Virology 456-457 (2014) 329–341

Contents lists available at ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/yviro

# A new look at adenovirus splicing

Hongxing Zhao *[,1], Maoshan Chen [1,2], Ulf Pettersson

*Department of Immunology, Genetics and Immunology, Uppsala University, S-751 85 Uppsala, Sweden*

ARTICLE INFO

*Article history:*
Received 20 December 2013
Returned to author for revisions
23 January 2014
Accepted 3 April 2014
Available online 25 April 2014

*Keywords:*
Adenovirus type 2 infection
RNA splicing
Polyadenylation
cDNA sequencing

ABSTRACT

Adenovirus type 2 RNA splicing events were quantitatively mapped by using deep cDNA sequencing. The majority of the previously identified splice sites were detected. The lack of complete consistency between the present and previous results is because of some sites which were incorrectly mapped in previous studies, such as the splice sites for pVII, pVIII and E3-11.6K. Several previously predicted splice sites such as that for E3-14.5K and E4ORF3/4 were not detected. In addition, several new splice sites were identified. The novel RNAs may code for hitherto undetected proteins or alternatively spliced mRNAs for known proteins. The open reading frames downstream of two novel splice sites, located in the major late transcription unit region, were shown to be highly conserved. Another interesting possibility is that some of them are non-coding RNAs. Finally, the adenovirus mRNA polyadenylation sites were accurately mapped and in some cases shown to be heterogeneous.



## Introduction

Adenoviruses are non-enveloped, icosahedral viruses containing a linear, double-stranded DNA molecule (Green et al., 1967). Both strands are transcribed with genes located on both the so-called rightward reading strand (r-strand) and leftward reading strand (l-strand). The genome consists of five early transcription units (E1A, E1B, E2, E3 and E4), two intermediate units (IX and IVa2) and one major late unit that generates five families of late mRNAs (for review (Shenk, 1996)). An additional late l-strand transcription unit encoding the U exon protein (UXP) has been identified recently (Tollefson et al., 2007; Ying et al., 2010). Each transcription unit contains its own promoter, and most of them encode more than two mRNAs by differential splicing of a single linear transcript. Expression of adenovirus genes is regulated during the productive infection in a step-wise manner. The immediate early gene E1a is expressed first followed by the expression of the delayed early genes, E1B, E2, E3 and E4. Then the intermediate early genes, IVa2 and IX are expressed. L1, a late transcript from the major late promoter, is also made during this phase. Soon after the onset of viral DNA synthesis, the transcription is switched from an early to a late mode for the production of viral structural proteins. Adenovirus makes an extensive use of alternative RNA splicing to produce a very complex set of mRNAs. Except for the polypeptide IX mRNA, all adenovirus primary transcripts undergo one or more splicing events which give rise to about fifty distinct mRNAs during a lytic infection. From early region E1A, three major mRNAs, the 13S, 12S and 9S mRNAs, and two minor mRNAs, the 10S and 11S mRNAs, are produced by alternative splicing (Berk and Sharp, 1978; Chow et al., 1979; Perricaudet et al., 1979; Schmitt et al., 1987; Spector et al., 1978; Virtanen and Pettersson, 1983). These mRNAs have common 5′ and 3′ ends, but differ from each other by the size of the intron. The E1A 13S and 12S mRNAs are the most abundant RNA species early after infection, while the 9S mRNA represents less than 5% of the total E1A mRNAs. At late times, a shift in the steady-state levels of the mRNAs occurs and the 9S mRNA becomes the most abundant species. Transcription of region E1B generates two major mRNAs (22S and 13S) and two minor mRNAs (14.5S and 14S) by splicing of one or two introns from a common precursor RNA (Berk and Sharp, 1978; Spector et al., 1978; Virtanen and Pettersson, 1983). The 22S and 13S mRNAs are the predominant species whereas the 14.5S and 14S mRNA represent less than 5% of the total steady-state level of E1B mRNAs (Babich and Nevins, 1981). Following the progression of the infection, the abundances of E1B 22S and 13S mRNAs are changed from equal amounts during the early stage to approximately 20-fold excess of 13S mRNA at late times. Region E2 is transcribed from the l-strand by using alternative promoter sites for the initiation of transcription. Two major classes of transcripts, E2A and E2B, have been described (Chow et al., 1979; Stillman et al., 1981). Although multiple mRNAs are produced from E2A, one mRNA encoding a single-stranded DNA-binding protein (DBP) is dominant. It is produced by the removal of two introns from the pre-mRNA. E2B precursor RNAs bypass the polyadenylation signal used for E2A mRNAs and extend to a downstream polyadenylation site. Two major proteins, the 87K

* Corresponding author. Fax: +46 18 471 4808.
*E-mail address:* Hongxing.Zhao@genpat.uu.se (H. Zhao).
[1] These authors contributed equally to this work.
[2] Present address: Department of Biochemistry, La Trobe Institute for Molecular Science, La Trobe University, Melbourne, Victoria 3083, Australia.

http://dx.doi.org/10.1016/j.virol.2014.04.006
0042-6822/

terminal protein (pTP) and the 140K DNA polymerase (Adv-Pol), are translated from the E2B mRNAs. They are both involved in adenovirus DNA replication (Smart and Stillman, 1982; Stillman et al., 1981, 1982). During the late stages of infection UXP is expressed abundantly from the l-strand from its own promoter, but its function remains unknown (Tollefson et al., 2007; Ying et al., 2010). The E3 transcription unit is embedded within the major late transcription unit and several of the splice sites are utilized in the maturation of the fiber mRNA. Previous studies have identified at least nine E3 mRNAs which are generated by differential processing of two major RNA precursors, E3A and E3B. The precursors share a common cap site, but differ from each other by having different 3′-ends. The E3B has one poly(A) addition site at 30846, whereas E3A has four poly(A) addition sites at 29792, 29799, 29801 and 29804. The E3 mRNAs are translated into seven proteins, namely 12.5K, 6.7K, gp19K, 11.6K, 10.4K, 14.5K and 14.7K (Wold et al., 1995). These proteins play a very important role in counteracting the antiviral defenses of the host (Gooding, 1992; Gooding and Wold, 1990). The E4 transcription unit is located at the right end of the l-strand (Pettersson et al., 1976; Sharp et al., 1974). A single primary transcript is spliced and generates a very complicated set of mRNAs and as many as 24 mRNAs with identical 5′ and 3′ ends have been reported (Freyer et al., 1984; Herisse et al., 1981; Rigolet and Galibert, 1984; Tigges and Raskas, 1984; Virtanen et al., 1984). E4 contains seven open reading frames (ORFs), and all except ORF3/4 polypeptides have been detected in infected cells.

Nearly all of transcription from the major late transcription unit (MLTU) starts at the onset of adenovirus DNA replication. Primary transcripts are generated from a single promoter and they are polyadenylated at one of five different sites, generating five families of RNAs, L1–L5. A minimum of 20 mRNAs are generated from the MLTU, and nearly all of them have a common 201-nucleotide tripartite leader sequence at their 5′ ends. Some of them contain an additional 440-nucleotide long i-leader exon. The L1 pre-mRNA is transcribed both at intermediate and late times, but its splicing pattern changes with time. The mRNA which codes for the two structurally related polypeptides 52K and 55K is produced both at intermediate and late times after infection, whereas the IIIa mRNA is produced exclusively in the late phase. Two poly(A) addition sites at 14113 (the most abundant) and 14119 were identified (Hales et al., 1988; Prescott and Falck-Pedersen, 1994).The L4-100K and L4-22K mRNAs are transcribed at a relatively early time of the late phase (Larsson et al., 1992). The L4-22K mRNA is transcribed from an internal L4 promoter embedded in the ORF of L4-100K (Morris et al., 2010). The L4-22K protein suppresses adenovirus early gene expression, but activates the full panel of L1-L5 transcription. The L4-33K protein is virus-encoded alternative RNA splicing factor and has been shown to activate splicing of Ad late gene transcripts with weak 3′ splice sites (Tormanen et al., 2006). Furthermore, L4-33K protein regulates selective accumulation of Ad late gene transcripts (Wu et al., 2013). All remaining late mRNAs are produced in the late phase. The L2 pre-mRNA is spliced into four major mRNAs with a common poly(A) addition site at nt 17969 and encoding polypeptides pIII (penton base), pV (major core protein), pVII (core protein) and pX (pμ) (Akusjarvi and Persson, 1981; Le Moullec et al., 1983); the L3 pre-mRNA is spliced into three major mRNAs with a common poly(A) site at nt 22443 which encode polypeptides pVI (hexon associated protein), pII (hexon) and the 23K viral protease (Prescott and Falck-Pedersen, 1994); the L4 pre-mRNA is spliced into four polypeptides, 100K, 22K, 33K and pVIII (hexon associated protein) with a common poly(A) site at nts 28223 and 28228 (Sittler et al., 1994); and finally, L5 includes transcripts encoding only pIV (fiber) with a poly(A) addition site at nt 32798. L5 mRNAs comprise a family of transcripts with a number of different 5′-leader sequences (x, y, z leaders) in various combinations in addition to leaders 1, 2, 3 and i (Le Moullec et al., 1983; Uhlen et al., 1982).

For decades, adenoviruses have served as an outstanding model system to study the molecular mechanisms of splicing due to the simplicity of their genomes and their efficient mode of replication. Most adenovirus mRNAs are generated by the removal of one or more introns and most of these introns are located in the 5′ or 3′ noncoding portion of pre-mRNA. Thus the viral introns do only in a few cases interrupt the ORFs. The development of high throughput sequencing methods has facilitated the discovery of many novel transcribed regions and splicing isoforms (Djebali et al., 2012). It is also a very powerful tool to study alternative splicing under different conditions at an unprecedented depth. Here we present a comprehensive analysis of adenovirus RNA splicing during different phases of the infection and a complete adenovirus splicing map. Two deep sequencing experiments, single-end and paired-end sequencing, were performed. Single-end sequencing was done by using the standard single-read DNA library preparation. The major shortcoming of this procedure is the short sequence reads and an exponential increase in error rates along the reads (Cox et al., 2010). The more recently developed paired-end sequencing allows for reading 255 nt long sequences from both ends of cDNA fragments. The data generated from paired-end sequencing, utilized in our second experiment, should thus be more reliable.

## Results and discussion

### *Summary of sequencing results*

Using mRNA single-end and total RNA paired-end sequencing technologies, the Ad2 RNA splicing profile during different phases was studied. Infection and RNA isolation were performed as in our previous study (Zhao et al., 2007). Briefly, synchronized human primary lung fibroblasts (IMR-90) were infected at a multiplicity of 100 FFU/cell. Infected cells were collected at 6, 12, 24, and 36 hpi. As shown in our previous study, these time points represent different stages of the infectious cycle, i.e. before any adenoviral gene expression, after immediate early gene (E1a) expression, after the onset of adenoviral DNA replication, and after late gene expression, respectively (Zhao et al., 2007). Thus, we could correlate the expression of RNAs with the progression of the infection. Single- and paired-end sequencing was performed. Only three RNA samples from 12 hpi, 24 hpi and mock were subjected to single-end sequencing in a pilot experiment whereas all RNA samples including 6 hpi, 12 hpi, 24 hpi, 36 hpi and mock were sequenced by paired-end sequencing. Single-end sequencing yielded 50–54 million 76 bp long sequence reads per sample, as shown in our previous publication (Zhao et al., 2012). The fraction of reads that aligned to the adenovirus genome increased dramatically, from 1.3 million reads at 12 hpi to 15.9 million reads at 24 hpi, indicating a very efficient infection (Table 1). The 401 reads that were aligned to the adenovirus genome in the mock sample represented the background noise. In the case of paired-end sequencing 30 million of 255 bp long sequence reads per sample were generated. The sequence reads that mapped to the adenovirus genome increased dramatically after 12 hpi. By using TopHat (Trapnell et al., 2009), an efficient read-mapping algorithm designed to align the sequence reads to reference genome without relying on known splice sites, 2228 and 1460 adenovirus splice junctions (with more than 1 sequence reads) were identified by single-end and paired-end sequencing, respectively. Although more splice junctions were identified by single-end sequencing, their sequence coverage was much lower than seen with paired-end sequencing. In general, the number of sequence reads

**Table 1**
Summary of sequencing data.

| Technique | Detection | Total | Mock | Ad-6 hpi | Ad2-12 hpi | Ad2-24 hpi | Ad2-36 hpi |
|---|---|---|---|---|---|---|---|
| **Single-end seq** | **Total reads** | | 53,146,621 | | 50,045,456 | 54,355,211 | |
| | **Aligned to human** | | 31,696,007 (59.6%)[a] | | 29,549,346 (59%) | 35,346,824 (33.9%) | |
| | **Aligned to Ad2 genome** | | 401 (0.0%)[b] | | 1,358,215 (2.7%) | 15,922,575 (29.3.0%) | |
| | **Splice junction** ( ≥ 1 reads) | 2228 | 2 | | 102 | 2,191 | |
| | **Splice junction** ( ≥ 100 reads) | 251 | 0 | | 23 | 238 | |
| | **Splice junction** ( ≥ 1000 reads) | 45 | 0 | | 9 | 44 | |
| **Paired-end seq** | **Total reads** | | 32,017,993 | 34,571,161 | 35,084,650 | 29,991,204 | 34,995,648 |
| | **Aligned to human** | | 22,906,437 | 23,590,900 | 17,354,900 | 15,276,823 | 6,750,352 |
| | | | 71.5% | 68.2% | 49.5% | 50.9% | 19.3% |
| | **Aligned to Ad2 genome** | | 7428[c] | 71,269 | 354,140 | 6,339,287 | 20,352,847 |
| | | | (0.03%) | (0.2%) | (1.0%) | (21.1%) | (58.2%) |
| | **Splice junction** ( ≥ 1 reads) | 1460 | 43 | 55 | 99 | 551 | 1064 |
| | **Splice junction** ( ≥ 100 reads) | 545 | 2 | 3 | 26 | 223 | 467 |
| | **Splice junction** ( ≥ 1000 reads) | 178 | 0 | 0 | 7 | 78 | 145 |

[a] Percentage of sequence reads that aligned to the human genome.
[b] Percentage of sequence reads that aligned to the adenovirus genome.
[c] The sequence reads in the sample from mock infected cells that aligned to the adenovirus genome are largely due to a minimal accidental contamination of RNA with the 36 hpi sample.

covering splice junction generated by paired-end sequencing was 10 times higher than generated by single-end sequencing. Thus suggesting that the specificity of the paired-end sequencing data was higher than obtained with single-end sequencing. Considering the difference in accuracy and sensitivity between the two data sets, different criteria were used for scoring new splice junctions. The splice junctions, covered by more than 100 sequence reads from single-end sequencing or more than 1000 reads from paired-end sequencing, were considered as significant. However, most of the result presented here are based on pair-end sequencing data. At 6 hpi, no or very low adenovirus gene expression was detected in IMR-90 cell as shown in our previous study (Zhao et al., 2007). Correspondingly, very few viral splice junctions were detected. Most splice junctions with significant sequence recovery were detected after 12 hpi and increased rapidly from 12 to 24 hpi, which correlated with the progression of the infection. The complete list of splice junctions aligned to adenovirus genome has been submitted to the National Center for Biotechnology Information Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) under accession no. GSE54583.

*Changes of adenovirus RNA splicing profile during the progression of the infection*

To reveal the changes in the adenovirus splicing profile the relative abundance of each splice junction, measured by paired-end sequencing, was calculated at each time point. Splice junctions representing more than 1% of all viral splice junctions at any given time point were considered as the most significant and included in Tables 2A–2C. The sequence reads that aligned to the adenovirus genome at 6 hpi were very low (see in Table 1), and were not included in this analysis. At 12 hpi, the most significant splice junctions were located on the l-strand (highlighted in bold text in Table 2A) whereas only two splice junctions were on the r-strand, in region E3. The splice junctions located within the E4 region were the most significant and the splice junction for E4ORF3 accounted for 53.8% of all adenovirus splice junctions at this time. The splice junctions situated in the E2 region were also prominent. Their relative abundance decreased rapidly with time and they accounted for 9.7% and 2.3% of all adenovirus splice junctions at 24 and 36 hpi, respectively. The splice junction for E4ORF3 fell to 0.66% at 24 hpi. The splice junctions within regions E1A and E1B were also detected, but they contributed to less than 1% of all viral splice junctions (see next section).

Following the progression of the infectious cycle into the S phase at 24 hpi, adenovirus late gene expression started. Consequently, a large number of splice junctions located within the major late transcription unit (MLTU) were detected. As shown in Table 2B (bold text), the most abundant splice junctions were situated within the leader region. The splice junctions for adenovirus late proteins, such as 55/52K, 13.6K, 33K and 100K, became prominent. It is noteworthy that L4 splices appeared earlier than splices in regions L2, L3, and L5. The explanation is presumably that L4 contains a promoter that is embedded in the coding sequence for the 100K protein. Although the percentage of splice junctions within the E2 and E3 regions was decreasing at 24 hpi and there was a noteworthy increase in their abundance. The splice junction for the recently discovered U gene on the l-strand (UXP) reached its highest expression at this time point.

In the late phase, the rate of transcription from the MLP increased to a level about 20 times higher than that from the early transcription units (Shaw and Ziff, 1980). All splice junctions which

**Table 2A**
The most significant splice junctions detected at 12 hpi (representing > 1% of all viral splice junctions at this time point) and their abundances at late times.

| Splice junction | | | Sequence reads by paired-end seq | | | | | Intron length | mRNA |
|---|---|---|---|---|---|---|---|---|---|
| Start | End | Strand | Mock | Ad-6 hpi | **Ad-12 hpi** | Ad-24 hpi | Ad-36 hpi | | |
| 35547 | 34736 | – | 18 | 177 | **19628 (53.8%)**[a] | 6043 (0.66)% | 681 (0.02%) | 812 | E4orf3 |
| 27981 | 28375 | + | 34 | 22 | **4605 (12.6%)** | 28,374 (3%) | 43,507 (1.4%) | 395 | E3 |
| 27024 | 24792 | – | 3 | 16 | **2135 (5.9%)** | 17,052 (1.8%) | 4201 (0.1%) | 2233 | E2A/E2B |
| 24714 | 24089 | – | 0 | 11 | **1652 (4.5%)** | 24,348 (2.6%) | 10,003 (0.3%) | 626 | E2A |
| 35547 | 34436 | – | 0 | 2 | **1124 (3.1%)** | 346 (0.04%) | 37 | 1112 | E4 region |
| 35547 | 34330 | – | 2 | 1 | **1059 (2.9%)** | 421 (0.05%) | 68 | 1218 | E4 region |
| 24714 | 24378 | – | 1 | 1 | **746 (2%)** | 3888 (0,4%) | 1450 (0.05%) | 337 | E2 region |
| 24196 | 24089 | – | 0 | 0 | **575 (2%)** | 1896 (0.2%) | 387 (0.01%) | 108 | E2 region |
| 24193 | 24089 | – | 0 | 0 | **498 (1.4%)** | 1981 (0.2%) | 372 (0.01%) | 105 | E2 region |
| 28560 | 30437 | + | 4 | 21 | **429 (1.2%)** | 6375 (0.7%) | 12392 (0.4%) | 1878 | E3 |

[a] Percentage of sequence reads covering a given splice junction relative to the total sequence reads covering all adenovirus splice junctions. For clarity the data at 12 hpi is highlighted in red.

**Table 2B**
The most significant splice junctions detected at 24 hpi (representing > 1% of all viral splice junctions at this time point) and their abundances at early and late times.

| Splice junction | | | Sequence reads by paired-end seq | | | | | Intron length | mRNA |
|---|---|---|---|---|---|---|---|---|---|
| Start | End | Strand | Mock | Ad-6 hpi | Ad-12 hpi | **Ad-24 hpi** | Ad-36 hpi | | |
| 7173 | 9633 | + | 486 | 425 | 343 (0.9%) | **281,568 (30.5%)**[a] | 1,426,350 (44.8%) | 2461 | Leaders 2–3 |
| 6080 | 7100 | + | 352 | 223 | 187 (0.5%) | **166,717 (18.1%)** | 573,713 (18%) | 1021 | Leaders 1–2 |
| 9724 | 11039 | + | 41 | 33 | 96 (0.26%) | **48,483 (5.3%)** | 75,495 (2.4%) | 1316 | Protein 52, 55K |
| 7173 | 7941 | + | 64 | 29 | 40 (0.1%) | **32,416 (3.5%)** | 45,039 (1.4%) | 769 | Leaders 2–i |
| 8382 | 9633 | + | 24 | 15 | 28 | **32,309 (3.5%)** | 54,467 (1.7%) | 1252 | Leaders i–3/13.6K protein |
| 26552 | 26753 | + | 9 | 0 | 15 | **31,039 (3.4%)** | 88,362 (2.8%) | 202 | 33k Protein |
| 27981 | 28375 | + | 34 | 22 | 4605 (12.6%) | **28,374 (3.1%)** | 43,507 (1.4%) | 395 | Leaders X–Y/E3 |
| 9724 | 24094 | + | 31 | 24 | 12 | **27,968 (3%)** | 61,766 (1.9%) | 14371 | 100K |
| 24714 | 24089 | - | 0 | 11 | 1652 (4.5%) | **24,348 (2.6%)** | 10,003 (0.3%) | 626 | E2A |
| 24792 | 27024 | - | 3 | 16 | 2135 (5.9%) | **17,052 (1.9%)** | 4201 (0.1%) | 2233 | E2A/E2B |
| 9724 | 16515 | + | 32 | 32 | 18 | **16,289 (1.8%)** | 89,029 (2.8%) | 6792 | Core protein V |
| 30855 | 24792 | - | 0 | 0 | 24 | **9348 (1%)** | 5599 (0.2%) | 6064 | UXP |

[a] As described in Table 2A.

**Table 2C**
The most significant splice junctions detected at 36 hpi (representing > 1% of all viral splice junctions) and their abundances at early times.

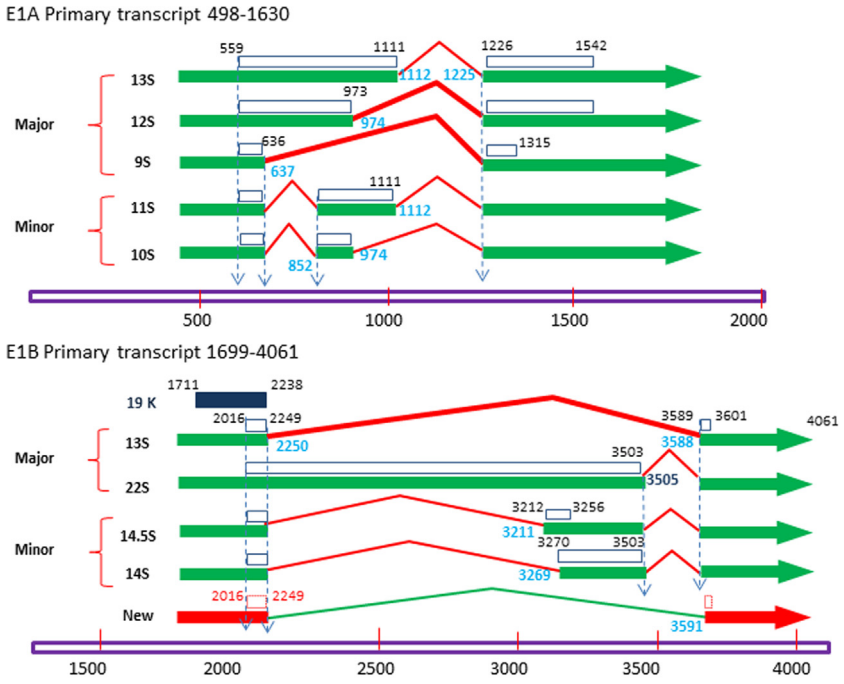
| Splice junction | | | Paired-end seq | | | | | Intron length | mRNA |
|---|---|---|---|---|---|---|---|---|---|
| Start | End | Strand | Mock | Ad-6 hpi | Ad-12 hpi | Ad-24 hpi | **Ad-36 hpi** | | |
| 7173 | 9633 | + | 486 | 425 | 343 (0.9%) | 281,568 (30.1%) | **1,426,350 (44.8%)**[a] | 2461 | Leaders 2–3 |
| 6080 | 7100 | + | 352 | 223 | 187 (0.5%) | 166,717 (18.1%) | **573,713 (18%)** | 1021 | Leaders 1–2 |
| 9724 | 16515 | + | 32 | 32 | 18 | 16,289 (1.8%) | **89,029 (2.8%)** | 6792 | Core protein V |
| 26552 | 26753 | + | 9 | 0 | 15 | 31,039 (3.4%) | **88,362 (2.8%)** | 202 | 33k Protein |
| 9724 | 11039 | + | 41 | 33 | 96 (0.3%) | 48,483 (5.3%) | **75,495 (2.4%)** | 1316 | 52, 55K |
| 9724 | 28375 | + | 35 | 17 | 7 | 6978 (0.8%) | **74,604 (2.3%)** | 18652 | Leaders 3–Y |
| 9724 | 24094 | + | 31 | 24 | 12 | 27,968 (3%) | **61,766 (1.9%)** | 14371 | 100K |
| 8382 | 9633 | + | 24 | 15 | 28 | 32,309 (3.5%) | **54,467 (1.7%)** | 1252 | 13.6K Protein |
| 7173 | 7941 | + | 64 | 29 | 40 (0.1%) | 32,416 (3.5%) | **45,039 (1.4%)** | 769 | Leaders 2–i |
| 27981 | 28375 | + | 34 | 22 | 4605 (12.6%) | 28,374 (3.1%) | **43,507 (1.4%)** | 395 | Leaders X–Y |
| 9724 | 18801 | + | 13 | 14 | 11 | 5418 (0.6%) | **39,521 (1.2%)** | 9078 | Hexon |
| 28560 | 31029 | + | 16 | 15 | 11 | 5447 (0.6%) | **36,841 (1.2%)** | 2470 | Leader Y–fiber |

[a] As described in Table 2A.

constituted > 1% of all viral splice sites at 36 hpi were located within the MLTU as shown in Table 2C. The splice junctions between leaders 1, 2 and 3 were still highest and accounted for more than 64% of all sequence reads covering adenovirus splice junctions. All remaining splice junctions which accounted for more than 1% of all viral splice junctions were located in L1–L5 regions.

*Splice junctions within the E1A and E1B transcription units*

E1A is an immediate early gene which is expressed first but at a very low level. Previous studies have shown that five mRNAs are produced from E1A region, three major products, 13S, 12S and 9S mRNAs, and two minor products, the 10S and 11S mRNAs (Schmitt et al., 1987). Here, we detected 4 splice junctions which matched exactly those identified in the earlier studies (Fig. 1). The sequence reads covering splice junctions for the 10S and 11S mRNAs were very low. Another consistency with earlier studies was the kinetics of E1A splicing. The 12S and 13S mRNAs were most abundant at 12 and 24 hpi, whereas 9S mRNA became the most abundant later (Schmitt et al., 1987). The relative abundance of 12S and 13S was different at different stages of infection. The 13S was more abundant at 12 hpi unlike at 24 hpi. These results consistent with




| Region | Splice site | | Sequence reads from paired-end seq | | | | | mRNA |
|---|---|---|---|---|---|---|---|---|
| | Start | End | Mock | Ad-6 hpi | Ad-12 hpi | Ad-24 hpi | Ad-36 hpi | |
| E1A | 637 | 852 | 0 | 0 | 43 | 165 | 182 | E1A-10S/11S |
| | 637 | 1225 | 3 | 0 | 112 | 2037 | 2249 | E1A-9S |
| | 974 | 1225 | 1 | 0 | 184 | 1531 | 1013 | E1A-12S |
| | 1112 | 1225 | 1 | 10 | 257 | 741 | 241 | E1A-13S |
| E1B | 2250 | 3211 | 0 | 0 | 0 | 28 | 268 | E1B-14.5S |
| | 2250 | 3269 | 0 | 0 | 0 | 139 | 135 | E1B-14S |
| | 2250 | 3588 | 5 | 1 | 28 | 4082 | 2646 | E1B-13S |
| | 2250 | 3591 | 1 | 0 | 1 | 589 | 312 | New |
| | 3505 | 3588 | 0 | 0 | 0 | 266 | 408 | E1B-22S/14.5/14S |

**Fig. 1.** A splicing map of the E1A and E1B transcription units. The green arrows represent the known E1A and E1B transcripts. The red arrow is a predicted transcript. The red lines indicate the known splice sites. The green line is a new splice site. The blue open boxes are ORFs. The numbers indicate the positions of ORFs and the splice junctions on the Ad2 genome. The red open boxes and numbers are predicted ORFs and their positions. For reference, the E1 region is shown at the bottom of figure with marked nucleotide positions. The sequence reads covering the splice sites are listed in the table at the bottom.

the kinetics of 12S and 13S mRNA transcription in a previous study (Svensson et al., 1983). The inefficient splicing of 10S and 11S mRNA is most likely due to the unusually long distance between the branch point sequence and the 3′ splice site in the first intron (Chebli et al., 1989; Gattoni et al., 1988).

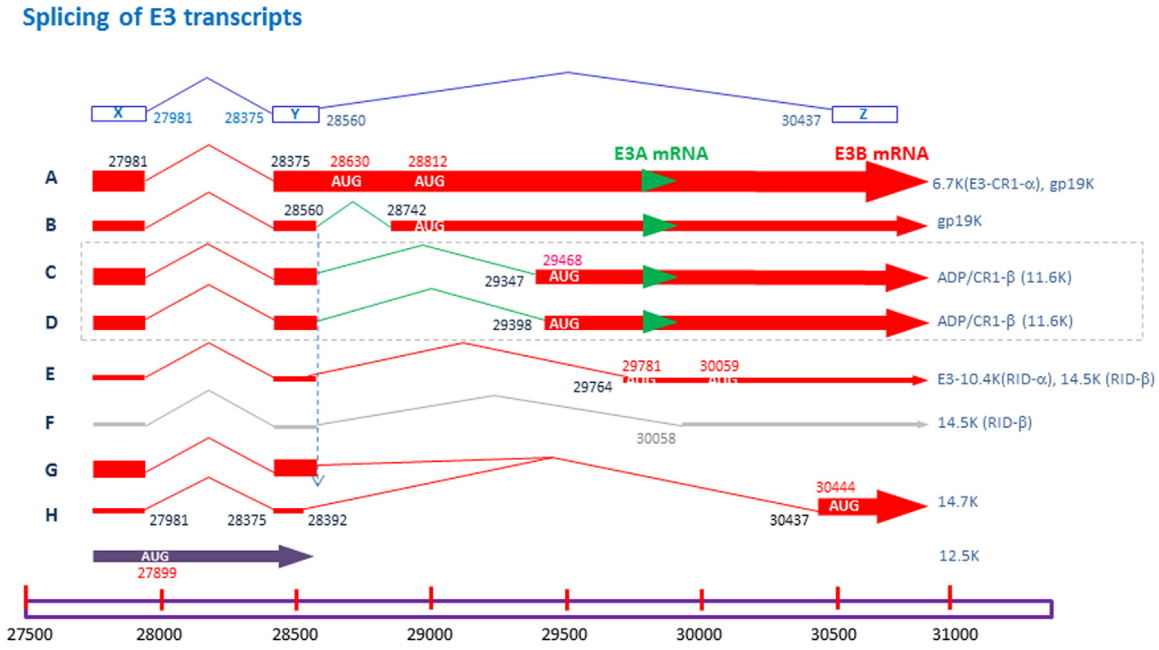
From E1B, a single primary transcript is produced which gives rise to two major mRNAs, 13S and 22S, and two minor mRNAs, the 14.5S and 14S mRNAs, by 4 alternative splicing events. Splicing in E1B peaked later than splicing in E1A starting at 24 hpi. Splicing of E1B RNAs was regulated over time. At 24 hpi, splicing occurred predominantly at 2250–3588 and accounted for 80% of all splicing within E1B region. It then decreased to 70% at 36 hpi, when splicing between 3505–3588 and 2250–3211 was increased. The splicing between 2250 and 3269 remained at similar level. In addition to the four previously described splice sites, a new splice site from 2250 to 3591 was detected. Its intron is three nucleotides longer at the 3′-end than that of the 13S mRNA. The resulting mRNA would encode a protein which is one amino acid shorter than that encoded by 13S mRNA. However, this splice junction was 7–8 fold less abundant than that of the authentic 13S mRNA, suggesting that it played a less significant functional role.

*Splicing in the E3 transcription unit*

The E3 region is located entirely within the MLTU (Wold et al., 1995). At least nine mRNAs have been reported from this region by alternative splicing of two major RNA precursors, E3A and E3B. Two splice sites coincide with the splice sites for the leaders X, Y and Z which are present on fiber mRNAs. Here, seven splice junctions were identified, four of which (A, E, G and H as shown in Fig. 2) were identical with previously described sites. Three 3′ splice sites were new (the second intron of B–D). The 3′splice site of the 2nd intron for gp19K was 28742 instead of 28772 (B). Two alternative splice sites, 29347 (C) and 29398 (D) seem to be used for the E3 11.6K (CR1-β) mRNA instead of 29465 which was undetectable. One previously reported splice site (F) for the 14.5K (RID-β) mRNA, 28560–30058, was also undetectable. The coverage of the two splice sites C and D for 11.6K (CR1-β) was similar at 24 hpi, although the site between 28560 and 29347 (C) was preferentially used at 36 hpi. A rare 5′ splice site (H) for 14.7K was identified, but the sequence reads were 9-fold lower than from the major site (G) and detectable only at 36 hpi. No specific mRNA for the RID-β protein was detected in the present study. It seems likely that the mRNA with a splice between 28560 and 29764 is polycistronic and encodes both the RID-alpha and -beta proteins (Tollefson et al., 1990a, 1990b). Most splicing in the E3 region occurred during the early phase of the infection (12 hpi) and continued until late (36 hpi).

*Splicing in the E4 transcription unit*

The E4 transcription unit is located at the extreme right end of the l-strand of the genome (Berk and Sharp, 1978; Chow et al., 1979; Kitchingman and Westphal, 1980; Pettersson et al., 1976).



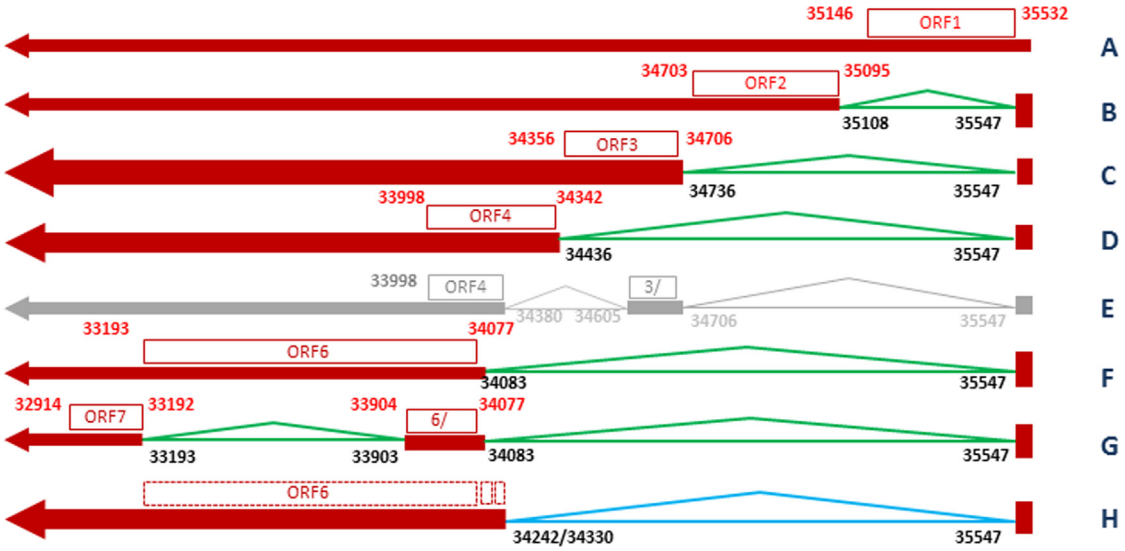

| | Splice junction | | | Sequence read from paired-end seq | | | | | Intron length | mRNA |
|---|---|---|---|---|---|---|---|---|---|---|
| | SS-Start | SS-End | Old site | Mock | Ad-6 hpi | Ad-12 hpi | Ad-24 hpi | Ad-36 hpi | | |
| A | 27981 | 28375 | | 34 | 22 | 4605 | 28374 | 43507 | 395 | X-Y/all E3 (except 12.5K) |
| B | 28560 | 28742 | 28772 | 0 | 0 | 38 | 280 | 2328 | 183 | gp19K |
| C | 28560 | 29347 | 29465 | 4 | 6 | 160 | 574 | 7211 | 788 | ADP (11.6K) |
| D | 28560 | 29398 | 29465 | 0 | 13 | 104 | 579 | 2309 | 839 | ADP (11.6K) |
| E | 28560 | 29764 | | 0 | 0 | 116 | 393 | 1225 | 1205 | E3 RID-α /β (10.4K/14.5K) |
| G | 28560 | 30437 | | 4 | 21 | 429 | 6375 | 12392 | 1878 | Y-Z/E3-14.7K |
| H | 28392 | 30437 | | 0 | 0 | 0 | 0 | 1374 | 2046 | New E3-14.7K |

**Fig. 2.** A splicing map of the E3 transcription unit. The green and red arrows represent the E3A and E3B mRNAs, respectively. The thickness of the arrow indicates the relative abundance of the sequence reads covering the splice junction. The thin lines indicate the splice sites. The thin green lines are a new splice sites identified here. The gray lines are the undetectable splice sites. The black and red numbers indicate the positions of the splice sites and ORFs on the Ad2 genome, respectively. For reference, a part of the Ad2 DNA is shown at the bottom of figure with marked nucleotide positions. The sequence reads covering the splice sites are listed in the table at the bottom.

Previous studies have shown that the primary transcript from the E4 region is spliced into a complicated set of mRNAs. At least 24 mRNAs have been detected by sequence analysis of cDNA clones and S1 nuclease analysis (Freyer et al., 1984; Herisse et al., 1981; Rigolet and Galibert, 1984; Tigges and Raskas, 1984; Virtanen et al., 1984). These mRNAs have common 5′ and 3′ ends but differ in their internal splicing patterns. Our sequencing results identified only ten splice junctions. The sequence reads for most of them were very low. The splice junctions for ORF2, ORF3, ORF4, ORF6 and ORF6/7 (B–D, F and G as shown in Fig. 3), as well as the splice junctions of 33903–33284, 34288–34083 and 34605–34436 were consistent with those previously reported. No sequence reads covering the splice junction (E) in the mRNA for the ORF3/4 protein was detected. Two new splice sites from 35547 to 34330 or 343242 (H) were identified. The sequence reads covering the splice site 35547–34330 was the third most abundant in the E4 region at 12 hpi. Although the sequence reads which covered the junction 35547–343242 were very low; they were comparable with those covering other well defined splice sites, such as that in the mRNA for the E4ORF2 protein. In addition, this splice junction was identified by both single- and paired-end sequencing. However, the predicted ORF downstream of the splice junction, 35547–34330 which starts at 34291 and ends at 34229 is only 63 nt long. Its biological function is thus uncertain. The predicted ORF which followed the splice junction 35547–34242 was from 34229 to 34211 and is thus unlikely to encode a protein. However, there follows an additional ORF located between 34077 and 33193 for ORF6. As mentioned early (shown in Table 2A), splicing in the E4 region was the most significant at 12 hpi and constituted about 60% of all adenovirus RNA splicing. Furthermore, the splicing of RNAs in E4 region was under a temporal control. Splicing of ORF2, ORF3, ORF4 and one of the new site 35547–34330 reached their maximums at 12 hpi, whereas ORF6, ORF6/7, as well as the site 33284–33903 and 35547–34242 reached their highest level at 24 hpi. Temporal regulation of the spicing pattern has been shown in previous studies (Dix and Leppard, 1993; Ross and Ziff, 1992; Tigges and Raskas, 1984). Based on their expression pattern, E4 mRNAs are divided into two classes: mRNAs of early classes included ORF2, ORF3, ORF4 and ORF6 synthesized during the early phase of infection; mRNAs of the later class included ORF1 and ORF6/7 and are synthesized later. Our result suggested that the synthesis of ORF6 and ORF6/7 mRNA started from the early phase at a low level, but their most efficient production occurred at the late phase.

*Splicing of RNA from the E2 transcription unit*

The splice junctions in the E2 region form a complicated pattern. All previously identified splice sites were detected in both data sets (Fig. 4). The splice junction for E2A (at nts 24714–24089) and the common splice junction for E2A and E2B (24792–27024) had the highest coverage. The sequence reads covering the splice junctions in the UXP mRNA (30855–24792) and in the mRNA that is transcribed from the late E2 promoter (25885–24792) E2A-L were also very high. They reached their peak at 24 hpi, and then decreased. The coverage of the splice junction that is unique for E2B was low. In particular, the sequence reads representing the pol mRNA were very low in both data sets. Many new splice junctions were identified within the E2 region, most of them being located within the large splice site between 24714 and 14293. However,




| Splice junction | | Sequence reads from paired-end seq | | | | | mRNA |
|---|---|---|---|---|---|---|---|
| Start | stop | Mock | Ad-6 hpi | Ad-12 hpi | Ad-24 hpi | Ad-36 hpi | |
| 35547 | 35108 | 0 | 7 | 164 | 56 | 11 | E4orf2 |
| 35547 | 34736 | 18 | 177 | 19628 | 6043 | 681 | E4orf3 |
| 35547 | 34436 | 0 | 2 | 1124 | 346 | 37 | E4orf4 |
| 35547 | 34083 | 1 | 4 | 149 | 709 | 337 | E4orf6 |
| 33903 | 33193 | 0 | 0 | 231 | 793 | 635 | E4orf6/7 |
| 33903 | 33284 | 0 | 0 | 44 | 117 | 85 | Previous identified |
| 34288 | 34083 | 0 | 0 | 0 | 21 | 0 | Previous identified |
| 34605 | 34436 | 0 | 0 | 56 | 8 | 0 | Previous identified |
| 35547 | 34242 | 0 | 0 | 51 | 160 | 55 | New |
| 35547 | 34330 | 2 | 1 | 1059 | 421 | 68 | New |

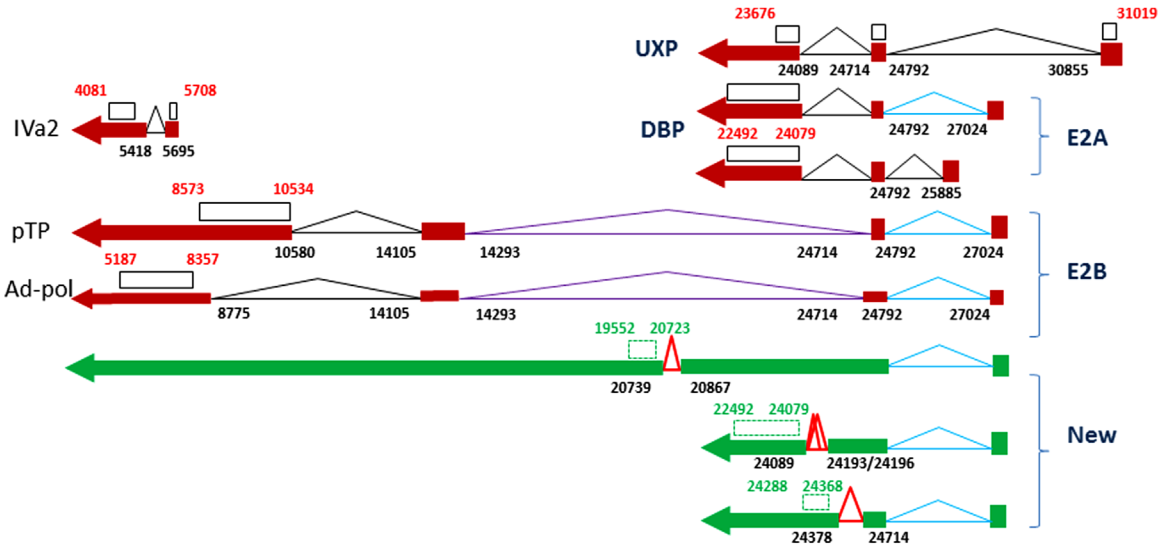
**Fig. 3.** A splicing map of the E4 transcription unit. The red arrows represent the E4 mRNAs. The thickness of the arrow indicates the relative abundance of the sequence reads covering the splice junction. The thin lines indicate the position of splice sites. The open boxes are ORFs. The black and red numbers indicate the positions of splice sites and ORFs, respectively. The dashed boxes are predicted ORFs for the new splice junction. The gray arrow, lines and box indicate an mRNA which was undetectable in the present study. The sequence reads covering the splice sites are listed in the table at the bottom.

most of them could only be detected by paired-end sequencing. Only four novel splice junctions, 20867-20739, 24193-24089, 24196-24089 and 24714-24378, could be detected in both data sets. Splicing at these sites occurred mostly during the early phase starting at 12 hpi and reaching the highest level at 24 hpi. The new splice junctions at 24193/24196-24089 and 24714-24378 might be alternative splices for E2A since the latter shares its splice donor site and the former shares its splice acceptor site with the major E2A mRNA. The sequence reads covering these two new splice junctions were, however, about six time lower than those covering the major E2A splice site (24714-24089), suggesting that they play a less significant role if they indeed are alternative splice sites for E2A. The sequence reads covering the splice junction 20867-20739 were high as compared with those for pTP and Ad-pol (14105-8775, 14105-10580 and 24714-14293).The E2B mRNAs share their poly(A) site with the IVa2 mRNA which is expressed during the intermediate phase of the infection. The previously described splice site (5695-5418) in the latter mRNA was confirmed by both single- and paired-end sequencing.

*Splices in the major late transcription unit*

The majority of the splice junctions at 24 and 36 hpi were located on the r-strand, and furthermore, about 80% of them were located within the MLTU. The sequence reads that covered the splice junctions within the tripartite leader region were most abundant after 24 hpi. Splicing occurred predominately between leaders 1 and 2 as well as leaders 2 and 3, and these splice junctions account for approximately 80% and 90% of all the detected splice junctions within the tripartite leader region at 24 and 36 hpi, respectively. However, a variety of alternative splice junctions were also detected within the leader region, especially in the data set obtained by paired-end sequencing and the most common splice junctions detected in both sequencing experiments are shown in Fig. 5. The splices between leaders 2 and i and between leaders i and 3 account for 11.6% and 4.5% at 24 and 36 hpi, respectively. Nearly all the remaining splice junctions contributed each to less than 1% of all splice junctions in this region, except the one between nts 9113 and 9197 which reached 1.2% at 24 hpi. It differs from most of the other aberrant splices in the leader region by not employing any of the known 5′ or 3′splice sites, involved in splicing of the tripartite leader. The low abundance of these splice sites makes it unlikely that these splice sites play important functions.

A large set of splice junctions had a common 5′ junction at 9724 as shown in Table 3. They coincided with the known splice sites for the capsid proteins except for pVII and pVIII. Previously identified 3′-splice sites at 15853 and 27194 for these polypeptides were not detected (Alestrom et al., 1984; Sung et al., 1983). Instead, we identified three new candidate 3′ splice sites at 15653, 15659, and 15723 for pVII and one site at 27030 for pVIII. The first downstream ATGs in the corresponding mRNAs were 15873 and 27215 for pVII and pVIII, respectively, suggesting that they are likely to be the correct splice sites for pVII and pVIII. To




| Splice junction | | Sequence reads from paired-end seq | | | | | mRNA |
|---|---|---|---|---|---|---|---|
| Start | End | Mock | Ad-6 hpi | Ad-12 hpi | Ad-24 hpi | Ad-36 hpi | |
| 30855 | 24792 | 0 | 0 | 24 | 9348 | 5599 | UXP |
| 27024 | 24792 | 3 | 16 | 2135 | 17052 | 4201 | E2A/E2B |
| 25885 | 24792 | 4 | 10 | 104 | 8729 | 3943 | E2A-L |
| 24714 | 24089 | 0 | 11 | 1652 | 24348 | 10003 | E2A |
| 24714 | 14293 | 0 | 0 | 93 | 2110 | 888 | Adv-pol & pTP |
| 14105 | 10580 | 0 | 2 | 0 | 674 | 1598 | pTP |
| 14105 | 8775 | 0 | 0 | 0 | 11 | 0 | pol |
| 5695 | 5418 | 1 | 0 | 6 | 2679 | 2429 | IVa2 |
| 20867 | 20739 | 0 | 0 | 0 | 7256 | 543 | New |
| 24196 | 24089 | 0 | 0 | 575 | 1896 | 387 | New |
| 24193 | 24089 | 0 | 0 | 498 | 1981 | 372 | New |
| 24714 | 24378 | 1 | 1 | 746 | 3888 | 1450 | New |

**Fig. 4.** A splicing map for the E2 transcription unit. The red arrows represent the known mRNAs. The green arrows are mRNAs predicted from new splice sites. The thin lines indicate the positions of splice sites on the Ad2 genome. The black and red numbers indicate the positions of splice sites and ORFs on the Ad2 genome, respectively. The red lines are newly identified splice sites. The green dashed boxes and numbers are predicted ORFs and their positions. The sequence reads covering the splice sites are listed in the table at the bottom.

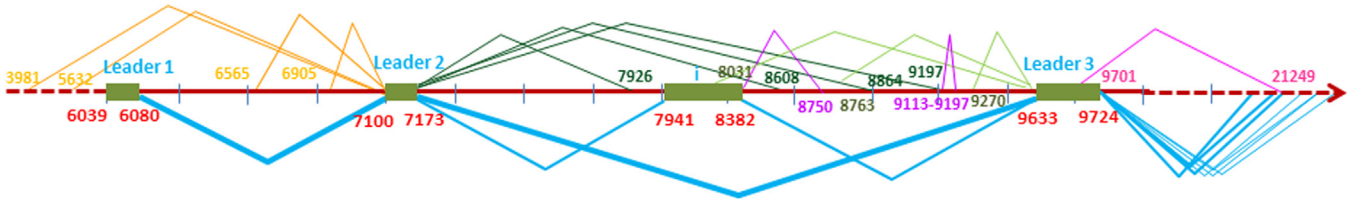
confirm these results, two PCR experiments were performed to analyze the new splice sites of pVII. Three primers were designed, two forward primers, F1- and F2-primer, and one reverse primer, R-primer. F1 primer covered the splice junction at 9724–15723, while F2-primer and R-primer located upstream and downstream splice of all three splice sites. As shown in Fig. 6, only one PCR product was produced when F1- and R-primers were used. The size of the band appeared to be close to 200 nt long, while the expected PCR product should be 193 nt. Two major bands were detected when F2-and R-primers were used. The upper band most likely represented two PCR products of 282 and 288 nt long fragment generated from mRNA with the splice junction at 9724–15653 and 9724–15659, respectively. The lower band was the PCR product of 212 nt long produced from mRNA with splice junction at 9724–15723.

Although the reported splice junction at 9724–26237 for p33K/22K was detected, it was not the dominant one. A new splice junction at 9724–26204 was identified which preceded the same ORF as 9724–26237. The sequence reads covering this junction were more than 3-fold higher. Many hitherto undetected splice junctions were identified which used 9724 as the splice donor site. Only those detected by both single-end sequencing and paired-end sequence analysis were considered and included in Table 3. The ORF prediction analysis showed that the downstream ORFs which followed the four 3′splice splice sites at 10433, 15029, 15056, and 26753 were less than 200 nt long, making it unlikely that the corresponding mRNAs encoded proteins. The ORFs following splice sites, 9724–14455 and 9724–17367, were of reasonable length, although the sequence reads covering them were low. Conservation analysis showed the downstream ORFs of splice sites 9724–14455 and 9724–10433 were highly conserved among adenoviruses belonging to different subgroups (see Supplementary material).

The fiber gene is the most distal gene in the MLTU. Previous studies have shown that the 5′end of fiber mRNA can be spliced in several different ways, resulting in 5 major species, each representing 5–70% of the total fiber mRNA (Chow and Broker, 1978; Uhlen et al., 1982). Since we could not assemble the full length mRNA, it was impossible to score all leader arrangements unambiguously. There are many combinations of putative splices as shown in Fig. 7. The sequence reads covering the splice junction between 28560 (3′-end Y leader) and 31029 were most abundant at 36 hpi, accounting for about 50% of all splice junctions between any leader and the fiber mRNA body. The highly abundant sequence reads covering the splice junction between leader 3 and leader Y indicate that the fiber mRNA with tripartite leader plus Y leader is the most abundant species late after infection.

*Mapping polyadenylation sites*

Our data also allowed us to accurately map poly(A) sites on the Ad2 genome. Furthermore, our time course experiment enabled us to quantitatively study the changes of polyadenylation following




| Splice junction | | Sequence reads from paired-end seq | | | | | Junction location |
|---|---|---|---|---|---|---|---|
| Start | End | Mock | Ad-6 hpi | Ad-12 hpi | Ad-24 hpi | Ad-36 hpi | |
| 6080 | 7100 | 352 | 223 | 187 | 166717 (30%) | 573713 (26%) | between Leader 1 and 2 |
| 7173 | 9633 | 486 | 425 | 343 | 281568 (50%) | 1426350 (64%) | between Leader 2 and 3 |
| 7173 | 7941 | 64 | 29 | 40 | 32416 (5.8%) | 45039 (2%) | between Leader 2 and i |
| 8382 | 9633 | 24 | 15 | 28 | 32309 (5.8%) | 54467 (2.5%) | Between leader i to 3/13.6K protein |
| 3981 | 7100 | 0 | 0 | 0 | 515 | 2141 | |
| 5632 | 7100 | 0 | 0 | 0 | 535 | 2184 | |
| 6565 | 7100 | 0 | 0 | 0 | 520 | 2490 | |
| 6905 | 7100 | 1 | 0 | 0 | 483 | 2118 | |
| 7173 | 7926 | 0 | 0 | 0 | 2163 | 161 | |
| 7173 | 8608 | 0 | 0 | 0 | 1428 | 262 | |
| 7173 | 8864 | 0 | 0 | 0 | 1463 | 182 | |
| 7173 | 9197 | 0 | 0 | 0 | 1598 | 1844 | |
| 8031 | 9633 | 0 | 0 | 0 | 517 | 1249 | |
| 8382 | 8750 | 1 | 0 | 0 | 108 | 1139 | |
| 8763 | 9633 | 0 | 0 | 0 | 1952 | 6443 | |
| 9113 | 9197 | 10 | 2 | 14 | 6775 (1.2%) | 9503 | |
| 9270 | 9633 | 0 | 0 | 0 | 2452 | 7239 | |
| 9701 | 21249 | 0 | 0 | 0 | 673 | 2624 | |

**Fig. 5.** A splicing map of the leader region. The blue lines represent the known splice junctions. The thickness of the lines indicates the relative abundance of the sequence reads which cover the splice junction. The numbers indicate the position of splice sites on theAd2 genome. The sequence reads covering these splice sites are listed in the table at the bottom.

**Table 3**
The most significant splice junctions in the MLTU.

| Region | Splice junction | | ORF | | | Sequence reads from paired-end seq | | | | | mRNA |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Start | End | Begin | Stop | length | Mock | Ad-6 hpi | Ad-12 hpi | Ad-24 hpi | Ad-36 hpi | |
| Leader | 6080 | 7100 | | | | 352 | 223 | 187 | 166,717 | 573,713 | Between leaders 1 and 2 |
| Leader | 7173 | 7941 | | | | 64 | 29 | 40 | 32,416 | 45,039 | Between leaders 2 and i |
| Leader | 7173 | 9633 | | | | 486 | 425 | 343 | 281,568 | 142,6350 | Between leaders 2 and 3 |
| Leader | 8382 | 9633 | | | | 24 | 15 | 28 | 32,309 | 54,467 | 13.6K protein |
| | 9724 | 10433 | 10434 | 10583 | 150 | 0 | 0 | 0 | 727 | 1198 | New |
| L1 | 9724 | 11039 | 11040 | 12287 | 1248 | 41 | 33 | 96 | 48,483 | 75,495 | 52, 55K |
| L1 | 9724 | 12307 | 12308 | 14065 | 1758 | 0 | 0 | 0 | 1332 | 4577 | IIIa |
| L2 | 9724 | 14149 | 14151 | 15866 | 1716 | 12 | 18 | 3 | 3683 | 25,497 | Capsid protein pIII |
| | 9724 | 14455 | 14541 | 15866 | 1326 | 0 | 0 | 0 | 453 | 2536 | New |
| | 9724 | 15029 | 15043 | 15153 | 111 | 0 | 0 | 0 | 1329 | 2531 | New |
| | 9724 | 15056 | 15064 | 15153 | 90 | 0 | 0 | 0 | 5435 | 1528 | New |
| L2 | 9724 | 15518 | 15873 | 16469 | 597 | 0 | 0 | 0 | 340 | 1413 | New |
| L2 | 9724 | 15653 | 15873 | 16469 | 597 | 0 | 0 | 0 | 674 | 4509 | New pVII (instead of 15853) |
| L2 | 9724 | 15659 | 15873 | 16469 | 597 | 8 | 16 | 11 | 2824 | 22,056 | New pVII (instead of 15853) |
| L2 | 9724 | 15723 | 15873 | 16469 | 597 | 5 | 9 | 2 | 3279 | 13,724 | New pVII (instead of 15853) |
| L2 | 9724 | 16515 | 16539 | 17648 | 1110 | 32 | 32 | 18 | 16289 | 89,029 | Core protein pV |
| | 9724 | 17367 | 17428 | 17763 | 336 | 0 | 0 | 0 | 463 | 1956 | New |
| L2 | 9724 | 17673 | 17676 | 17918 | 243 | 6 | 12 | 5 | 3142 | 17,648 | Core prot precursor pX |
| L3 | 9724 | 17999 | 18001 | 18753 | 753 | 7 | 10 | 11 | 3092 | 26,407 | Capsid prot precursor pVI |
| L3 | 9724 | 18801 | 18838 | 21744 | 2907 | 13 | 14 | 11 | 5418 | 39,521 | Hexon pII |
| L3 | 9724 | 21649 | 21778 | 22392 | 615 | 10 | 1 | 3 | 807 | 6104 | Protease |
| L4 | 9724 | 24094 | 24108 | 26525 | 2418 | 31 | 24 | 12 | 27968 | 61,766 | 100K protein |
| L4 | 9724 | 26204 | 26239 | 26826 | 588 | 13 | 2 | 0 | 4957 | 14,558 | New L4-33K (instead of 26237) |
| L4 | 9724 | 26237 | 26239 | 26826 | 588 | 5 | 2 | 3 | 1673 | 4627 | L4-33K and encapsidation prot 22K |
| L4 | 9724 | 26753 | 26952 | 26993 | 42 | 5 | 2 | 2 | 1455 | 9498 | New |
| L4 | 26552 | 26753 | | | | 9 | 0 | 15 | 31,039 | 88,362 | L4-33K protein, 2nd intron |
| L4 | 9724 | 27030 | 27215 | 27898 | 684 | 0 | 0 | 0 | 593 | 3728 | pVIII instead of 27194 |
| L5 | 9724 | 27847 | | | | 1 | 1 | 27 | 573 | 7784 | Leaders 3–X |
| L5 | 9724 | 28375 | | | | 35 | 17 | 7 | 6978 | 74,604 | Leaders 3–Y |
| L5 | 9724 | 30437 | | | | 0 | 1 | 11 | 1111 | 3503 | Leaders 3–Z |
| L5 | 9724 | 31029 | | | | 0 | 0 | 6 | 1685 | 14,441 | Leader 3–fiber |
| L5 | 27981 | 31029 | | | | 0 | 1 | 0 | 187 | 1228 | X–fiber |
| L5 | 28560 | 31029 | | | | 16 | 15 | 11 | 5447 | 36,841 | Y–fiber |
| L5 | 30583 | 31029 | | | | 0 | 0 | 0 | 4738 | 3021 | Z–fiber |

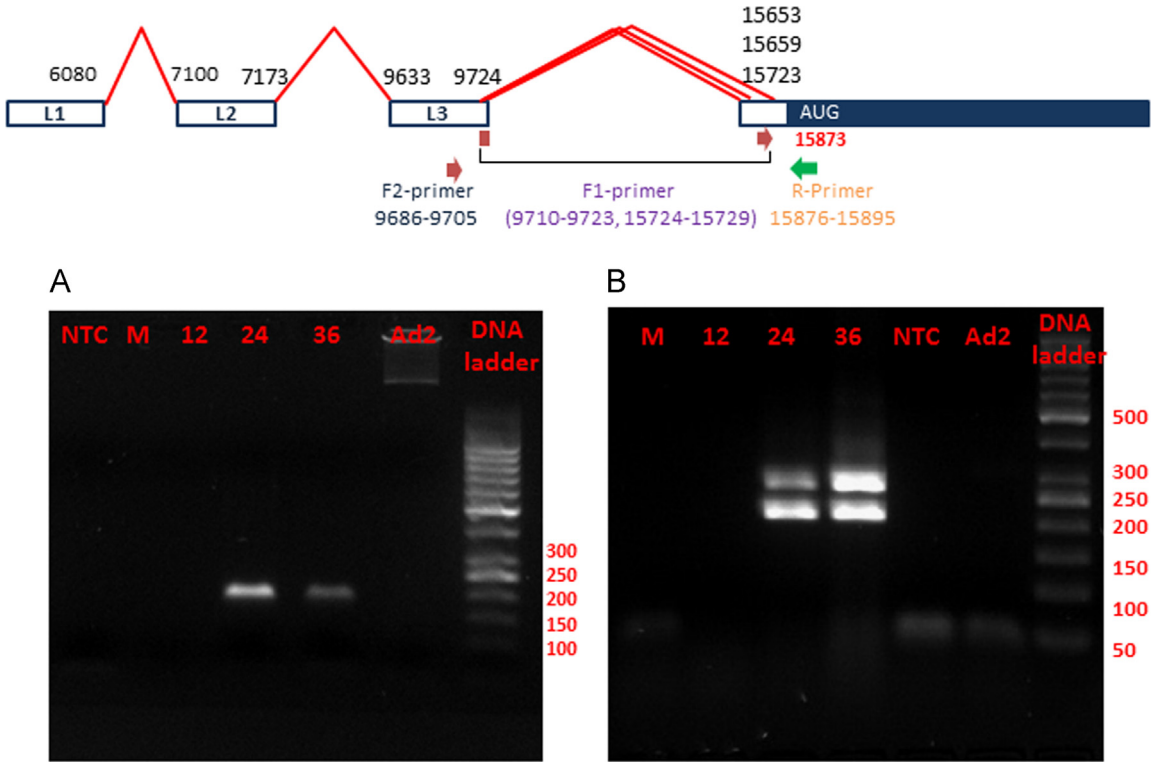
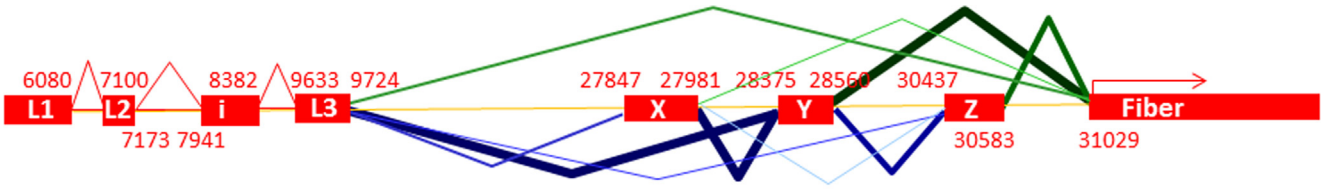



**Fig. 6.** PCR verification of new splice sites for pVII. The locations of the primers are shown at the top. Total RNA was extracted from mock (M) and Ad2-infected IMR-90 cells at 12, 24 and 36 hpi and transcribed into cDNA. PCR was performed by using F1- and R-primer (A) or F2- and R-primer (B). NTC stands for control without template. Purified DNA from adenovirus was used as control (Ad2).




| Splice junction | | Sequence reads from paired-end seq | | | | | Junction location |
|---|---|---|---|---|---|---|---|
| Start | End | Mock | Ad-6 hpi | Ad-12 hpi | Ad-24 hpi | Ad-36 hpi | |
| 6080 | 7100 | 352 | 223 | 187 | 166717 | 573713 | Leader 1-2 |
| 7173 | 7941 | 64 | 29 | 40 | 32416 | 45039 | Leader 2-i |
| 7173 | 9633 | 486 | 425 | 343 | 281568 | 1426350 | Leader 2-3 |
| 8382 | 9633 | 24 | 15 | 28 | 32309 | 54467 | Leader i-3 |
| 9724 | 27847 | 1 | 1 | 27 | 573 | 7784 | Leader 3-X |
| 9724 | 28375 | 35 | 17 | 7 | 6978 | 74604 | Leader 3- Y |
| 27981 | 28375 | 34 | 22 | 4605 | 28374 | 43507 | Leader X-Y |
| 28560 | 30437 | 4 | 21 | 429 | 6375 | 12392 | Leader Y-Z |
| 9724 | 31029 | 0 | 0 | 6 | 1685 | 14441 | Leader 3- fiber |
| 27981 | 31029 | 0 | 1 | 0 | 187 | 1228 | Leader X-fiber |
| 28560 | 31029 | 16 | 15 | 11 | 5447 | 36841 | Leader Y-Fiber |
| 30583 | 31029 | 0 | 0 | 9 | 3021 | 23336 | Leader Z-Fiber |

**Fig. 7.** A splice map of fiber RNA. The red bars represent fiber transcript. The blue and green lines indicate splice sites. The numbers indicate the position of splice sites on the Ad2 genome. The thickness of the lines indicates the relative abundance of the sequence reads covering the splice junctions. The sequence reads covering these splice sites are listed in the table.

the progression of the infection. As shown in Table 4, most of our results were in agreement with previous reports, including the poly(A) sites for E1B, L2, L3 E2B, E4 and E3B mRNAs. There was only one dominant poly(A) site for these mRNAs, i.e. more than 96% of the reads covered each respective site. However, several new poly(A) sites were identified. For example, five new L5 poly

**Table 4**
Poly(A) addition sites and the efficiency of their usage.

| mRNA | Previous identified poly-A addition site | Detected by sequencing | Percentage | | |
|---|---|---|---|---|---|
| | | | 12 hpi (%) | 24 hpi (%) | 36 hpi (%) |
| E1A | | 1626 | –[a] | 22.9[b] | – |
| | | 1628 | – | 18.6 | – |
| | 1630 | 1630 | – | 58.6 | 100 |
| E1B | 4061 | 4061 | – | 96.8 | 97 |
| L1 | 14113 (The most abundant site) | 14113 | – | 13.7 | 13.7 |
| | 14118 | 14118 | – | 73 | 69.7 |
| | – | 14112 | | 4.5 | 4.5 |
| | – | 14117 | – | 3 | 8 |
| L2 | 17969 | 17969 | – | 99.3 | 98.8 |
| L3 | 22443 | 22443 | – | 99.1 | 94.9 |
| | | 22437 | – | 0.9 | 3 |
| L4 | 28228 | 28228 | – | 97.5 | 97.2 |
| | 28223 (The most abundant site) | 28223 | – | 2.2 | 2.5 |
| L5 | | 32794 | – | 4.8 | 4.4 |
| | | 32795 | – | 32.8 | 34.4 |
| | | 32796 | – | 7.4 | 7.1 |
| | | 32797 | – | 6.9 | 6.4 |
| | 32798 | 32798 | – | 44.1 | 40 |
| | | 32803 | – | 4.1 | 7.2 |
| E3A | 29792/29799/29801/ 29804 | 29788 | – | 8.8 | – |
| | | 29791 | – | 10.2 | – |
| | | 29792 | 65.5 | 59.1 | 66.4 |
| | | 29799 | 34.5 | 21.9 | 33.6 |
| E3B | 30846 | 30846 | – | 100 | 100 |
| E2A | 22420 | 22420 | 56 | 51.2 | 54.5 |
| | | 22416 | 44 | 47 | 45.5 |
| E2B and IVa2 | 4050 | 4049 | – | 100 | 100 |
| E4 | | 32799 | 2.8 | | |
| | 32802 | 32802 | 96.8 | 100 | 100 |

[a] Not detected.
[b] Percentage of sequence reads covering this poly(A) site.

(A) addition sites were found in addition to the previously reported site at 32798 (Le Moullec et al., 1983). About 40% of the L5 mRNAs had a poly(A) site at 32798 and about 32% at 32795. Two new L1 poly (A) sites at 14112 and 14117 were identified, but they represented only 7.5% and 12.5% of all L1 poly(A) sites at 24 and 36 hpi, respectively. The poly(A) site at 14118 was the most abundant and accounted for over 70% of all L1 poly(A) sites, whereas the previously reported site at 14113 accounted for less than 14% (Le Moullec et al., 1983). Two previously identified L4 poly(A) sites at position nts 28228 and 28223 could be confirmed here, although the site at 28228, was shown to be the most abundant site for L3 (Le Moullec et al., 1983; Prescott and Falck-Pedersen, 1994). Among four E3A poly(A) sites detected here, two were consistent with earlier studies (Ahmed et al., 1982) and they were the most significant sites. A novel E2A poly(A) site at 22416 was identified but poly(A) addition at 22420 was slightly higher than at 22416.

## Conclusion

Adenovirus splice and polyadenylation sites were quantitatively mapped by using deep cDNA sequencing. As a result of our findings the functional map of the adenovirus genome needs revision. Although all adenovirus splice junction could be mapped with great precision, the full structure of all adenovirus mRNAs could not be defined. An attempt to assemble full length mRNA sequences proved impossible with available bioinformatics tools due to the extreme overlapping nature of adenovirus mRNAs. It is clear from the present results that adenovirus splicing is an inaccurate process since more than 500 different splice sites could be identified using sequence coverage of 100 as a criterion. Although a few of the novel splices are likely to generate hitherto undetected functional mRNAs it seems highly improbable that all these sites are functional. This is most apparent in the region which includes the exons of the tripartite leader. Here the strong splice donor and acceptor sites are used in combinations with numerous weaker partners. Polyadenylation, in contrast, appears very precise and at some sites the poly(A) tail was added at the same position in 97% of the cases or more. In a number of cases our results are inconsistent with previously reported findings. Although this is likely to be due to our use of a superior method, it cannot be excluded that the differences are related to the use of other host cells and virus strains.

## Materials and methods

### *Cell culture and adenovirus infection*

Human primary lung fibroblast cells (IMR-90) purchased from American Type Culture Collection (ATCC) were cultured in Eagle's minimum essential medium (ATCC) supplemented with 10% fetal bovine serum, 100 U/ml penicillin, and 100 μg/ml streptomycin. After reaching confluence, the cells were cultured for two more days in order to synchronize them. Over 95% of the cells were in the G0/G1 phase as indicated by FACS analysis (Zhao et al., 2007). Synchronized cells were mock-infected or infected with Ad2 at a multiplicity of 100 fluorescence-forming units (FFU) per cell in serum-free medium (Philipson, 1961). After 1 h adsorption at 37 °C, the medium was replaced with complete EMEM containing 10% FBS and incubated at 37 °C. Infected cells were collected at 6, 12, 24, and 36 hours post infection (hpi). Mock-infected cells were collected at 6 hpi.

### *RNA extraction, cDNA library preparation, and sequencing*

Total RNA from adenovirus or mock-infected IMR-90 cells was extracted using TRIZOL Reagent (Invitrogen). The quality of the input RNA was controlled by the Bioanalyzer (Agilent Technologies). RNA was treated with RiboZero (Epicentre) to remove ribosomal RNA and the libraries were constructed using the ScriptSeq™ v2 RNA-Seq library preparation kit according to the manufacturer's protocol (Epicentre). The cDNA libraries were sequenced on the Genome Analyzer II or the Illumina HiSeq 2000.

### *Computational analysis of sequencing data*

To evaluate the sequencing data and split virus and host cellular reads, the reads of each sample were aligned to adenovirus type 2 (http://www.ncbi.nlm.nih.gov/nuccore/NC_001405) and human (hg19, http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/chromFa.tar.gz) genome using SOAP2 (Li et al., 2009) with maximum of two mismatches. Then, the TopHat software (Trapnell et al., 2009) was used to detect and predict splice junctions in the Ad2 genome sequence also with less than 2 mismatches. Furthermore, the start and end positions of the intron should be in perfect match with those of known introns. Finally, splice junctions, detected in only one analysis, were considered as less reliable and removed from list.

Identification of the poly(A) addition site was done by using the first FASTQ file generated from strand-specific sequencing. The sequence reads that covered more than 18 adenosines (poly (A) tail) in a row were selected. The poly(A) sequence was then removed and the upstream sequences were mapped to the adenovirus genome.

*PCR validation of new pVII splice sites*

Three new alternative 3′ splice sites, 15653, 15659 and 15723 for pVII, were identified by deep sequencing. To confirm this result, a PCR experiment was performed. Three primers were designed, two forward primers (GTCACAGTCGCAAGATCACG and CCTCTCGAGAAAGGCGTCTA), and one reverse primer (CTGGGCGA-TATAAGGATGGA). The first forward primer covered the splice junction. Reverse transcription was performed by using Super-Script™ III Reverse Transcriptase (Invitrogen by life Technologies). PCR was done by using DreamTaq DNA polymerse (Thermo SCIENTIFIC) under the condition of 2 min of denaturing, then 40 cycles of 30 s at 95 °C, 30 s at 60 °C and 15 s at 72 °C and finally 10 min extension at 72 °C. The PCR products were analysis by electrophoresis on a 2% agarose gel.

## Acknowledgments

Sequencing was performed at the SNP&SEQ Technology Platform in the Uppsala University and the University Hospital. We thank Ulrika Liljedahl for excellent sequencing, and Olof Karlberg and Martin Dahlo for the help with data analysis. We thank Goran Akusjarvi for the critical reading of this manuscript and for valuable comments. NN at UPPMAX is acknowledged for the assistance concerning technical and implementational aspects in making the code run on the UPPMAX resources. This work was supported by the Kjell and Märta Beijer Foundation.

## Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.virol.2014.04.006.

## References

Ahmed, C.M., Chanda, R., Stow, N., Zain, B.S., 1982. The sequence of 3′-termini of mRNAs from early region III of adenovirus 2. Gene 19, 297–301.
Akusjarvi, G., Persson, H., 1981. Gene and mRNA for precursor polypeptide VI from adenovirus type 2. J. Virol. 38, 469–482.
Alestrom, P., Akusjarvi, G., Lager, M., Yeh-kai, L., Pettersson, U., 1984. Genes encoding the core proteins of adenovirus type 2. J. Biol. Chem. 259, 13980–13985.
Babich, A., Nevins, J.R., 1981. The stability of early adenovirus mRNA is controlled by the viral 72 kD DNA-binding protein. Cell 26, 371–379.
Berk, A.J., Sharp, P.A., 1978. Structure of the adenovirus 2 early mRNAs. Cell 14, 695–711.
Chebli, K., Gattoni, R., Schmitt, P., Hildwein, G., Stevenin, J., 1989. The 216-nucleotide intron of the E1A pre-mRNA contains a hairpin structure that permits utilization of unusually distant branch acceptors. Mol. Cell. Biol. 9, 4852–4861.
Chow, L.T., Broker, T.R., 1978. The spliced structures of adenovirus 2 fiber message and the other late mRNAs. Cell 15, 497–510.
Chow, L.T., Broker, T.R., Lewis, J.B., 1979. Complex splicing patterns of RNAs from the early regions of adenovirus-2. J. Mol. Biol. 134, 265–303.
Cox, M.P., Peterson, D.A., Biggs, P.J., 2010. SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinform. 11, 485.
Dix, I., Leppard, K.N., 1993. Regulated splicing of adenovirus type 5 E4 transcripts and regulated cytoplasmic accumulation of E4 mRNA. J. Virol. 67, 3226–3231.
Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G.K., Khatun, J., Williams, B.A., Zaleski, C., Rozowsky, J., Roder, M., Kokocinski, F., Abdelhamid, R.F., Alioto, T., Antoshechkin, I., Baer, M.T., Bar, N.S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M.J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O.J., Park, E., Persaud, K., Preall, J.B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.H., Shahab, A., Skancke, J., Suzuki, A.M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S.E., Hannon, G., Giddings, M.C., Ruan, Y., Wold, B., Carninci, P., Guigo, R., Gingeras, T.R., 2012. Landscape of transcription in human cells. Nature 489, 101–108.
Freyer, G.A., Katoh, Y., Roberts, R.J., 1984. Characterization of the major mRNAs from adenovirus 2 early region 4 by cDNA cloning and sequencing. Nucleic Acids Res. 12, 3503–3519.
Gattoni, R., Schmitt, P., Stevenin, J., 1988. in vitro splicing of adenovirus E1A transcripts: characterization of novel reactions and of multiple branch points abnormally far from the 3′ splice site. Nucleic Acids Res. 16, 2389–2409.
Gooding, L.R., 1992. Virus proteins that counteract host immune defenses. Cell 71, 5–7.
Gooding, L.R., Wold, W.S., 1990. Molecular mechanisms by which adenoviruses counteract antiviral immune defenses. Crit. Rev. Immunol. 10, 53–71.
Green, M., Pina, M., Kimes, R., Wensink, P.C., MacHattie, L.A., Thomas Jr., C.A., 1967. Adenovirus DNA. I. Molecular weight and conformation. Proc. Natl. Acad. Sci. USA 57, 1302–1309.
Hales, K.H., Birk, J.M., Imperiale, M.J., 1988. Analysis of adenovirus type 2 L1 RNA 3′-end formation in vivo and in vitro. J. Virol. 62, 1464–1468.
Herisse, J., Rigolet, M., de Dinechin, S.D., Galibert, F., 1981. Nucleotide sequence of adenovirus 2 DNA fragment encoding for the carboxylic region of the fiber protein and the entire E4 region. Nucleic Acids Res. 9, 4023–4042.
Kitchingman, G.R., Westphal, H., 1980. The structure of adenovirus 2 early nuclear and cytoplasmic RNAs. J. Mol. Biol. 137, 23–48.
Larsson, S., Svensson, C., Akusjarvi, G., 1992. Control of adenovirus major late gene expression at multiple levels. J. Mol. Biol. 225, 287–298.
Le Moullec, J.M., Akusjarvi, G., Stalhandske, P., Pettersson, U., Chambraud, B., Gilardi, P., Nasri, M., Perricaudet, M., 1983. Polyadenylic acid addition sites in the adenovirus type 2 major late transcription unit. J. Virol. 48, 127–134.
Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., Wang, J., 2009. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25, 1966–1967.
Morris, S.J., Scott, G.E., Leppard, K.N., 2010. Adenovirus late-phase infection is controlled by a novel L4 promoter. J. Virol. 84, 7096–7104.
Perricaudet, M., Akusjarvi, G., Virtanen, A., Pettersson, U., 1979. Structure of two spliced mRNAs from the transforming region of human subgroup C adenoviruses. Nature 281, 694–696.
Pettersson, U., Tibbetts, C., Philipson, L., 1976. Hybridization maps of early and late messenger RNA sequences on the adenovirus type 2 genome. J. Mol. Biol. 101, 479–501.
Philipson, L., 1961. Adenovirus assay by the fluorescent cellcounting procedure. Virology 15, 263–268.
Prescott, J., Falck-Pedersen, E., 1994. Sequence elements upstream of the 3′ cleavage site confer substrate strength to the adenovirus L1 and L3 polyadenylation sites. Mol. Cell. Biol. 14, 4682–4693.
Rigolet, M., Galibert, F., 1984. Organization and expression of the E4 region of adenovirus 2. Nucleic Acids Res. 12, 7649–7661.
Ross, D., Ziff, E., 1992. Defective synthesis of early region 4 mRNAs during abortive adenovirus infections in monkey cells. J. Virol. 66, 3110–3117.
Schmitt, P., Gattoni, R., Keohavong, P., Stevenin, J., 1987. Alternative splicing of E1A transcripts of adenovirus requires appropriate ionic conditions in vitro. Cell 50, 31–39.
Sharp, P.M., Meador, R.C., Martin, R.R., 1974. A case of mixed anaerobic infection of the jaw. J. Oral Surg. 32, 457–459.
Shaw, A.R., Ziff, E.B., 1980. Transcripts from the adenovirus-2 major late promoter yield a single early family of 3′ coterminal mRNAs and five late families. Cell 22, 905–916.
Shenk, T., 1996. Adenoviridae: The Viruses and Their Replication, 3rd ed. Lippincott-Raven, Publishers, Philadelphia.
Sittler, A., Gallinaro, H., Jacob, M., 1994. Upstream and downstream cis-acting elements for cleavage at the L4 polyadenylation site of adenovirus-2. Nucleic Acids Res. 22, 222–231.
Smart, J.E., Stillman, B.W., 1982. Adenovirus terminal protein precursor. Partial amino acid sequence and the site of covalent linkage to virus DNA. J. Biol. Chem. 257, 13499–13506.
Spector, D.J., McGrogan, M., Raskas, H.J., 1978. Regulation of the appearance of cytoplasmic RNAs from region 1 of the adenovirus 2 genome. J. Mol. Biol. 126, 395–414.
Stillman, B.W., Lewis, J.B., Chow, L.T., Mathews, M.B., Smart, J.E., 1981. Identification of the gene and mRNA for the adenovirus terminal protein precursor. Cell 23, 497–508.
Stillman, B.W., Tamanoi, F., Mathews, M.B., 1982. Purification of an adenovirus-coded DNA polymerase that is required for initiation of DNA replication. Cell 31, 613–623.
Sung, M.T., Cao, T.M., Coleman, R.T., Budelier, K.A., 1983. Gene and protein sequences of adenovirus protein VII, a hybrid basic chromosomal protein. Proc. Natl. Acad. Sci. USA 80, 2902–2906.
Svensson, C., Pettersson, U., Akusjarvi, G., 1983. Splicing of adenovirus 2 early region 1A mRNAs is non-sequential. J. Mol. Biol. 165, 475–495.
Tigges, M.A., Raskas, H.J., 1984. Splice junctions in adenovirus 2 early region 4 mRNAs: multiple splice sites produce 18 to 24 RNAs. J. Virol. 50, 106–117.
Tollefson, A.E., Krajcsi, P., Pursley, M.H., Gooding, L.R., Wold, W.S., 1990a. A 14,500 MW protein is coded by region E3 of group C human adenoviruses. Virology 175, 19–29.
Tollefson, A.E., Krajcsi, P., Yei, S.P., Carlin, C.R., Wold, W.S., 1990b. A 10,400-molecular-weight membrane protein is coded by region E3 of adenovirus. J. Virol. 64, 794–801.

Tollefson, A.E., Ying, B., Doronin, K., Sidor, P.D., Wold, W.S., 2007. Identification of a new human adenovirus protein encoded by a novel late l-strand transcription unit. J. Virol. 81, 12918–12926.
Tormanen, H., Backstrom, E., Carlsson, A., Akusjarvi, G., 2006. L4-33K, an adenovirus-encoded alternative RNA splicing factor. J. Biol. Chem. 281, 36510–36517.
Trapnell, C., Pachter, L., Salzberg, S.L., 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105–1111.
Uhlen, M., Svensson, C., Josephson, S., Alestrom, P., Chattapadhyaya, J.B., Pettersson, U., Philipson, L., 1982. Leader arrangement in the adenovirus fiber mRNA. EMBO J. 1, 249–254.
Virtanen, A., Gilardi, P., Naslund, A., LeMoullec, J.M., Pettersson, U., Perricaudet, M., 1984. mRNAs from human adenovirus 2 early region 4. J. Virol. 51, 822–831.
Virtanen, A., Pettersson, U., 1983. The molecular structure of the 9S mRNA from early region 1A of adenovirus serotype 2. J. Mol. Biol. 165, 496–499.
Wold, W.S., Tollefson, A.E., Hermiston, T.W., 1995. E3 transcription unit of adenovirus. Curr. Top. Microbiol. Immunol. 199 (Pt. 1), 237–274.
Wu, K., Guimet, D., Hearing, P., 2013. The adenovirus L4-33K protein regulates both late gene expression patterns and viral DNA packaging. J. Virol. 87, 6739–6747.
Ying, B., Tollefson, A.E., Wold, W.S., 2010. Identification of a previously unrecognized promoter that drives expression of the UXP transcription unit in the human adenovirus type 5 genome. J. Virol. 84, 11470–11478.
Zhao, H., Dahlo, M., Isaksson, A., Syvanen, A.C., Pettersson, U., 2012. The transcriptome of the adenovirus infected cell. Virology 424, 115–128.
Zhao, H., Granberg, F., Pettersson, U., 2007. How adenovirus strives to control cellular gene expression. Virology 363, 357–375.