues and presumably carry out other functions (Weiss et al., 1997). Evolution of E4 in mastadenoviruses has evidently involved duplication or deletion events resulting in variable numbers of genes. Occasionally, alternative routes of functional partition for duplicated 34K or dUTPase-derived genes have led to a situation where sequence similarity between two viruses is greatest between ORFs that do not correspond in location. In atadenoviruses, knowledge of E4 is less extensive, but splicing in leader–exon fashion and gene duplication are evident. Two tandem 34K genes (34K-1 and 34K-2) are present, and the RH family of duplicated genes (termed ORF8 and ORF9 in DAdV-1) encoding F-box proteins has developed upstream (Both, 2002b). Two related genes (ORF5 and ORF6) resulting from duplication are present at the right end of the DAdV-1 genome.

Table 3. Genus-specific genes captured during adenovirus evolution

<table>
<thead>
<tr>
<th>Virus</th>
<th>Gene</th>
<th>Homologue</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>MASTADENOVIRUS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Most</td>
<td>E4 ORF1</td>
<td>Cellular dUTPase</td>
<td>Duplicated, active site motifs retained in some</td>
</tr>
<tr>
<td>Primate</td>
<td>E3 CR1</td>
<td>Primate cytomegalovirus CR1 protein</td>
<td>Duplicated</td>
</tr>
<tr>
<td>ATADENOVIRUS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>p32K</td>
<td>Bacterial small acid-soluble protein</td>
<td>Tenuous relationship</td>
</tr>
<tr>
<td>All</td>
<td>RH</td>
<td>Cellular F-box protein</td>
<td>Duplicated</td>
</tr>
<tr>
<td>DAdV-1</td>
<td>ORF4</td>
<td>Cellular OX-2 glycoprotein</td>
<td></td>
</tr>
<tr>
<td>DAdV-1</td>
<td>ORF7</td>
<td>Cellular RING-finger protein</td>
<td></td>
</tr>
<tr>
<td>AVIADENOVIRUS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>ORF1</td>
<td>Cellular dUTPase</td>
<td>Active site motifs retained</td>
</tr>
<tr>
<td>All</td>
<td>ORF2</td>
<td>Parvovirus Rep</td>
<td>Duplicated</td>
</tr>
<tr>
<td>All</td>
<td>ORF11</td>
<td>Cellular leukocyte adhesion molecule</td>
<td>Duplicated</td>
</tr>
<tr>
<td>All</td>
<td>ORF19</td>
<td>Putative avian herpesvirus lipase</td>
<td></td>
</tr>
<tr>
<td>FAdV-1</td>
<td>ORF26</td>
<td>Cellular C-type lectin</td>
<td></td>
</tr>
<tr>
<td>FAdV-10</td>
<td>ORF4</td>
<td>Avian herpesvirus US22 protein</td>
<td></td>
</tr>
<tr>
<td>SIADENOVIRUS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>sialidase</td>
<td>Cellular sialidase</td>
<td></td>
</tr>
</tbody>
</table>
These observations lead to a speculative model for the evolution of E4 in mastadenoviruses and atadenoviruses in which 34K was present in the common ancestor and other genus-specific genes were inserted nearer to the right genome terminus. These included a dUTPase gene in the mastadenovirus lineage and an F-box gene in the atadenovirus lineage. The 34K, dUTPase and F-box genes then proliferated at various stages by duplication, divergence and functional partition, and sometimes by deletion, to give rise, in concert with development of unrelated additional genes, to the E4 formats observed. The dUTPase-derived proteins in modern mastadenoviruses exhibit a range of degrees of relatedness. Most are probably not active dUTPases, but a functional enzyme appears to have been retained in the lineages leading to TSAdV-1 and PAdV-5/BAdV-2. Other clues of gene duplication in E4 also remain, such as in two 34K-derived genes in PAdV-5 and in three genes in TSAdV-1 that are related to ORF2 of primate mastadenoviruses. However, we suspect that divergence of duplicated genes may have proceeded so far in many instances as to obliterate evidence of common origin. It is also interesting to note that duplication of a dUTPase gene has occurred during herpesvirus evolution (McGeoch, 1990), where it is postulated that the initial duplication event resulted in a fused, double-length gene encoding two sets of active site residues. This was followed by loss of one active site to yield a still-active enzyme, and in one lineage by loss of the remaining active site to end in a protein that is presumably no longer a dUTPase (McGeoch & Davison, 1999).

In atadenoviruses, E4 is substantially larger than in other genera. Understanding of the gene content is incomplete, but useful information on transcription of the FAdV-1 and FAdV-9 genomes (Ojkic et al., 2002; Payet et al., 1998) has aided our extension of previous analyses. It is clear that splicing is common and that gene capture, duplication and functional partition have occurred. A family of putative glycoprotein genes is located near the right genome end in FAdV-1 and FAdV-9. It consists of three rightward-oriented genes (ORF9, ORF10 and ORF11) in the former and one rightward- (ORF11) and one leftward-oriented gene (ORF23) in the latter. The origin of this gene family is revealed by the clear similarity between the ORF11 protein from each virus and cellular leukocyte adhesion molecules. In the absence of obvious cellular forerunners, two observations currently imply that genetic exchange may also have occurred between avian adenoviruses and herpesviruses. Firstly, the ORF19 proteins in FAdV-1 and FAdV-9 are most closely related to membrane proteins (putative lipases) encoded by members of the genus Mardivirus, an avian lineage of the Herpesviridae (subfamily Alphaherpesvirinae) that includes Marek’s disease virus (Ojkic & Nagy, 2000). Secondly, partial sequence data indicate that the gene (ORF4) at the right end of the FAdV-10 genome (AF160185) lacks counterparts in FAdV-1 and FAdV-9, but has relatives in several other viruses (Davison et al., 2003). ORF4 is predicted to be spliced, and has closest relatives again in mardiviruses. Distant relatives of this gene are detectable in two other lineages of the Herpesviridae (subfamily Betaherpesvirinae and an amphibian herpesvirus) and in one lineage of the Poxviridae (fowlpox virus).

Gene capture and duplication also feature in E1. An atadenovirus protein (p32K) bears a tenuous relationship to small acid-soluble proteins of bacteria (Elö, et al., 2003), and the siadenoviruses are so named because they encode a putative sialidase (Davison & Harrach, 2002). The aviadeno-viruses have a dUTPase (ORF1), which retains active site residues, plus a parovirus Rep protein (ORF2) (Chiocca et al., 1996). Although Chiocca et al. (1996) speculated that the dUTPase genes of mastadenoviruses and aviadeno-viruses may have evolved via transfer of an ancestral gene from one genome terminus to the other, it is equally likely that they were acquired by separate capture events. Duplicates of aviadenvirus ORF2 appear to be present on the opposing strand (ORF12 and ORF13), and are expressed by splicing from the short coding exon utilized by pTP.

In contrast to E1 and E4, which are present in all genera, E3 features only in mastadenoviruses and siadenoviruses. This location for genus-specific genes may have been arrived at independently, or may represent an ancient locus for rapid gene evolution that has been lost in two genera. A substantial set of genes, numbering up to nine, has evolved in mastadenovirus E3. The two genes at the ends of this block (12-5K and 14-7K) encode distantly related proteins that probably arose via gene duplication at an early stage in mastadenovirus evolution with subsequent loss in some lineages. The intervening genes encode membrane proteins, certain of which (CR1 genes) in primate adenoviruses have relatives in primate cytomegaloviruses (a lineage of subfamily Betaherpesvirinae in the Herpesviridae). CR1 genes share a common motif but are highly variable in number and sequence in both virus families (Davison et al., 2003). Interestingly, TSAdV-1 lacks CR1 genes in E3, but the cognate domain is present in the putative glycoprotein encoded by a gene (105R-T), which is unique to this virus and situated at the right end of the genome. Both (2002b) detected a weak similarity between the single siadenovirus E3 protein and RH5 of atadenoviruses OAdV-7 and BAdV-4, but the evolutionary significance of this is not clear.

In the genomes of most primate mastadenovirus species (including all human and chimpanzee viruses), the region between the pTP and 52K genes contains one or two rightward-oriented VA RNA genes, which are transcribed by RNA polymerase III (pol III) (Kidd et al., 1995; Ma & Mathews, 1996; Mathews & Shenk, 1991). The encoded partially double-stranded RNAs are approximately 160 nucleotides in size, and are involved in translational control and inhibition of the interferon response (Mathews & Shenk, 1991; Mori et al., 1996). A 90 nucleotide pol III RNA is also produced by a leftward-oriented gene located between ORF16 and ORF9 in FAdV-1 (Larsson et al., 1986). Thus, acquisition of VA RNA genes is likely to have occurred at least twice during adenovirus evolution, once in primate mastadenoviruses (perhaps from a tRNA gene,
followed by gene duplication in some lineages) and once in aviadeno-viruses. On the basis of limited sequence similarity to FAdV-1, a candidate VA RNA gene was identified in the DAdV-1 genome, overlapping the 3′ end of ORF4 (Hess et al., 1997). This region is absent from the other two atadenoviruses sequenced, and functional investigations have failed to detect a VA RNA in OAdV-7 (Venktesh et al., 1998). If the identified DAdV-1 VA RNA gene is genuine, its evolution might have been independent or might have involved transfer to or from an aviadeno-virus.

**Evolution**

Similarities between phylogenetic trees for adenovirus protease and the small subunit of host mitochondrial rRNA led to the conclusion that adenoviruses have largely co-speciated with their hosts (Benkó & Harrach, 2003). Mastadenoviruses infect mammalian hosts exclusively, and aviadeno-viruses have been found only in birds. The picture becomes blurred for the other two genera. Atadenoviruses have been identified in distantly related hosts, such as various species of poultry, a variety of domestic and wild ruminants, and a marsupial. More recently, atadenoviruses have been shown to be present in snakes and lizards, in support of a reptilian source for this genus (Harrach, 2000). The presence of atadenoviruses in birds and mammals could be explained by a couple of (supposedly independent) host switching events. Although one of the siadenoviruses infects a bird, it is tempting to speculate that this genus corresponds to the original amphibian lineage. The partial genome sequence of a fish adenovirus implies a fifth genus (Fig. 1), thus reinforcing the idea that major vertebrate lineages are characterized by distinct adenovirus genera (Kovács et al., 2003).

The most recent ancestor of all modern adenoviruses is likely have been an adenovirus that existed before the divergence of bony fish from other vertebrates. This virus possessed a substantial prior evolutionary history. Since adenoviruses of invertebrates have not yet been discovered, the characteristics of adenoviruses predating vertebrates are unknown. Nonetheless, tantalizing glimpses may be had into earlier epochs. There are clear similarities in overall architecture of the virion and in the structure of its principal protein component (hexon) between adenoviruses and bacteriophage PRD1 (Belnap & Steven, 2000; Benson et al., 1999, 2002; San Martin & Burnett, 2003). PRD1 belongs to the family Tectiviridae, infects Gram-negative hosts and has a linear, double-stranded DNA genome of 15 kbp that is linked to a TP (Bamford, 2002; Bamford & Ackermann, 2000). Adenovirus hexon possesses the ‘jellyroll fold’ common to capsid proteins of many viruses (Chelvanayagam et al., 1992), but the structural relationship is closest to that of PRD1. Parallels also extend to other features, such as fiber and its PRD1 counterpart (spike), which are present at the virion vertices. These findings provide evidence for divergent evolution of adenoviruses and tectiviruses from an ancestor that pre-dated eukaryotes and exhibited the adenovirus morphology. The ancient origins of this morphology are highlighted further by recent findings on bacteriophage Bam35, which infects a Gram-positive host. PRD1 and Bam35 are morphologically indistinguishable and share the hexon fold, even though they are thought to have diverged over a billion years ago (Ravantti et al., 2003).

Protein-primed DNA replication is employed by adenoviruses and certain bacteriophages (Berencsi et al., 1995; de Jong et al., 2003; Liu et al., 2003; Salas, 1991). The components are a linear double-stranded DNA template with ITRs, pol, TP and at least one DBP. This strategy is used by PRD1 (Bamford et al., 1991) and by bacteriophage φ29 (19 kbp), which belongs to the Podoviridae and has a different capsid morphology from PRD1 (Pečenková & Pačes, 1999). Moreover, in adenoviruses and these two bacteriophages, TP and pol are early genes arranged in tandem near the left genome end, and the late genes are located more centrally, adjacent to TP and pol and arrayed rightward. The antiquity of protein-primed DNA replication is further underscored by its occurrence in linear plasmids of fungi and plants, which generally encode the required pol and TP (Paillard et al., 1985; Rohe et al., 1991), and by linear bacterial genomes of Streptomyces species (Chen, 1996).

In summary, we may speculate that the protein-primed DNA replication strategy originated at a very early stage in evolution, to be followed by acquisition of the adenovirus morphology during a pre-eukaryotic era. Introduction of splicing and additional replicative and structural genes resulted in an adenovirus from which extant members of the family have inherited at least 16 genes. Subsequent lineages developed specific subsets of genes that fit them to particular biological niches.

**ACKNOWLEDGEMENTS**

This work was supported by the Royal Society, the Hungarian Scientific Research Fund (OTKA T043422), the Hungarian Prime Minister’s Office (MEH 4676/1/2003), and the Medical Research Council. We thank Duncan McGeoch for critical reading of the manuscript.

**REFERENCES**


First molecular evidence for the existence of


Cell

messenger RNA.

J Virol

organization of the avian adenovirus CELO.

42

2906

Journal of General Virology


http://vir.sgmjournals.org


