BMC Bioinformatics

BioMed Central

Research article

Reannotation of the CELO genome characterizes a set of previously unassigned open reading frames and points to novel modes of host interaction in avian adenoviruses

Stefan Washietl[1,2] and Frank Eisenhaber*[1]

Address: [1]Research Institute of Molecular Pathology, Dr. Bohrgasse 7, A-1030 Vienna, Austria and [2]Current address: Institute for Theoretical Chemistry and Structural Biology, University of Vienna, Waehringerstrasse 17, A-1090 Vienna, Austria

Email: Stefan Washietl - wash@tbi.univie.ac.at; Frank Eisenhaber* - Frank.Eisenhaber@imp.univie.ac.at

* Corresponding author

Published: 07 November 2003

BMC Bioinformatics 2003, **4**:55

Received: 02 September 2003

Accepted: 07 November 2003

This article is available from: http://www.biomedcentral.com/1471-2105/4/55

Abstract

Background: The genome of the avian adenovirus Chicken Embryo Lethal Orphan (CELO) has two terminal regions without detectable homology in mammalian adenoviruses that are left without annotation in the initial analysis. Since adenoviruses have been a rich source of new insights into molecular cell biology and practical applications of CELO as gene a delivery vector are being considered, this genome appeared worth revisiting. We conducted a systematic reannotation and in-depth sequence analysis of the CELO genome.

Results: We describe a strongly diverged paralogous cluster including ORF-2, ORF-12, ORF-13, and ORF-14 with an ATPase/helicase domain most likely acquired from adeno-associated parvoviruses. None of these ORFs appear to have retained ATPase/helicase function and alternative functions (e.g. modulation of gene expression during the early life-cycle) must be considered in an adenoviral context. Further, we identified a cluster of three putative type-1-transmembrane glycoproteins with IG-like domains (ORF-9, ORF-10, ORF-11) which are good candidates to substitute for the missing immunomodulatory functions of mammalian adenoviruses. ORF-16 (located directly adjacent) displays distant homology to vertebrate mono-ADP-ribosyltransferases. Members of this family are known to be involved in immuno-regulation and similiar functions during CELO life cycle can be considered for this ORF. Finally, we describe a putative triglyceride lipase (merged ORF-18/19) with additional domains, which can be expected to have specific roles during the infection of birds, since they are unique to avian adenoviruses and Marek's disease-like viruses, a group of pathogenic avian herpesviruses.

Conclusions: We could characterize most of the previously unassigned ORFs pointing to functions in host-virus interaction. The results provide new directives for rationally designed experiments.

Background

Chicken embryo lethal orphan virus (CELO) is an adenovirus infecting avian species [1,2]. It is a member of the genus *Aviadenovirus* and also referred to as Fowl Adenovirus 1 (FAdV-1). Compared to mammalian and, in particular, human adenoviruses of the genus *Mastadenovirus*, which have been studied extensively over the years (reviewed in [3]), relatively little information is available

on avian adenoviruses. In 1996, CELO was the first virus of this group to be completely sequenced [4].

The analysis of the sequence revealed that the central portion of the 43.8 kb long, double-stranded, linear DNA genome is organized similar to mammalian adenoviruses. Genes for the major structural proteins (e.g. IIIa, hexon, penton base) as well as crucial functional proteins (e.g. DNA-polymerase, protease) are well conserved with respect to amino acid sequence and location. However, the important E1A, E1B, E3 and E4 regions, mainly responsible for host cell interaction and immune modulation/evasion in mammalian adenoviruses, could not be identified. Instead, two unique terminal regions of about 6 kb and 12 kb rich in open reading frames with no homologs in mammalian adenoviruses could be found. This surprising result suggests that the basic properties of the replication cycle are similar in both groups whereas they encode a completely different set of proteins for host interaction. Only a few of these proteins have been functionally characterized so far.

ORF-1 is significantly homologous to dUTP-pyrophosphatases and was reported to have this enzymatic activity [4]. ORF-1 is the only sequence in the terminal regions which has homologs in mastadenoviruses (ORF-1 of early region 4). In human adenovirus 9, this protein has growth-transforming properties and is an important oncogenic determinant [5].

ORF-8, which has been designated Gam1, is probably the most intriguing protein found in CELO. Originally identified as a novel antiapoptotic protein [6] and further shown to induce heat shock response necessary for replication [7], it is now known to influence host gene expression by inactivation of histone deacetylase 1 [4,8,9]. Together with another unique protein (ORF-22), Gam1 influences also the pRb/E2F pathway crucial for cell-cycle progression. Both proteins bind pRb and, thus, act as functional analogs of the prominent adenoviral E1A protein [10].

For the rest of the unique ORFs, experimental data is sparse if available at all. Mutational studies found most of them to be dispensable for viral replication under different experimental settings [11,12]. In an attempt to characterize the transcriptional organisation of CELO, the corresponding RNAs for some of the ORFs together with their expression kinetics could be identified [13]. However, the functions of these proteins during the viral life cycle are still completely unknown. Since they are thought to be implicated in such critical areas of biology as for example cell cycle control and immune response to viral infections, these proteins are of special interest. Moreover, CELO has been considered for use as a gene delivery vector with promising features for both human gene therapy and vaccination applications in aviculture [11,12,14]. A better understanding of CELO biology could help to promote such applications.

In this contribution, we report a complete, systematic, in-depth sequence analysis of all potential coding sequences in the CELO genome. Applying a relevant subset of the most advanced analyzing methods available at present, we determined the molecular architecture of the putative proteins and uncovered distant homologies, evolutionary relationships and possible molecular and cellular functions. If available, we also analyzed homologous sequences of closely related avian adenoviruses. These are (i) Fowl Adenovirus 9 (FAdV-9, formerly known in literature as FAdV-8) [15-17], (ii) strain CFA40, a hypervirulent variant of FAdV-9 [18] and (iii) FAdV-10. For FAdV-9, the complete genomic sequence is available, for CFA40 and FAdV-10 only fragments of the nucleic acid sequence are known. We anticipate that our results will stimulate experimental studies of CELO ORFs with newly assigned molecular and/or cellular functions.

Results

Refinement and analysis of potential coding regions

The complete CELO sequence has been analyzed upon its initial sequencing [4]. In the central region ranging from approximately nt 6000 to 31000, most of the ORFs could be reliably assigned to proteins that have been previously described for mastadenoviruses. In the terminal regions (appr. nt 0–6000 and 31000–43804) no sequence similarity to known adenoviral sequences could be detected on the nucleic acid or protein level. Originally, 22 potential protein coding sequences were proposed to reside in the unique terminal regions [4]. They have found their way into public databases and are referred to throughout literature. Those putative proteins are exclusively ORFs which are longer than 99 amino acids and start with a methionine. This is a rather arbitrary approach and, since also the experimental studies fall short in detecting and characterizing all RNAs of these regions [13], we had to refine the prediction of protein coding regions in order not to miss important information due to wrong conceptual translations. We did a complete retranslation of the genome in all six frames also considering ORFs shorter than 99 amino acids and without a starting methionine, we further compared the potential coding regions to the related avian adenoviruses, especially to the complete genome of FAdV-9, and integrated all available experimental data [13,15-17] as well as the results of our subsequent protein sequence analysis. Table 1 and Fig. 1 list the most likely coding regions that could be identified. If possible, we adhere to the nomenclature introduced by Chiocca et al. [4].

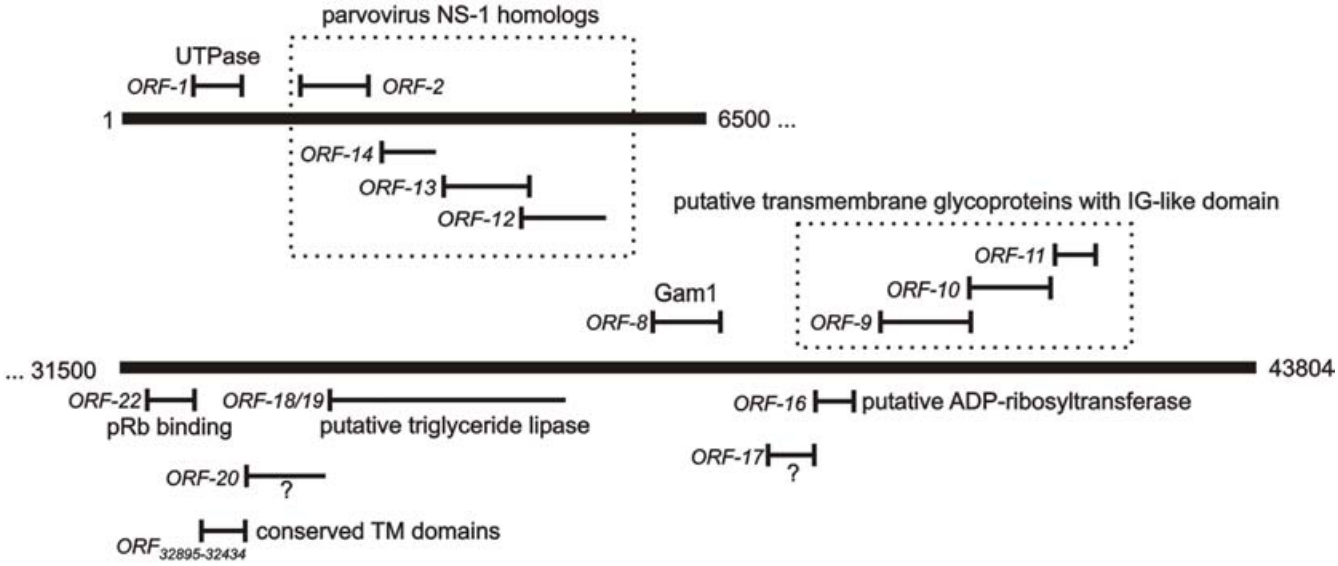

Figure 1
Coding regions in the terminal segments of the CELO genome. The 15 ORFs listed in Table 1, representing the most likely protein coding regions, are indicated. ORFs being transcribed from the forward and reverse strand are shown above or below the bold line representing the double-stranded DNA, respectively. Open lines denote ORFs without a start codon in the genomic sequence. ORF-1, ORF-8 and ORF-22 are annotated based on experimental results. The detailed annotation and results of the sequence analysis for all other ORFs are described in the text and Fig. 3.

In four cases (ORF-12, ORF-14, ORF-20, ORF-18/19) the translation of the ORFs was extended in the amino terminus mainly because of significant similarity to homologous sequences in FAdV-9 and CFA40 or the existence of known domains in this extended region. ORF-18 and ORF-19 were merged to one single ORF-18/19 for reasons detailed in the discussion below.

Furthermore, we could find two new ORFs. $ORF_{28115-27765}$ is not located in the terminal regions but is located between the fibre and pVIII gene and was, therefore, not described and numbered by Chiocca et al. Since it is conserved in CELO, FAdV-9, CFA40 and FAdV-10 but unique to this group, it was of special interest for this study. It is noteworthy that this is the only unique ORF in the central portion of the genome, all others are exclusively found in the terminal regions.

We further identified $ORF_{32895-32434}$, which overlaps with ORF-21 in a different frame. Since $ORF_{32895-32434}$ has homologous sequences in FAdV-9 and CFA40, it appears more likely to be expressed than the originally described ORF-21.

Also some other originally described ORFs overlap with each other (e.g. ORF-3 with ORF-13 or ORF-7 with ORF-18/19). In adenoviruses, genes usually do not overlap and it is unlikely that heavy usage of overlapping genes does occur in CELO. It can be rather expected that, if two or more ORFs overlap in substantial parts of their coding sequence, only one ORF is expressed. After our analysis, we propose that the originally described ORF-3,4,5,6,7,15,21 do not code for proteins because (i) there are no homologs in the closely related avian adenoviruses or in other viruses/organisms, (ii) sequence analysis did not yield reasonable protein features, (iii) no corresponding transcript could be experimentally detected [13] (iv) they overlap with alternative ORFs that meet most of these criteria.

Taken together, we have to expect that the CELO genome has at least 15 ORFs of functional importance without homologs in mammalian adenoviruses. The amino acid sequences of all the ORFs can be found together with homologous sequences from related avian adenoviruses on our website http://mendel.imp.univie.ac.at/SEQUENCES/CELO/. All these sequences were subject of an in-depth sequence analysis. The general strategy that was used is outlined in Fig. 2 and the major results are summarized below.

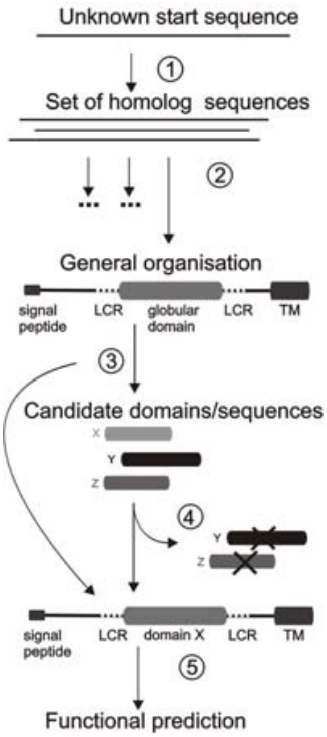

Figure 2
Outline of the analysis process illustrating basic steps from an unknown protein sequence towards a functional interpretation. (1) Starting with the unknown CELO sequence, significantly homologous sequences featuring relatively high identity/similarity are searched. Usually, only sequences from related avian adenoviruses could be found at this step. This results in a set of homologous proteins likely to have the same or at least similar function. The following steps are carried out for each of these sequences. This comparative approach can bring up additional information which might be missed if only one sequence is analyzed. (2) Intrinsic sequence features are investigated. This includes a statistical analysis of amino acid contents, the search for low complexity regions (LCRs), coiled coil domains, transmembrane domains (TM), amino- and carboxy-terminal signal sequences and internal repeats. An important output of this step is the rough discrimination between globular and non-globular regions in the protein. (3) The globular regions are further analyzed. These domains present the most useful level on which to understand protein function and their identification is, therefore, one of the major issues during the whole analysis process. Comparison to different databases using various algorithms (see Material and Methods) can either find significant homologs, or proposes a set of candidate domains with borderline statistical significance. In the latter case (4), those hits must be further verified or excluded be additional investigations (conservation of critical functional or structural residues, secondary structure prediction, fold recognition, consensus of different methods, consensus of prediction results within the group of close homologs,...). (5) Finally, all the results are integrated and can be interpreted in the context of the CELO infection cycle.

ORF-2, ORF-12, ORF-13: homologs of parvovirus non-structural proteins with an inactive ATPase/helicase domain

In ORF-2, homology to parvoviral non-structural proteins is significant and was noted previously [4]. ORF-2 is most similar to Rep78 of adeno associated virus (AAV) 3B (BLASTP expectation value: $E = 8 \cdot 10^{-32}$) and is already member of the corresponding Pfam family (Parvovirus non-structural protein NS1: PF01057). This domain spans the complete sequence of ORF-2 (Fig. 3a). We also found that ORF-12 and ORF-13 are distantly related to this family of proteins. A PSI-BLAST search with inclusion threshold 0.05 was initiated with ORF-12. After the first run, only the FAdV-9 homolog $ORF_{6190\text{-}5243}$ was found ($E = 8 \cdot 10^{-38}$). The second run did also bring up CELO-ORF-13 (E = 0.016). After the inclusion of FAdV-9-$ORF_{5058\text{-}4261}$ (the FAdV-9 homolog of ORF-13) in round 3, CELO-ORF-2 was found among the top hits (E = 0.55) after four iterations.

So, PSI-BLAST suggests distant links between ORF-12, ORF-13 and ORF-2 and, thus, to the NS-1 family. Those three ORFs are likely to form a paralogous group which originates from an acquired parvoviral NS-1 protein (see supplementary material for a more detailed phylogenetic analysis). Since (i) BLAST searches initiated with ORF-2 clearly hit AAV Rep proteins and (ii) interactions between adenoviruses and AAVs, which depend in their replication on a helper adeno- or herpesvirus [19], are naturally occuring, an AAV Rep protein is the most plausible candidate.

Rep proteins are multifunctional proteins and have a variety of enzymatic activities: DNA-binding activity, endonuclease activity, helicase activity and ATPase activity [20,21]. The regions of the Rep proteins responsible for the distinct activities have been functionally mapped in a variety of mutational studies [22-26] (Fig. 4).

Endonuclease activity is located in the 200 amino-terminal residues. This region is missing completely in the CELO/FAdV-9 sequences. ATPase/helicase activity was found to be located in the central region of the Rep proteins. This region is covered by the Pfam NS-1 domain which is conserved between other parvoviral non-structural proteins and the CELO/FAdV-9 ORFs. In other words, ORF-2, ORF-12, ORF-13 and their FAdV-9 homologs mainly consist of a domain derived from an ATPase/helicase domain.

The ATPase/helicase domain was previously classified as a superfamily III helicase [27]. This sort of helicase proteins can be found in small viruses. These proteins have three conserved sequence motifs tightly packed in an approximately 100-amino-acid domain. The first two of them

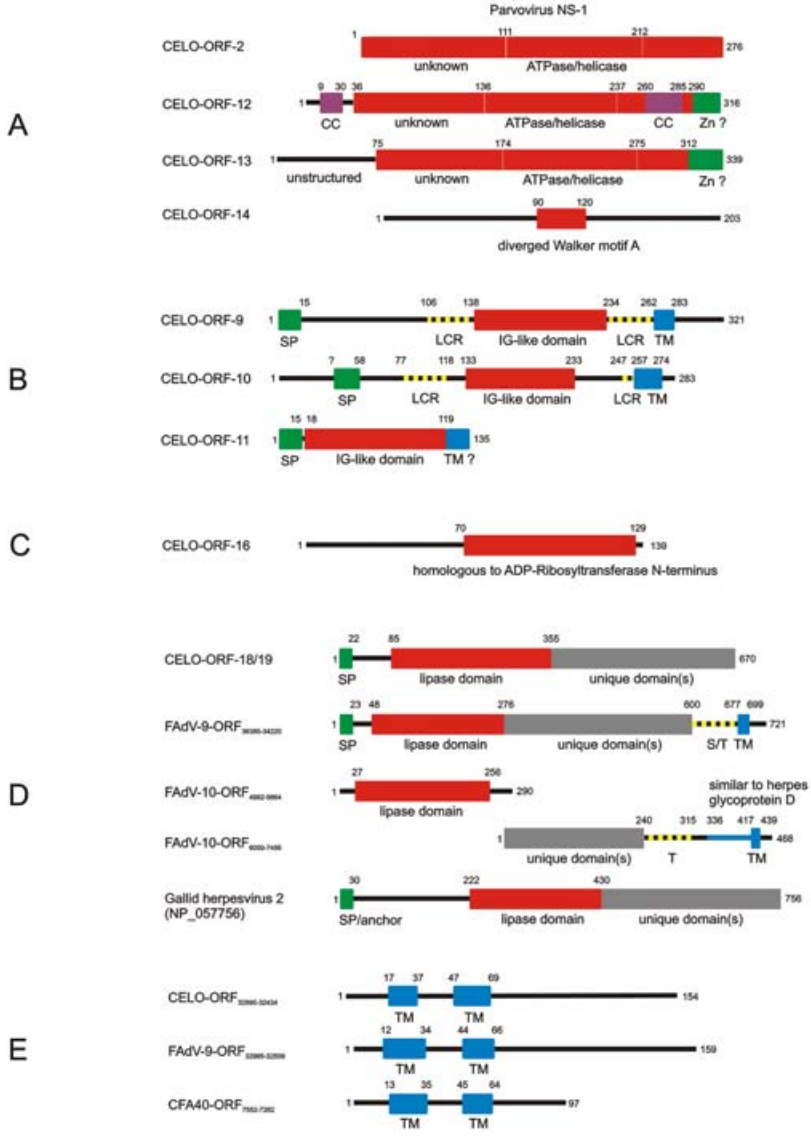

Figure 3

Molecular architecture of CELO ORFs and selected homologs. (A) The red bar in ORF-2, ORF-12 and ORF-13 denotes homologous regions to the parvovirus NS-1 domain (Pfam PF01057). These domains are subdivided in a inactive ATPase/helicase domain of the helicase superfamily III and a region with no clearly defined function. CC: potential coiled-coil segments as reported by "COILS". Zn: region with four conserved cysteines in ORF-12 and ORF-13. ORF-13 has an extended and presumably unstructured amino-terminus rich in prolines and charged residues. In ORF-14, a distant homology to the superfamily III helicases could be detected in the region of the "Motif A" (see text). (B) Putative type-1 transmembrane glycoproteins. SP: signal peptides predicted by SignalP. TM: transmembrane domains predicted by TMHMM. In ORF-11, the TM is not significantly predicted. LCR: low complexity regions reported by SEG with parameters 25, 3.0, 3.3. The red bar denotes homology to the immunoglobulin-like domain (SMART SM00409). Potential Asn-glycosylation sites (PROSITE PS00001) could be found in all three ORFs: ORF-9 (41, 89, 114, 135, 181), ORF-10 (75, 92, 121, 157, 179, 198, 223, 229), ORF-11 (74, 84, 89). (C) The red bar indicates homology in ORF-16 to a family of vertebrate mono-ADP-ribosyltransferases (Pfam PF01129) as reported by CD-Search. (D) CELO-ORF-18/19 and homologous sequences in FAdV-9, FAdV-10 and Gallid herpesvirus 2 (NP_057756, a representative of Marek-disease like viruses). All have a lipase domain (Pfam PF00151) and a region unique to this group of avian viruses. The domain boundaries were estimated according to the location of PSI-BLAST hits to known lipases. "S/T" and "T" denote regions rich in serine/threonine and threonine, respectively. These domains are predicted to be highly O-glycosylated. In FAdV-10-ORF$_{6050\text{-}7456}$ the blue colored region indicates a region of similarity to herpes glycoprotein D (Pfam PF01537) as reported by CD-Search. (E) Conserved transmembrane domains in CELO-ORF$_{32895\text{-}32434}$ and its homologs in FAdV-9 and CFA40 predicted by TMHMM.

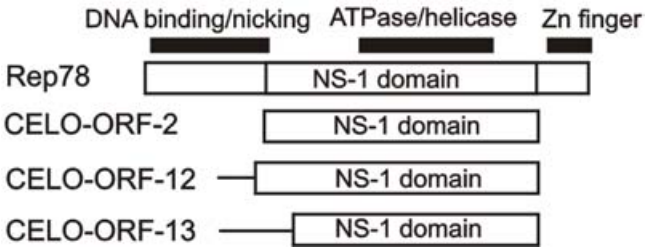

Figure 4
Functional regions mapped to Rep78 of adeno-associated virus in comparison to the location of the Pfam NS-1 domain present in CELO ORFs.

(motif A and B) form the NTP binding site and are specific versions of a NTP binding pattern common to many families of helicases. The third motif (C) is unique to superfamily III helicases [27]. In parvoviral sequences, an additional motif B' between B and C was identified [28].

Fig. 5 shows a multiple sequence alignment of the central region of Rep78 from AAV-3B to the NS-1 domains found in CELO and FAdV-9 sequences. The superfamily III helicase motifs are indicated. Motif A (also known as the Walker motif or P-loop, [29]) has the consensus [AG]-x(4)-G-K-[ST] (PROSITE PS00017) and forms a NTP interacting loop which connects a beta-sheet and an alpha-helix. In Rep78, this motif is perfectly represented, while in the CELO/FAdV-9 sequences critical residues are not conserved. The lysine and the serine/threonine are substituted in all cases. Only the glycines are partly conserved indicating the existence of a loop which is confirmed by the secondary structure prediction. Although some variations of the Motif A might be compatible with ATPase function if the typical sheet-loop-helix conformation is maintained [28], it is unlikely that this is the case here. The lysine and serine/threonine are strictly conserved throughout the superfamily III but also in related superfamilies [28] and, in the special case of AAV-Rep proteins, it was shown that mutation of either of these residues abolishes ATPase and helicase activity completely [24]. Also in the other three motifs, critical residues required for enzymatic activity are not or only partly conserved. This is most obvious for B' where a substantial part of the motif including three essential residues for helicase function [25] is deleted. To conclude, none of the sequences appear to be Rep-like enzymatically active, not even ORF-2 and FAdV-9-$ORF_{1950-2753}$, which are significantly similar to Rep proteins.

Interestingly, the ATPase/helicase motifs only cover 100 amino acids in the central part of the conserved NS-1 domain (Fig. 5). There are appr. 100 additional residues in the amino terminus. We could not find data that shows that this region is directly involved in ATPase/helicase activity and it is definitely not part of the amino-terminal endonuclease domain of the AAV Rep proteins [26]. Therefore, taking also into account the relatively high sequence conservation, we assume that the amino-terminal appr. 100 residues form another globular domain with additional yet unknown functions.

Also, the identity of the appr. 80 carboxy-terminal residues is unclear. Compared to the rest of the sequence, this region is not that well conserved and the CELO/FAdV-9 ORFs cannot be reliably aligned in this region. AAV Rep proteins have a carboxy-terminal domain which contains several zinc binding motifs (Fig. 4). This domain is known to bind zinc in vitro [30] but little is known about its function. In the CELO/FAdV-9 sequences, a distinct domain with pronounced zinc binding motifs is missing. However, for CELO-ORF-12, CELO-ORF-13 and their FAdV-9 homologs, some weak hits in the comparison with domain libraries (PFAM, SMART) point to various C4 zing finger domains. Those hits can be explained by the existence of four conserved cysteines in the very carboxy-terminus of the sequences (cysteine is a rare amino-acid type and, if cysteines match, they yield high scores). It can be speculated that these residues have zinc binding capability, although no further data can support this.

Furthermore, there is good evidence that AAV Rep proteins function as oligomers [31] and important interaction sites have been mapped to two putative coiled-coil regions [25,31]. All sequences were routinely scanned for regions with the potential to form coiled-coils. In the case of ORF-12 and its FAdV-9 homolog, two such regions are found (Fig. 3a). The signal in the carboxy-terminus lies exactly in the region corresponding to the experimentally determined interaction site. Closer inspection shows that this region is predicted with maximum confidence to form a helix which has amphipathic properties indicated by the typical distribution pattern of hydrophobic and hydrophilic residues. This result might suggest that also some of the adenovirus NS-1 proteins interact with each other.

ORF-14: an additional putative NS-1 domain protein

ORF-14 is located within the cluster of NS-1 proteins between ORF-2 and ORF-13 (Fig. 1). This genomic arrangement suggests a connection for ORF-14 to the NS-1 proteins. We have, indeed, evidence that ORF-14 is related to this protein family. In this case, however, the degree of divergence has almost reached the limit of detection and a homology could only be indirectly inferred in a short region of ORF-14.

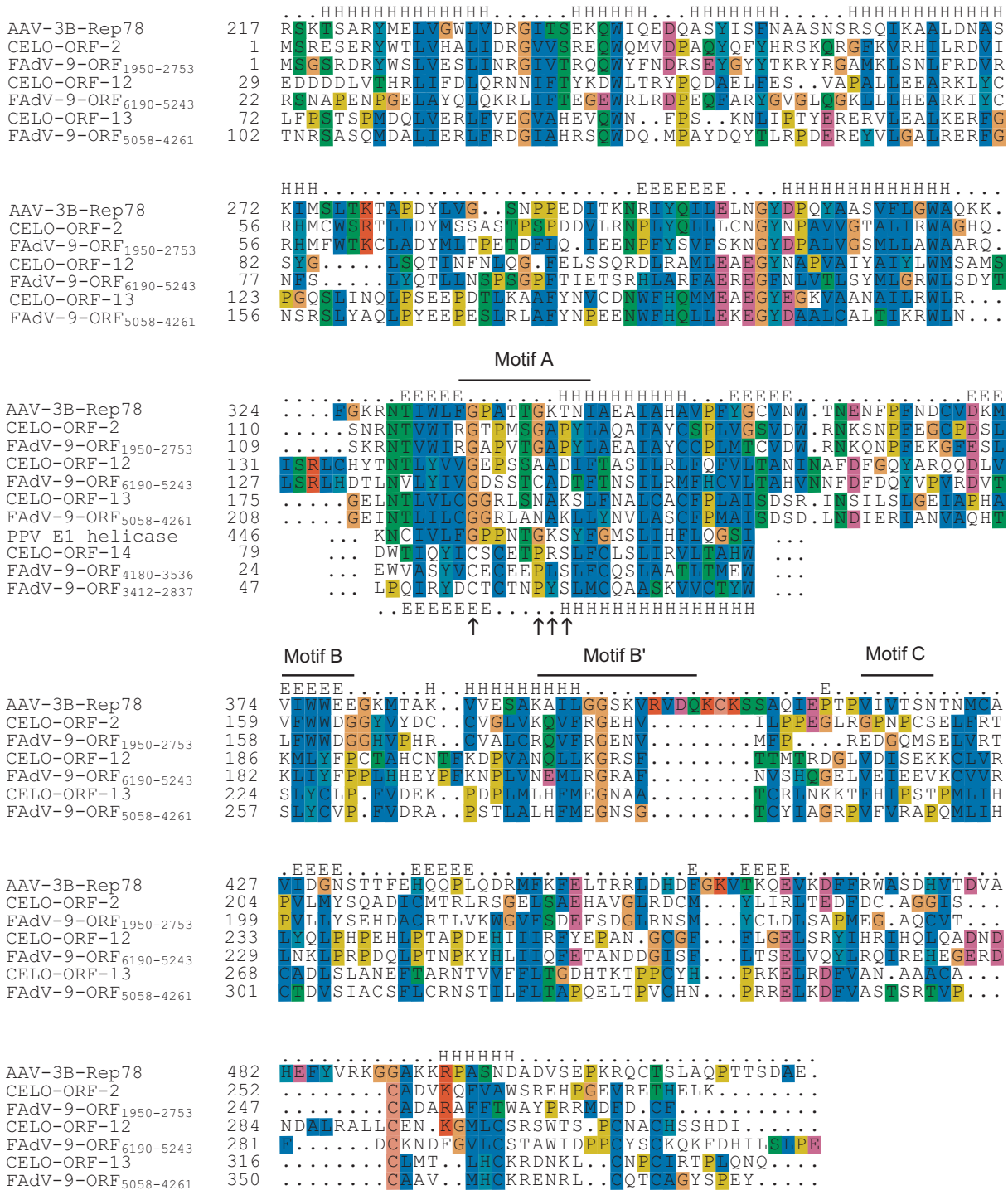

Figure 5
Multiple sequence alignment of parvovirus NS-1 domains found in CELO and FAdV-9. As a reference sequence, the Rep78 protein of adeno-associated virus 3B (acc. no. AAB95451) is included. JPred secondary structure prediction for CELO-ORF-2 is shown in the top line (H: alpha-helix, E: beta-sheet). Superfamily III ATPase/helicase motifs (see text) are indicated. Critical residues for NTP-binding in motif A are marked by arrows. In the region of motif A, CELO-ORF-14 and two homologous sequences from FAdV-9 were included in the alignment. In this region of CELO-ORF-14, homology to papillomavirus helicases is reported by CD-Search. As a reference sequence, papillomavirus E1 helicase (acc. no. P22154) is included. JPred secondary structure prediction for CELO-ORF-14 is shown in the bottom line.

In ORF-14, CD-search detected sequence similarity to E1 papillomavirus helicases (Pfam PF00519, pos. 90–120, E = 0.57). Although a borderline hit of limited statistical significance, it turned out to be of special interest. The E1 helicase (reviewed in [32]) is member of the same superfamily as the parvoviral NS-1 helicases [28]. Both have the Walker A-motif discussed above, and the short CD-search hit matches the region of this motif. Interestingly, there are two ORFs related to CELO-ORF-14 in FAdV-9. One full length homolog (ORF) can be easily found by BLASTP with $E = 6 \cdot 10^{-8}$. If this ORF is included in a PSI-BLAST query, another homolog (FAdV-9-$ORF_{3412-2837}$), which is encoded directly adjacent to FAdV-9-$ORF_{4180-3536}$, is detected (E = 1.8). The PSI-BLAST hit only matches a short region, which corresponds, again, to the Walker A motif. In the alignment in Fig. 5, the relevant stretches of CELO-ORF-14 and the two FAdV-9 sequences have been aligned to the A motif of the sequences with the parvoviral NS-1 domains. The motif itself is hardly recognizable but the hydrophobic pattern and also the typical sheet-loop-helix succession seems to be present.

To conclude, these remnants of the Walker A-motif indicate that there are additional ORFs in CELO and FAdV-9 which are likely to be derived from superfamily III helicases. Together with ORF-2, ORF-12 and ORF-13 they form a cluster which dominates the left terminal region in both genomes.

ORF-9, ORF-10, ORF-11: Putative type-1 transmembrane glycoproteins with an immunoglobulin-like domain

The analysis results for ORF-9, ORF-10 and ORF-11 show that the three ORFs, which are arranged directly adjacent to each other, are similarly organized and encode putative type-1 transmembrane glycoproteins (Fig. 3b). In all sequences, an amino terminal signal peptide is significantly predicted (probabilities of the SignalP hidden Markov model >0.9). In the case of ORF-10, a signal peptide is only predicted if the second methionine in the sequence is used as start (P = 0.996 in contrast to P = 0.027 if the complete sequence is used). This suggests that the start codon is at pos. 41113 rather than at pos. 41002. In ORF-9 and ORF-10, transmembrane regions (TM) are significantly predicted (classified as "certain" by Toppred with scores near 2 and TMHMM probabilities near 1). In ORF-11, no significant TM is reported. There is only a hydrophobic region in the carboxy-terminus labelled as a "putative" TM by Toppred.

In all three sequences, the Prosite Asn-glycosylation motif PS00001 was detected several times (see legend of Fig. 3b). This is a short and thus very common motif but the number of occurrences is unusual high for proteins of this length, and so some of them can be expected to be real glycosylation sites rather than mere statistical artifacts.

There is, apparently, one distinct globular domain common for all three ORFs. In ORF-11, this domain spans almost the complete sequence. In ORF-9 and ORF-10, this central domain is flanked by presumably unstructured low complexity regions. Detailed sequence analysis revealed that it is an immunoglobulin-like domain: In ORF-11, the SMART IG-domain (SMART SM00409) is predicted by CD-Search and HMMER (19–119, $E = 21 \cdot 10^{-7}$ and 18–119, $E = 3 \cdot 10^{-6}$, respectively). In the other two sequences, the prediction is not that clear but the domain can be plausibly assigned. In ORF-9, CD-Search predicts the SMART IG-domain in region 192–227 with E = 1.0. In ORF-10, it is detected by CD-Search (135–233, E = 0.71) and HMMER (166–233, E = 0.36). Furthermore, the 3D-PSSM fold recognition server proposes for all three sequences almost exclusively structures of the immunoglobulin superfamily. A multiple sequence alignment of the IG-like domains found in the CELO virus genome and in related viruses is available as part of the supplementary material on our website.

The IG-like fold is probably the most abundant protein fold that exists. As a consequence, public databases are full of proteins with IG-like domains and this makes homology searches with ORF-9, ORF-10 and ORF-11 difficult. In all cases, BLASTP detects a wide variety of different glycoproteins and surface receptors with borderline E-values. However, those hits most likely only reflect the fact that the proteins have the same fold and a closer evolutionary relationship could not be inferred for any of the three sequences to other known proteins. On the other hand, the results show that ORF-9, ORF-10 and ORF-11 are closer related to each other. A BLASTP search with ORF-9 against the NCBI non-redundant protein database finds ORF-10 with $E = 5 \cdot 10^{-4}$. A PSI-BLAST profile search initiated with ORF-11 (inclusion E-value 0.05) finds ORF-9 with E = 0.04 after the second iteration. These results suggest a common origin for these ORFs. Further database searches propose a candidate for a possible ancestor. We could find an expressed sequence tag from a chicken library which is highly similar to ORF-9 (acc.no. BM491231, TBLASTN against the NCBI EST database: $E = 6 \cdot 10^{-14}$). So, it is likely that this cluster of three similarly organized proteins form a paralogous group derived from a cellular gene that has been acquired from an avian host.

ORF-16: a putative ADP-ribosyltransferase

In ORF-16, an unexpected homology to ADP-ribosyltransferases (ARTs) could be detected. ARTs (reviewed in [33]) transfer the ADP-ribose moiety of NAD onto specific protein targets. ARTs have been long known in prokaryotes but an ART family could also be found in vertebrates [34-36]. In ORF-16, CD-search reported a hit from pos. 70 to 129 to this family of vertebrate ARTs (Pfam PF01129). The hit is statistically of borderline significance (E = 0.23)

but there are additional arguments which consistently support this finding.

(i) The hit matches the region of the ART NAD-binding pocket which constitutes the important region for enzymatic activity. This binding pocket is structurally conserved (see below) and characteristic for all ART enzymes of known structure [37-39].

(ii) Critical residues for enzymatic activity are conserved. Although the structural properties of the catalytic core are similar in distantly related ARTs, the conservation in primary sequence is remarkably low. Only typical fingerprint residues are conserved between the distantly related ARTs [37]. Vertebrate ARTs belong to a subgroup which is characterized by an Arg-Ser-Glu motif [37]. This motif can be found in ORF-16 (Fig. 6). The first arginine (Arg93) is well conserved together with other surrounding residues. The serine (Ser108) is also conserved and part of a short S/T rich stretch which is characteristic for the other ART sequences too. The relevant region of the glutamate in the Arg-Ser-Glu motif was not part of the CD-search hit. But there is a charged motif in the very carboxy-terminus of ORF-16 including a glutamate (Glu136) which can be plausibly aligned to the mainly acidic stretch found in the ART sequences which contains the critical glutamate.

(iii) Predicted secondary structural features of ORF-16 are compatible with the ART fold. The 3D-structure of a vertebrate ART of this family (ART2.2 from rat) has been determined recently [39]. Secondary structure predictions for ORF-16 are consistent with it (Fig. 6). The amino-terminal part is predicted to form mainly alpha-helices. Especially, α-4 and α-5 immediately upstream of the catalytic core are well predicted by different methods. In contrast, the catalytic core itself is, again in accordance with the ART2.2 structure, predicted to form mainly beta sheets. There is only one clear alpha-helix predicted in this region which matches exactly the α-6 of the ART2.2 structure. Furthermore, the gaps in ORF-16 match exactly the loop regions of the ART structure and no important secondary structures are broken or missing. Only β-9 and β-10 are missing due to the end of the sequence but both are not critical for the formation of the typical four stranded NAD-binding core which is made up by β-2, β-5, β-6 and β-8 [39].

(iv) For ART2.2 it was found that the fold of the catalytic core is stabilized by a disulfide bond tying together the two ends of the strands β-2 and β-6. The responsible cysteines are marked in the alignment. Both are conserved in ORF-16 (C88 and C128).

Taken together, there is sufficient evidence to suggest that ORF-16 is related to ADP-ribosyltransferases. To our surprise, ORF-16 has no homolog in FAdV-9. We could only detect a short homology in FAdV-10 ($ORF_{4550-4209}$). This ORF is similar to the amino-terminus of ORF-16, but it stops prematurely and the rest of the sequence including the relevant part showing ART homology in ORF-16 is unclear.

ORF-18/19: a putative triglyceride lipase with an additional domain unique to avian adeno- and herpesviruses

The sequence analysis of ORF-18 and ORF-19 suggested that both ORFs encode one single protein. A sequencing error was suspected and could be confirmed by comparison to an alternative nucleic acid sequence of CELO covering this region (acc.no. S33490). In the sequence of Chiocca et al., a single A is obviously missing at pos. 35749. Insertion of the missing nucleotide leads to a continuous open reading frame (ORF-18/19).

There are homologs of the merged ORF-18/19 in FAdV-9, CFA40 and FAdV-10 (Table 1) but also in Marek's disease-like viruses (MDV), a group of pathogenic avian herpesviruses [40]. Fig. 3d shows the architecture of the different proteins. In ORF-18/19, significant homology to triglyceride lipases (Pfam PF00151) could be detected by different methods (e.g. CD-Search reports a hit to this family in the region of 125–306 with $E = 3 \cdot 10^{-7}$). This homology to lipases has been noted previously in the CFA40 homolog [18] and also in the MDV sequences [41,42]. The active site serine and the surrounding residues (Prosite motif PS00120) are well conserved among all sequences, suggesting enzymatic activity (see supplementary material). However, only part of the Pfam lipase domain, which is widely distributed among animals, plants and prokaryotes, can be found in the viral proteins. Instead, there are about 300 residues unique to the avian and adenoviral proteins. PSI-BLAST and HMMER profile searches with this region did not find a connection to any other known proteins. Some of these residues may contribute to lipase function but additional functional domains can be expected. Interestingly, in FAdV-10 the lipase domain and the unique region is encoded by two distinct ORFs. It must be noted that this cannot be explained by a simple sequencing error as in the case of the CELO sequence.

Further results of the comparative analysis indicate that the proteins of this group are possibly membrane glycoproteins. Signal peptides and transmembrane regions could be identified (Fig. 3d). In the CELO sequence, no signal peptide could be found (SignalP: P = 0.005). However, Payet et al. report a short leader sequence which is spliced together with ORF-18/19 [13]. If this leader is included in the translation and an alternative ATG encoded by this leader is used as the start codon, the new amino terminus has significant signal peptide properties (P = 0.996). This suggests that the short 5'-leader

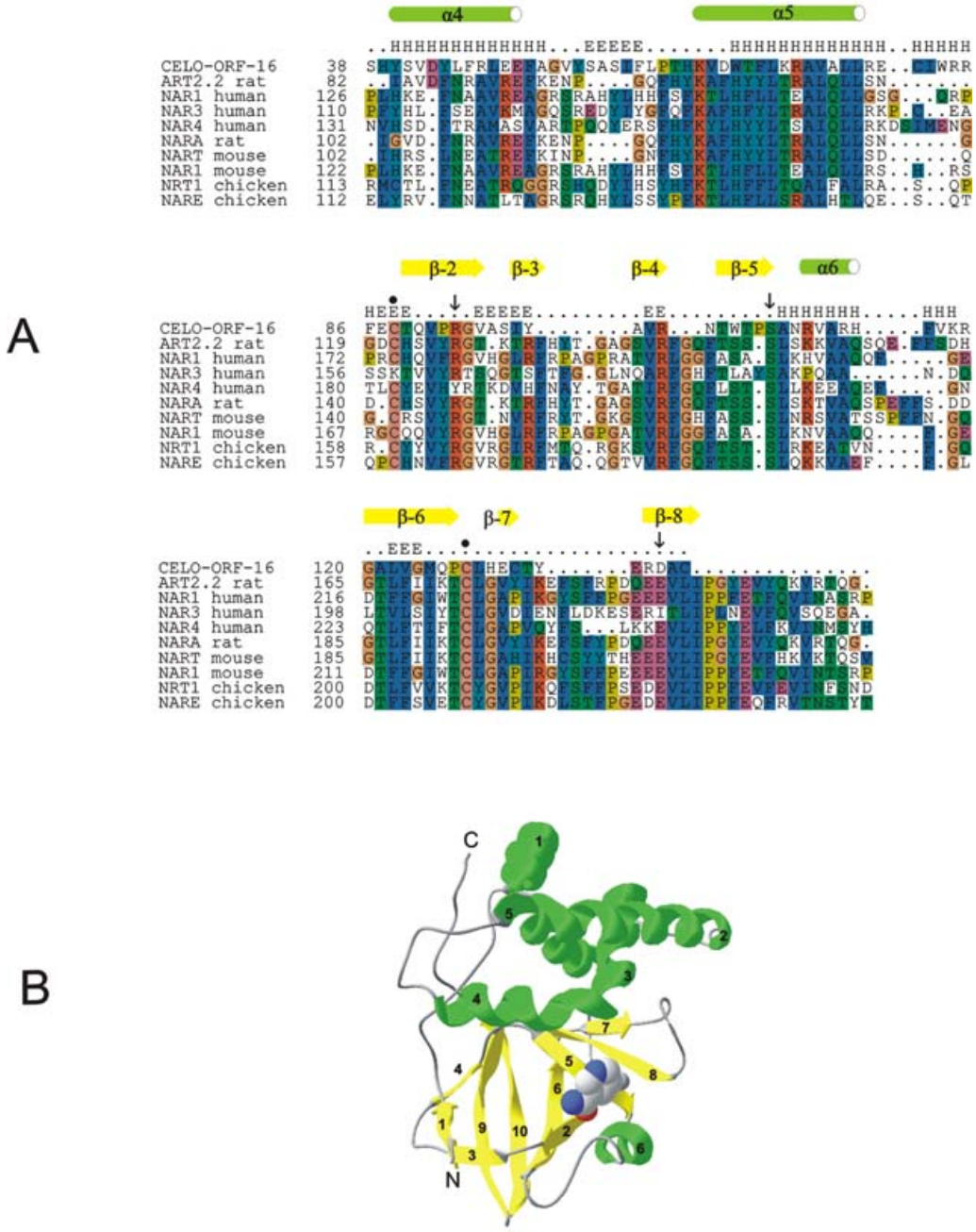

Figure 6
(A) Multiple sequence alignment of ORF-16 and diverse members of the vertebrate ADP-ribosyltransferase family (Pfam PF01129). Sequences are indicated by their SwissProt names. The established secondary structure for ART2.2 (PDB entry 1GXZ [39]) is indicated by colored bars. PHD secondary structure prediction for ORF-16 is indicated below. (The results of a consensus secondary structure prediction applying various other methods can be found as part of the supplementary material on our website.) The sequence alignment was basically adopted from the RPS-BLAST alignment of the CD-Search hit. The first 30 amino-terminal and the last 10 carboxy-terminal residues not covered by the CD-Search hit were manually aligned. It must be noted, that the alignment is, thus, based on sequence similarity alone and was not edited considering any structural information. Critical residues of the typical R-S-E motif (see text) are marked by an arrow. Disulfide forming cysteines are marked by a filled circle. (B) Structure of ART2.2 from rat. The structure consists of a mainly alpha helical amino-terminal part and a carboxy-terminal part dominated by beta-sheets. The NAD binding site is formed by β-2, β-5, β-6 and β-8. Numbering and colors of the secondary structural elements are the same as in (A).

sequences which are common during the transcription in CELO and FAdV-9 [13,17] are, at least in some cases, part of the coding sequence and must be regarded as short exons rather than untranslated leaders. Interestingly, also in the homologous sequence of Marek's disease virus 1 the signal peptide is encoded in a very short exon which is spliced together with a much longer second exon encoding the rest of the protein [41].

In FAdV-9, CFA40 and FAdV-10 an extended carboxy-terminus including S/T rich regions can be observed. In FAdV-10, there is a run of about 60 threonines interspersed only with some prolines. Such S/T rich domains are typical sites for O-glycosylation of the mucin type [43]. Moreover, the carboxy-terminus of FAdV-10-ORF was found by CD-Search to be similar to the carboxy-terminus of herpes glycoprotein D (Pfam PF01537, E = 0.007). In CELO this extended glycoprotein-like carboxy-terminus is missing. It might be encoded by another exon or might have been lost completely.

$ORF_{32895-32434}$: two conserved transmembrane domains

This ORF overlaps with the originally described ORF-21 and is read in a different frame on the same strand. It is conserved in CELO, FAdV-9 and CFA40 with respect to amino acid sequence and genomic location (in all three viruses it is located between ORF-20 and ORF-22). The analysis of $ORF_{32895-32434}$ found only one interesting feature in this sequence. There are two significantly predicted transmembrane segments (TMHMM probabilities > 0.9 and TopPred2 scores > 2). Also the homologous ORFs in FAdV-9 and CFA40 contain two transmembrane segments each (Fig. 3e). We do not have the impression that $ORF_{32895-32434}$ encodes a functional protein on its own but is conceivable that this conserved coding region is an exon which provides one or two transmembrane segments for some other ORFs. Candidate sequences are for example ORF-20 and ORF-18/19 which are located on the same strand directly upstream of $ORF_{32895-32434}$ and which are likely to be membrane located (indicated by signal peptides or transmembrane domains in close homologs).

Other ORFs

In the case of ORF-17 and $ORF_{28115-27765}$, the sequence analysis did not yield reasonable new results. For ORF-20, it can be noted that an amino-terminal signal peptide is significantly predicted in the FAdV-9 homolog. In ORF-20 and also in the CFA40 homolog, the amino terminus is unclear since the homology goes beyond the only methionine and another methionine cannot be observed. It can be speculated that ORF-20 is provided with a leader peptide by another exon, presumably the same as in the case of ORF-18/19. This assumption is supported by the genomic location and could account for the missing start codon.

Also in the case of the UTPase (ORF-1), GAM-1 (ORF-8) and ORF-22, which have been characterized experimentally [4-10], the sequence analysis could not add new aspects to what has already been known.

Discussion

We report the reannotation of the genome of the avian adenovirus CELO with emphasis on the unique terminal regions. In view of the unsatisfactory state of the previous annotation and the rapidly improving sequence analyzing techniques, this genome appeared worth to be revisited. So, we conducted a comprehensive sequence analysis on the protein level aimed towards a better understanding of the unique features of CELO biology.

In a first step, we had to refine the prediction of the coding regions and propose 15 ORFs which can be expected to be of functional importance. Interestingly, we found several ORFs without a start codon. This possibly indicates that some of these proteins are not encoded by one contiguous ORF and splicing is necessary to form the complete coding sequence. Also, simple errors in the genomic sequence can result in wrong or missing start codons which in turn can obscure the identity of ORFs remarkably. Both issues are difficult to deal with by theoretical methods. Therefore, protein sequences cannot be reliably determined in all cases. However, the relevant regions for this study have a manageable size of about 18 kb which could be examined manually. Thus, obvious pitfalls of an automatic ORF prediction could be avoided which resulted in a prediction which is in some cases quite different from what has been proposed before but which is likely to reflect the expression situation in vivo more precisely.

The subsequent in-depth sequence analysis of these new ORFs could shed new light on the identity of most of them. An unexpected result is that the majority of the ORFs are related to each other and cluster in paralogous groups.

The terminal region on the left side of the map (Fig. 1) is dominated by a group of ORFs with a conserved domain homologous to Rep proteins of adeno-associated viruses. This parvoviral domain is completely unusual in adenoviruses. Within this family, it can be exclusively found in CELO and its close relative FAdV-9. The very fact that the generally tightly packed and economically arranged CELO genome contains several copies of this domain suggests major functional importance for it.

The function of the adenoviral Rep proteins, however, must be different from the primary function of the Rep protein in AAVs. There, they are essential for a successful life cycle and are required for DNA nicking and subsequent priming of DNA replication, for site specific integra-

Table 1: Unique coding sequences in CELO and related avian adenoviruses

CELO-ORF[a]	Region[b]	Strand	Length (aa)	Homologous sequences[c]	Comment
ORF-1	794–1330	forward	178	FadV-9-ORF$_{847–1338}$	functional dUTP pyrophosphatase [4], homologous to ORF-1 proteins of mastadenovirus E4 region [5]
ORF-2	1999–2829	forward	276	FadV-9-ORF$_{1950–2753}$	
ORF-8	37391–38239	forward	282	FadV-9-ORF$_{37859–38668}$ FadV-10-ORF$_{2147–2911}$	Gam1, antiapoptotic, induction of heat shock response, inactivation of histone deacetylase 1, pRb/E2F pathway [6-10]
ORF-9	40037–41002	forward	321	FadV-9-ORF$_{43595–42660}$ CFA40-ORF$_{17739-16381}$ EST-ORF-9	amino terminus unclear, since transcript was shown to be spliced up to 40133 [13]
ORF-10	41002–41853	forward	283	no homologs	Translation start is likely to be at pos. 41113
ORF-11	41958–42365	forward	135	FadV-9-ORF$_{41461–41853}$	
ORF-12	5412-4462	reverse	315	FadV-9-ORF$_{6190-5243}$	amino terminus extended by 106 residues
ORF-13	4568-3549	reverse	339	FadV-9-ORF$_{5058-4261}$	
ORF-14	3503-2892	reverse	203	FadV-9-ORF$_{4180-3536}$ FadV-9-ORF$_{3412-2837}$	amino terminus extended by 43 residues
ORF-16	39705-39286	reverse	139	FadV-10-ORF$_{4550-4209}$	
ORF-17	39256-38717	reverse	179	FadV-9-ORF$_{41096-40596}$ CFA40-ORF$_{15112-14642}$ FadV-10-ORF$_{4023-3574}$	
ORF-18/19	(36144)-34238	reverse	635	FadV-9-ORF$_{36385-34220}$ CFA40-ORF$_{10653-8782}$ FadV-10-ORF$_{4992-5864}$ FadV-10-ORF$_{6050-7456}$	Correction of an obvious error in the genomic sequence combines ORF-18 and ORF-19 to one single ORF. The amino terminus was extended by the translation of a short leader sequence that was shown to be spliced ahead of the original ORF-18 [13].
ORF-20	33832-32892	reverse	313	FadV-9-ORF$_{33963-32986}$ CFA40-ORF$_{8466-7741}$	amino terminus extended by 42 residues
ORF-22	32429-31812	reverse	205	FadV-9-ORF$_{32502-31930}$ CFA40-ORF$_{3306-2729}$	involved in pRb/E2F pathway [10]
ORF$_{28115-27765}$	28115-27765	reverse	117	FadV-9-ORF$_{30192-29797}$ CFA40-ORF$_{4777-4478}$ FadV-10-ORF$_{1814-1637}$	New conserved ORF, located between fibre and pVIII gene
ORF$_{32895-32434}$	32895-32434	reverse	154	FadV-9-ORF$_{32985-32509}$ CFA40-ORF$_{7552-7262}$	new conserved ORF, is translation of the original ORF-21 in a different frame

[a] ORF-numbering in accordance to Chiocca et al. [4][b] coordinates in the genomic sequence of CELO (acc.no. NC_001720) [c] ORFs are named by their coordinates in the following nucleic acid sequences: NC_000899 (FAdV-9), AF155911 (CFA40) and AF160185 (FAdV-10). FAdV-10-ORF$_{1814-1637}$ is derived from entry AF006739, CFA40-ORF$_{3306-2729}$ from entry U40587. EST-ORF-9 is the translation of an EST sequence (BM491231) of a chicken EST library.

tion into the host genome and for packaging the single stranded DNA into the capsid [21,44,45]. These functions are useless for CELO simply because these processes do not occur or are solved in a different way during the life cycle of adenoviruses. This is consistent with the results of our sequence analysis which found that only the central region of the AAV-Rep proteins containing the ATPase/helicase function is present in CELO and FAdV-9 while the regions with DNA-binding and endonuclease activity are missing. Furthermore, the ATPase/helicase domain is most likely not functional indicated by the fact that critical residues which are conserved throughout the corresponding helicase-superfamily and which are known to be essential for enyzmatic activity in AAV Rep proteins are not conserved.

Therefore, other functions for this diverged non-functional domain must be envisaged. In AAVs, the *rep* gene is the only non-structural gene. This might be the reason why *rep* products have taken over a wide variety of other functions. Rep proteins are known, in different contexts, to act as transcriptional activators and repressors of homologous and heterologous promoters [46-49]. Several interaction partners have been identified including different transcription factors [50-54]. These results point to a general role in transcriptional regulation. Moreover, Rep proteins are also implicated in other cellular pathways as for example the p53 and pRB-E2F pathways where they exhibit onco-suppressive functions and hinder cell cycle progression [55,56]. Rep proteins are also known to induce apoptosis [57]. Interestingly, these functions are contrary to CELO physiology in which proliferation is enhanced and apoptosis is prevented with the help of Gam1 and ORF-22 [6,10].

In most of the cases, the exact molecular basis of all those Rep functions are not established yet. Diffuse mappings do not allow the identification of new functional domains or motifs. This situation is of course unfavourable for a detailed functional prediction for the CELO Rep proteins. However, CELO apparently makes use of the great functional plasticity of this protein family and we must expect that ORF-2, ORF-12, ORF-13 (and possibly also ORF-14) interact with a number of cellular targets resulting in implications for various pathways. They might be involved in transcriptional control as it can be seen in a rather general fashion for AAV Rep products. CELO possibly uses those early proteins to modulate the host's gene expression machinery in order to render cellular conditions more favourable.

In the right terminal region (Fig. 1), we could identify a cluster of three putative type-1 transmembrane glycoproteins with (partly diverged) immunoglobulin-like domains. IG-like domains are multi-purpose interaction domains and characteristic for proteins involved in recognition processes in the immune-system [58]. Also in the case of the CELO proteins, a connection to the immune system must be considered.

A virus is always threatened by the host's immune response and adenoviruses have evolved multiple strategies to escape the immune mechanisms (reviewed in [59]). In human adenoviruses, most of these functions are encoded by the E3 transcription unit which is not present in avian adenoviruses. Detailed E3 functions have primarily been described for human adenoviruses of the subgenus C. The E3 regions of different human subgenera differ remarkably and there are many E3 proteins of unknown function which are unique to distinct subgenera. It is noteworthy that several E3 products were shown to be type-1 transmembrane glycoproteins. Also a conserved domain which is thought to have an IG-like fold was found in some E3 proteins of subgenera B and D [60,61].

Although no closer evolutionary relationship between any of these known E3 proteins and the ORFs of the CELO IG-cluster could be detected, these ORFs are strong candidates to substitute for the missing immunomodulatory functions. The fact that not a single E3 protein is conserved in CELO, may be explained by the different immunological requirements that a virus faces in an avian host. This avian specificity is evident if we consider the origin of this gene cluster. We have found an expressed sequence tag from a chicken library which is a direct homolog to ORF-9. Although the corresponding gene/protein has not been characterized yet, this shows that an ORF-9 homolog must exist in the chicken genome. This chicken gene is likely to be present also in other avian species and is presumably the origin of the IG-like proteins in avian adenoviruses. It is an interesting scenario that a virus could have acquired an immune-receptor from the host and uses it, in course of its efforts to escape the immune mechanisms, to its own advantage.

Directly adjacent to the IG-cluster, ORF-16 can be found. We have well-founded evidence that ORF-16 is homologous to a family of vertebrate mono-ADP-ribosyltransferases. Although the overall sequence similarity is only within the twilight zone, the conservation of invariant fingerprint residues together with structural considerations including secondary structure prediction and conserved disulfide bond forming cysteines, strongly suggest that ORF-16 has a NAD-binding fold which is characteristic for all known ARTs. Interestingly, it has been speculated before that there might exist unrecognized ARTs in known genomes which could have evaded detection by standard methods due to the low conservation of primary sequence [35].

To our knowledge, this putative CELO ART would be the first occurrence of such an enzymatic activity in a vertebrate virus and this raises the question of its function in such a viral context.

ADP-ribosylation is well known as the pathogenic mechanism of some potent bacterial toxins such as pertussis, cholera and clostridial toxins [62]. On the other hand, the functions of vertebrate ARTs are still ill-defined. However, data is emerging that members of this family which can be found in mammalian and avian species play an important role in cell signaling and the modulation of inflammatory and immune response (reviewed in [63]). Different surface receptors (mostly expressed on cells of the immune system) have been identified as targets for ART mediated ADP-ribosylation. Such immuno-regulatory functions, based on the posttranslational modification of cell-surface receptors, would also make sense in the context of CELO infection. Considering the existence of three potential IG-like surface receptors in the CELO genome, it is of course tempting to speculate that CELO uses the ART activity to modify them. It must be noted, however, that the known members of the vertebrate ART family are localized in the extracellular space (secreted or glycosylphosphatidylinositol-anchored [34,35]). The sequence of ORF-16 has no features which indicate extracellular localization. It is possible that the amino terminus is not complete and a signal peptide is missing, as we can see it for other CELO ORFs. Alternatively, it is conceivable that the putative ART has changed target specifity and is located intracellularly. In any case, such an unusual enzymatic activity is of broader interest and appears worth to be pursued experimentally.

Finally, we have characterized the merged ORF-18/19 which is expected to encode a triglyceride lipase. Comparison to homologous sequences of other avian adeno- and herpesviruses show that these lipases are likely to be transmembrane glycoproteins and have an additional domain of unknown function unique to those viruses. It is difficult to speculate on a possible role of these lipases. Some ideas have been put forward previously [42].

Conclusions

Taken together, our results give a new picture of the unique terminal regions of the CELO genome. Even the use of different highly sensitive methods could not detect homologies to any known sequences of mastadenoviruses in these regions. In contrast, those methods could elucidate unexpected relationships to various other proteins. We found that CELO has acquired several genes from other viruses and also from its host. Apparently, these proteins form, partly after duplications and heavy diversification, a novel set of functions for host interaction in avian adenoviruses. This reannotation provides an important source of new information which can readily direct and assist experimental work. The detailed sequence analysis of the CELO gene products can help to devise new experiments and to interpret existing and forthcoming experimental results.

Materials and Methods

Sequences

The complete genomic sequences of CELO and FAdV-9 described by Chiocca et al. [4] and Ojkic et al. [15] were taken from the RefSeq [64] entries with GenBank accession numbers NC_001720 and NC_000899, respectively. Partial genomic sequences of the hypervirulent FAdV-9 strain CFA40 were taken from entry AF155911 [18], and in the case of FAdV-10 from entries AF160185 [65] and AF006739 [66].

Searching for homologous sequences

Public available sequence databases (National Center for Biotechnology Information, NIH, Bethesda) were scanned using the BLAST suite of programs, including BLASTP, TBLASTN and PSI-BLAST [67,68]. To enhance sensitivity during clustering and comparing of protein sequences among the avian adenoviruses, a custom library of all available sequence data for this group was created and searched as well.

Identifications of known domains and motifs

Sequences were compared to the NCBI conserved domain database [69] using the CD-search server http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi which uses the RPS-BLAST algorithm. The E-value cutoff was set to 100, forcing that all (also insignificant) hits were reported and could be critically inspected. Additionally, the Pfam [70] and SMART [71,72] collections of hidden Markov models of known protein domains and families were searched using the HMMER package (version 2.1.1, Sean Eddy, Dept. of Genetics, Washington university School of Medicine) in both global and fragmentary mode. All sequences were scanned for PROSITE [73] patterns and motifs using PPSEARCH (European Bioinformatics Institute).

Intrinsic protein features

Regions of biased amino acid content and regions of low complexity were detected with SAPS [74] and SEG [75]. Sequences were scanned for transmembrane regions using TopPred 2 [76] and TMHMM 2 [77]. Amino-terminal signal peptides were predicted with SignalP 2, applying both the neural network and the hidden Markov model [78].

Secondary and tertiary structure prediction

Secondary structure was predicted using PHD [79] and JPred [80]. The existence of coiled-coil structures was examined with COILS [81]. All sequences were submitted to the 3D-PSSM fold recognition server [82].

Sequence manipulation and multiple sequence alignments

All sequence manipulations, especially translation operations, were carried out with the appropriate programs of the EMBOSS package [83]. Multiple sequence alignments were created with the help of ClustalW [84] and T_coffee [85]. The alignments were automatically shaded according to the default settings of the ClustalX [86] interface.

In addition to the programs, servers and databases listed here, the sequences were also analyzed with a variety of other methods described previously [87,88]. However, they did not yield relevant results for this special study and, therefore, their description is omitted here.

List of abbreviations

CELO: Chicken embryo lethal orphan virus

ORF: Open reading frame

FAdV: Fowl adenovirus

AAV: Adeno-associated virus

TM: Transmembrane region

IG: Immunoglobulin

ART: ADP-ribosyltransferase

MDV: Marek's disease like virus

Authors' Contributions

The sequence analytic work was executed by SW. Both authors (SW, FE) contributed to evaluating the results and making the discoveries reported here. SW prepared all the figures and, together with FE, the manuscript text. Both authors read and approved the final manuscript.

Acknowledgments

The authors are grateful for generous support from Boehringer Ingelheim. Discussions with Sebastian Maurer-Stroh on methodical aspects and computational support from Werner Kubina are thankfully acknowledged. This project has been partly funded by the Austrian Gen-AU bioinformatics integration network sponsored by BM-BWK and BMWA.

References

1. Yates VJ, Fry DE: **Observations on a chicken embryo lethal orphan (CELO) virus.** *Am J Vet Res* 1957, **18:**657-660.
2. Laver WG, Younghusband HB, Wrigley NG: **Purification and properties of chick embryo lethal orphan virus (an avian adenovirus).** *Virology* 1971, **45:**598-614.
3. Russell WC: **Update on adenovirus and its vectors.** *J Gen Virol* 2000, **81:**2573-2604.
4. Chiocca S, Kurzbauer R, Schaffner G, Baker A, Mautner V, Cotten M: **The complete DNA sequence and genomic organization of the avian adenovirus CELO.** *J Virol* 1996, **70:**2939-2949.
5. Weiss RS, Lee SS, Prasad BV, Javier RT: **Human adenovirus early region 4 open reading frame 1 genes encode growth-transforming proteins that may be distantly related to dUTP pyrophosphatase enzymes.** *J Virol* 1997, **71:**1857-1870.
6. Chiocca S, Baker A, Cotten M: **Identification of a novel antiapoptotic protein, GAM-1, encoded by the CELO adenovirus.** *J Virol* 1997, **71:**3168-3177.
7. Glotzer JB, Saltik M, Chiocca S, Michou AI, Moseley P, Cotten M: **Activation of heat-shock response by an adenovirus is essential for virus replication.** *Nature* 2000, **407:**207-211.
8. Colombo R, Boggio R, Seiser C, Draetta GF, Chiocca S: **The adenovirus protein Gam1 interferes with sumoylation of histone deacetylase 1.** *EMBO Rep* 2002, **3:**1062-1068.
9. Chiocca S, Kurtev V, Colombo R, Boggio R, Sciurpi MT, Brosch G, Seiser C, Draetta GF, Cotten M: **Histone deacetylase 1 inactivation by an adenovirus early gene product.** *Curr Biol* 2002, **12:**594-598.
10. Lehrmann H, Cotten M: **Characterization of CELO virus proteins that modulate the pRb/E2F pathway.** *J Virol* 1999, **73:**6517-6525.
11. Francois A, Eterradossi N, Delmas B, Payet V, Langlois P: **Construction of avian adenovirus CELO recombinants in cosmids.** *J Virol* 2001, **75:**5288-5301.
12. Michou AI, Lehrmann H, Saltik M, Cotten M: **Mutational analysis of the avian adenovirus CELO, which provides a basis for gene delivery vectors.** *J Virol* 1999, **73:**1399-1410.
13. Payet V, Arnauld C, Picault JP, Jestin A, Langlois P: **Transcriptional organization of the avian adenovirus CELO.** *J Virol* 1998, **72:**9278-9285.
14. Wick N, Luedemann S, Vietor I, Cotten M, Wildpaner M, Schneider G, Eisenhaber F, Huber LA: **Induction of short interspersed nuclear repeat-containing transcripts in epithelial cells upon infection with a chicken adenovirus.** *J Mol Biol* 2003, **328:**779-790.
15. Ojkic D, Nagy E: **The complete nucleotide sequence of fowl adenovirus type 8.** *J Gen Virol* 2000, **81:**1833-1837.
16. Cao JX, Krell PJ, Nagy E: **Sequence and transcriptional analysis of terminal regions of the fowl adenovirus type 8 genome.** *J Gen Virol* 1998, **79 (Pt 10):**2507-2516.
17. Ojkic D, Krell PJ, Nagy E: **Unique features of fowl adenovirus 9 gene transcription.** *Virology* 2002, **302:**274-285.
18. Johnson MA, Pooley C, Lowenthal JW: **Delivery of avian cytokines by adenovirus vectors.** *Dev Comp Immunol* 2000, **24:**343-354.
19. Berns KI, Bohenzky RA: **Adeno-associated viruses: an update.** *Adv Virus Res* 1987, **32:**243-306.
20. Zhou X, Zolotukhin I, Im DS, Muzyczka N: **Biochemical characterization of adeno-associated virus rep68 DNA helicase and ATPase activities.** *J Virol* 1999, **73:**1580-1590.
21. Im DS, Muzyczka N: **The AAV origin binding protein Rep68 is an ATP-dependent site-specific endonuclease with DNA helicase activity.** *Cell* 1990, **61:**447-457.
22. Davis MD, Wonderling RS, Walker SL, Owens RA: **Analysis of the effects of charge cluster mutations in adeno-associated virus Rep68 protein in vitro.** *J Virol* 1999, **73:**2084-2093.
23. Walker SL, Wonderling RS, Owens RA: **Mutational analysis of the adeno-associated virus type 2 Rep68 protein helicase motifs.** *J Virol* 1997, **71:**6996-7004.
24. Walker SL, Wonderling RS, Owens RA: **Mutational analysis of the adeno-associated virus Rep68 protein: identification of critical residues necessary for site-specific endonuclease activity.** *J Virol* 1997, **71:**2722-2730.
25. Davis MD, Wu J, Owens RA: **Mutational analysis of adeno-associated virus type 2 Rep68 protein endonuclease activity on partially single-stranded substrates.** *J Virol* 2000, **74:**2936-2942.
26. Yoon M, Smith DH, Ward P, Medrano FJ, Aggarwal AK, Linden RM: **Amino-terminal domain exchange redirects origin-specific interactions of adeno-associated virus rep78 in vitro.** *J Virol* 2001, **75:**3230-3239.
27. Gorbalenya AE, Koonin EV, Wolf YI: **A new superfamily of putative NTP-binding domains encoded by genomes of small DNA and RNA viruses.** *FEBS Lett* 1990, **262:**145-148.
28. Koonin EV: **A common set of conserved motifs in a vast variety of putative nucleic acid-dependent ATPases including MCM proteins involved in the initiation of eukaryotic DNA replication.** *Nucleic Acids Res* 1993, **21:**2541-2547.
29. Walker JE: **Distantly related sequences in the alpha and beta subunits of ATP-synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold.** *EMBO J* 1982, **1:**945-951.
30. Horer M, Weger S, Butz K, Hoppe-Seyler F, Geisen C, Kleinschmidt JA: **Mutational analysis of adeno-associated virus Rep protein-mediated inhibition of heterologous and homologous promoters.** *J Virol* 1995, **69:**5485-5496.
31. Smith RH, Spano AJ, Kotin RM: **The Rep78 gene product of adeno-associated virus (AAV) self-associates to form a hexameric complex in the presence of AAV ori sequences.** *J Virol* 1997, **71:**4461-4471.
32. Wilson VG, West M, Woytek K, Rangasamy D: **Papillomavirus E1 proteins: form, function, and features.** *Virus Genes* 2002, **24:**275-290.
33. Koch-Nolte F, Reche P, Haag F, Bazan F: **ADP-ribosyltransferases: plastic tools for inactivating protein and small molecular weight targets.** *J Biotechnol* 2001, **92:**81-87.
34. Okazaki IJ, Zolkiewska A, Nightingale MS, Moss J: **Immunological and structural conservation of mammalian skeletal muscle glycosylphosphatidylinositol-linked ADP-ribosyltransferases.** *Biochemistry* 1994, **33:**12828-12836.
35. Glowacki G, Braren R, Firner K, Nissen M, Kuhl M, Reche P, Bazan F, Cetkovic-Cvrlje M, Leiter E, Haag F, Koch-Nolte F: **The family of toxin-related ecto-ADP-ribosyltransferases in humans and the mouse.** *Protein Sci* 2002, **11:**1657-1670.
36. Okazaki IJ, Kim HJ, Moss J: **Cloning and characterization of a novel membrane-associated lymphocyte NAD:arginine ADP-ribosyltransferase.** *J Biol Chem* 1996, **271:**22052-22057.
37. Domenighini M, Rappuoli R: **Three conserved consensus sequences identify the NAD-binding site of ADP-ribosylating enzymes, expressed by eukaryotes, bacteria and T-even bacteriophages.** *Mol Microbiol* 1996, **21:**667-674.
38. Han S, Craig JA, Putnam CD, Carozzi NB, Tainer JA: **Evolution and mechanism from structures of an ADP-ribosylating toxin and NAD complex.** *Nat Struct Biol* 1999, **6:**932-936.
39. Mueller-Dieckmann C, Ritter H, Haag F, Koch-Nolte F, Schulz GE: **Structure of the ecto-ADP-ribosyl transferase ART2.2 from rat.** *J Mol Biol* 2002, **322:**687-696.
40. Kato S, Hirai K: **Marek's disease virus.** *Adv Virus Res* 1985, **30:**225-277.
41. Lee LF, Wu P, Sui D, Ren D, Kamil J, Kung HJ, Witter RL: **The complete unique long sequence and the overall genomic organization of the GA strain of Marek's disease virus.** *Proc Natl Acad Sci U S A* 2000, **97:**6091-6096.

42. Tulman ER, Afonso CL, Lu Z, Zsak L, Rock DL, Kutish GF: **The genome of a very virulent Marek's disease virus.** *J Virol* 2000, **74:**7980-7988.
43. Hanisch FG: **O-glycosylation of the mucin type.** *Biol Chem* 2001, **382:**143-149.
44. Linden RM, Ward P, Giraud C, Winocour E, Berns KI: **Site-specific integration by adeno-associated virus.** *Proc Natl Acad Sci U S A* 1996, **93:**11288-11294.
45. King JA, Dubielzig R, Grimm D, Kleinschmidt JA: **DNA helicase-mediated packaging of adeno-associated virus type 2 genomes into preformed capsids.** *EMBO J* 2001, **20:**3282-3291.
46. Lackner DF, Muzyczka N: **Studies of the mechanism of transactivation of the adeno-associated virus p19 promoter by Rep protein.** *J Virol* 2002, **76:**8225-8235.
47. Kyostio SR, Wonderling RS, Owens RA: **Negative regulation of the adeno-associated virus (AAV) P5 promoter involves both the P5 rep binding site and the consensus ATP-binding motif of the AAV Rep68 protein.** *J Virol* 1995, **69:**6787-6796.
48. Zhan D, Santin AD, Liu Y, Parham GP, Li C, Meyers C, Hermonat PL: **Binding of the human papillomavirus type 16 p97 promoter by the adeno-associated virus Rep78 major regulatory protein correlates with inhibition.** *J Biol Chem* 1999, **274:**31619-31624.
49. Batchu RB, Shammas MA, Wang JY, Munshi NC: **Interaction of adeno-associated virus Rep78 with p53: implications in growth inhibition.** *Cancer Res* 1999, **59:**3592-3595.
50. Hermonat PL, Santin AD, Zhan D: **Binding of the human papillomavirus type 16 E7 oncoprotein and the adeno-associated virus Rep78 major regulatory protein in vitro and in yeast and the potential for downstream effects.** *J Hum Virol* 2000, **3:**113-124.
51. Hermonat PL, Santin AD, Batchu RB, Zhan D: **The adeno-associated virus Rep78 major regulatory protein binds the cellular TATA-binding protein in vitro and in vivo.** *Virology* 1998, **245:**120-127.
52. Hermonat PL, Santin AD, Batchu RB: **The adeno-associated virus Rep78 major regulatory/transformation suppressor protein binds cellular Sp1 in vitro and evidence of a biological effect.** *Cancer Res* 1996, **56:**5299-5304.
53. Weger S, Wendland M, Kleinschmidt JA, Heilbronn R: **The adeno-associated virus type 2 regulatory proteins rep78 and rep68 interact with the transcriptional coactivator PC4.** *J Virol* 1999, **73:**260-269.
54. Weger S, Hammer E, Heilbronn R: **Topors, a p53 and topoisomerase I binding protein, interacts with the adeno-associated virus (AAV-2) Rep78/68 proteins and enhances AAV-2 gene expression.** *J Gen Virol* 2002, **83:**511-516.
55. Batchu RB, Shammas MA, Wang JY, Munshi NC: **Dual level inhibition of E2F-1 activity by adeno-associated virus Rep78.** *J Biol Chem* 2001, **276:**24315-24322.
56. Batchu RB, Kotin RM, Hermonat PL: **The regulatory rep protein of adeno-associated virus binds to sequences within the c-H-ras promoter.** *Cancer Lett* 1994, **86:**23-31.
57. Schmidt M, Afione S, Kotin RM: **Adeno-associated virus type 2 Rep78 induces apoptosis through caspase activation independently of p53.** *J Virol* 2000, **74:**9441-9450.
58. Williams AF, Barclay AN: **The immunoglobulin superfamily--domains for cell surface recognition.** *Annu Rev Immunol* 1988, **6:**381-405.
59. Burgert HG, Blusch JH: **Immunomodulatory functions encoded by the E3 transcription unit of adenoviruses.** *Virus Genes* 2000, **21:**13-25.
60. Deryckere F, Burgert HG: **Early region 3 of adenovirus type 19 (subgroup D) encodes an HLA-binding protein distinct from that of subgroups B and C.** *J Virol* 1996, **70:**2832-2841.
61. Windheim M, Burgert HG: **Characterization of E3/49K, a novel, highly glycosylated E3 protein of the epidemic keratoconjunctivitis-causing adenovirus type 19a.** *J Virol* 2002, **76:**755-766.
62. Ueda K, Hayaishi O: **ADP-ribosylation.** *Annu Rev Biochem* 1985, **54:**73-100.
63. Corda D, Di Girolamo M: **Mono-ADP-ribosylation: a tool for modulating immune response and cell signaling.** *Sci STKE* 2002, **2002:**PE53.
64. Pruitt KD, Maglott DR: **RefSeq and LocusLink: NCBI gene-centered resources.** *Nucleic Acids Res* 2001, **29:**137-140.
65. Sheppard M, Werner W, Tsatas E, McCoy R, Prowse S, Johnson M: **Fowl adenovirus recombinant expressing VP2 of infectious bursal disease virus induces protective immunity against bursal disease.** *Arch Virol* 1998, **143:**915-930.
66. Sheppard M, Tsatas E, Johnson M: **DNA sequence analysis of the genes for the fowl adenovirus serotype 10 putative 33K and pVIII.** *DNA Seq* 1998, **9:**37-43.
67. Altschul SF, Koonin EV: **Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases.** *Trends Biochem Sci* 1998, **23:**444-447.
68. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
69. Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH: **CDD: a database of conserved domain alignments with links to domain three-dimensional structure.** *Nucleic Acids Res* 2002, **30:**281-283.
70. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL: **The Pfam protein families database.** *Nucleic Acids Res* 2002, **30:**276-280.
71. Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP, Bork P: **Recent improvements to the SMART domain-based sequence annotation resource.** *Nucleic Acids Res* 2002, **30:**242-244.
72. Schultz J, Milpetz F, Bork P, Ponting CP: **SMART, a simple modular architecture research tool: identification of signaling domains.** *Proc Natl Acad Sci U S A* 1998, **95:**5857-5864.
73. Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K, Bairoch A: **The PROSITE database, its status in 2002.** *Nucleic Acids Res* 2002, **30:**235-238.
74. Brendel V, Bucher P, Nourbakhsh IR, Blaisdell BE, Karlin S: **Methods and algorithms for statistical analysis of protein sequences.** *Proc Natl Acad Sci U S A* 1992, **89:**2002-2006.
75. Wootton JC: **Non-globular domains in protein sequences: automated segmentation using complexity measures.** *Comput Chem* 1994, **18:**269-285.
76. Claros MG, von Heijne G: **TopPred II: an improved software for membrane protein structure predictions.** *Comput Appl Biosci* 1994, **10:**685-686.
77. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.** *J Mol Biol* 2001, **305:**567-580.
78. Nielsen H, Brunak S, von Heijne G: **Machine learning approaches for the prediction of signal peptides and other protein sorting signals.** *Protein Eng* 1999, **12:**3-9.
79. Rost B: **PHD: predicting one-dimensional protein structure by profile-based neural networks.** *Methods Enzymol* 1996, **266:**525-539.
80. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ: **JPred: a consensus secondary structure prediction server.** *Bioinformatics* 1998, **14:**892-893.
81. Lupas A, Van Dyke M, Stock J: **Predicting coiled coils from protein sequences.** *Science* 1991, **252:**1162-1164.
82. Kelley LA, MacCallum RM, Sternberg MJ: **Enhanced genome annotation using structural profiles in the program 3D-PSSM.** *J Mol Biol* 2000, **299:**499-520.
83. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends Genet* 2000, **16:**276-277.
84. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22:**4673-4680.
85. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment.** *J Mol Biol* 2000, **302:**205-217.
86. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.** *Nucleic Acids Res* 1997, **25:**4876-4882.
87. Novatchkova M, Eisenhaber F: **Can molecular mechanisms of biological processes be extracted from expression profiles? Case study: endothelial contribution to tumor-induced angiogenesis.** *Bioessays* 2001, **23:**1159-1175.

88. Eisenhaber B, Maurer-Stroh S, Novatchkova M, Schneider G, Eisenhaber F: **Enzymes and auxiliary factors for GPI lipid anchor biosynthesis and post-translational transfer to proteins.** *Bioessays* 2003, **25:**367-385.