Quaye *et al. Virology Journal* (2024) 21:175
https://doi.org/10.1186/s12985-024-02449-0

Virology Journal

RESEARCH Open Access

Characterizing the splice map of Turkey Hemorrhagic Enteritis Virus

Abraham Quaye[1†], Brett E. Pickett[1], Joel S. Griffitts[1], Bradford K. Berges[1] and Brian D. Poole[1*]

Abstract

Background Hemorrhagic enteritis, caused by *Turkey Hemorrhagic Enteritis Virus (THEV)*, is a disease affecting turkey poults characterized by immunosuppression and bloody diarrhea. An avirulent THEV strain that retains the immunosuppressive ability is used as a live vaccine. Characterizing the splice map of THEV is an essential step that would allow studies of individual genes mediating its immunosuppressive functions. We used RNA sequencing to characterize the splice map of THEV for the first time, providing key insights into the THEV gene expression and mRNA structures.

Methods After infecting a turkey B-cell line with the vaccine strain, samples in triplicates were collected at 4-, 12-, 24-, and 72-hours post-infection. Total RNA was extracted, and poly-A-tailed mRNA sequenced. Reads were mapped to the THEV genome after trimming and transcripts assembled with StringTie. We performed PCR of THEV cDNA, cloned the PCR products, and used Sanger sequencing to validate all identified splice junctions.

Results Researchers previously annotated the THEV genome as encoding 23 open reading frames (ORFs). We identified 29 spliced transcripts from our RNA sequencing data, all containing novel exons although some exons matched some previously annotated ORFs. The three annotated splice junctions were also corroborated by our data. During validation we identified five additional unique transcripts, a subset of which were further validated by 3' rapid amplification of cDNA ends (3' RACE). Thus, we report that the genome of THEV contains 34 transcripts with the coding capacity for all annotated ORFs. However, we found six of the previously annotated ORFs to be truncated ORFs on the basis of the identification of an in-frame upstream start codon or the detection of additional coding exons. We also identified three of the annotated ORFs with longer or shorter isoforms, and seven novel unannotated ORFs that could potentially be translated; although it is beyond the scope of this manuscript to investigate whether they are translated.

Conclusions Similar to human adenoviruses, all THEV transcripts are spliced and organized into five transcription units under the control of their cognate promoters. The genes are expressed under temporal regulation and THEV also produces multiple distinctly spliced transcripts that code for the same protein. Studies of the newly identified potential proteins should be urgently performed as these proteins may have roles in THEV-induced

[†]Abraham Quaye first author.

*Correspondence:
Brian D. Poole
brian_poole@byu.edu

Full list of author information is available at the end of the article

immunosuppression. Also, knowing the splicing of THEV genes should be invaluable to future research focusing on studying THEV genes, as this will allow accurate cloning of the mRNAs.

Keywords Alternative splicing, Turkey hemorrhagic enteritis virus, Adenovirus, Transcriptome, RNA sequencing

Background

Adenoviruses (AdVs) are non-enveloped icosahedral-shaped DNA viruses, causing infection in virtually all types of vertebrates studied to date. Their double-stranded linear DNA genomes range between 26 and 45 kb in size, producing a broad repertoire of transcripts via highly complex alternative splicing patterns [1, 2]. The AdV genome is one of the most optimally economized; both the forward and reverse DNA strands harbor protein-coding genes, making it highly gene-dense. There are 16 genes termed "genus-common" that are homologous in all AdVs, presumably inherited from a common ancestor. All other genes are termed "genus-specific". The genus-specific genes tend to be located at the termini of the genome while genus-common genes are usually towards the center of the genome [1]. This pattern is also observed in *Poxviridae* and *Herpesviridae*, which also have linear DNA genomes [1, 3, 4]. The family *Adenoviridae* consists of five genera: *Mastadenovirus* (MAdV), *Aviadenovirus*, *Atadenovirus*, *Ichtadenovirus*, and *Siadenovirus* (SiAdV) to which turkey adenovirus 3 also called turkey hemorrhagic enteritis virus (THEV) belongs [5–10]. Members of SiAdV have the smallest genome size (~26 kb) and gene content of all known AdVs, with five genus-specific genes of undefined functions (see Fig. 1) [1, 2, 6].

Virulent THEV strains (THEV-V) and avirulent strains (THEV-A) of THEV both infect turkeys, with THEV-V causing hemorrhagic enteritis (HE), a debilitating acute disease predominantly affecting turkey poults characterized by immunosuppression, intestinal lesions leading to bloody diarrhea, and up to 80% mortality [2, 11–13]. While the current vaccine strain (a THEV-A called Virginia Avirulent Strain [VAS]) has proven effective at preventing HE in turkey poults, it still retains its immunosuppressive ability. Thus, vaccinated birds are rendered more susceptible to opportunistic infections and death than unvaccinated birds leading to substantial economic losses [11, 14–16]. To eliminate the immunosuppressive effect of the vaccine strain, a thorough investigation of the culprit viral genes mediating this phenomenon is essential. However, the transcriptome (splicing and gene expression patterns) of THEV has not been characterized, making an investigation of specific immunosuppressive viral genes impractical.

A myriad of studies have elucidated the AdV transcriptome in fine detail [17, 18]. However, a large preponderance of studies focus on MAdVs – specifically human AdVs. Thus, most of the current AdV gene expression and replication knowledge is based on MAdV studies, which is generalized for all other AdVs [10, 19]. MAdV transcription is temporally regulated; their genes are categorized into five early transcription units (E1A, E1B, E2, E3, and E4), two intermediate (IM) units (pIX and IVa2), and one major late transcription unit (MLTU or major late promoter [MLP] region), which generates five families of late mRNAs (L1-L5) based on the polyadenylation site. An additional gene (UXP or U exon) is located on the reverse strand. The early genes encode non-structural proteins such as enzymes or host-cell modulating proteins, primarily involved in DNA replication, or providing the necessary intracellular niche for optimal replication while late genes encode structural proteins that act as capsid proteins, promote virion assembly, or direct genome packaging. The immediate early genes E1A are expressed first, followed by the delayed early genes, E1B, E2, E3 and E4. Then the intermediate early genes, IVa2 and pIX are expressed followed by the late genes [10, 17, 18]. It is noteworthy that the MLP shows basal transcriptional activity during early infection (before DNA replication), with a comparable efficiency to other early viral promoters, but it reaches its maximal activity during late infection (after DNA replication). However, during early infection only a subset of the MLP-derived transcripts are expressed [10]. MAdVs make extensive use of alternative RNA splicing and polyadenylation to produce a very complex array of mRNAs. All but the pIX mRNA undergo at least one splicing event. For instance, the MLTU produces over 20 distinct splice variants all containing three non-coding exons at the 5'-end (collectively known as the tripartite leader; TPL) [17, 18]. There is also an alternate three-exon 5' non-coding leader sequence present in varying amounts on a subset of MLTU mRNAs (known as the x-, y-, and z-leaders). Lastly, there is the i-leader exon, which is infrequently included between the second and third TPL exons, and codes for the i-leader protein [20]. Thus, the MLTU produces a complex repertoire of mRNA with diverse 5' untranslated regions (UTRs) spliced onto different 3' coding exons which are grouped into five different 3'-end classes (L1-L5) based on polyadenylation site. Each transcription unit (TU) contains its own promoter driving the expression of the array of mRNA transcripts produced via alternative splicing in the unit [10, 17, 18]. The promoters are activated at different phases of infection by proteins from previously activated TUs. Paradoxically, the early-to-late phase transition during infection requires the L4 gene products, 22 K and 33 K, which should only be available after the

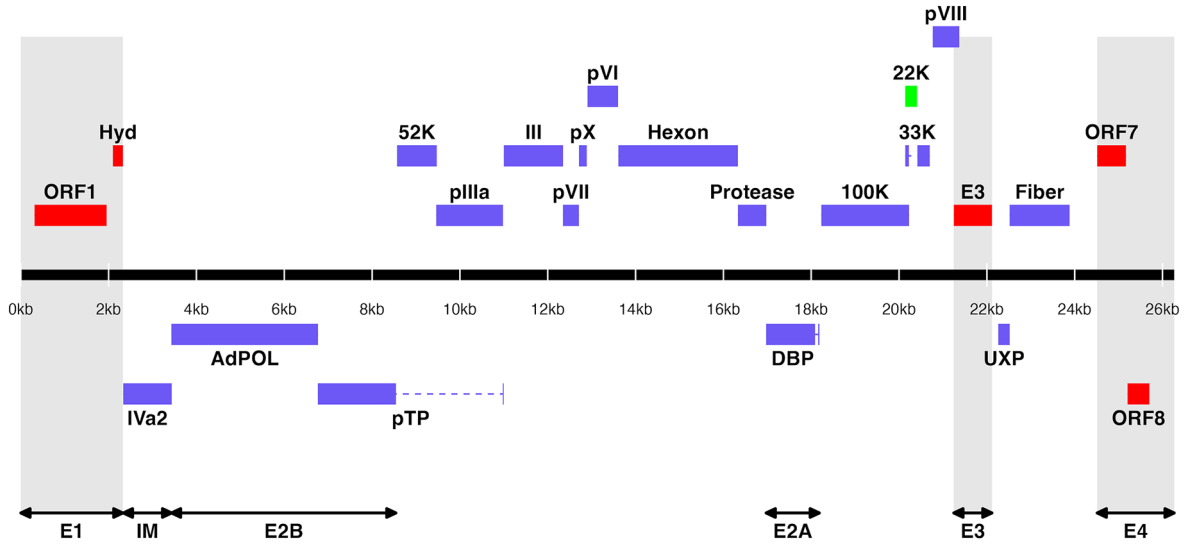


Fig. 1 Predicted ORF map of THEV avirulent strain. The central horizontal line represents the double-stranded DNA marked at 2 kb intervals as white line breaks. Colored blocks represent viral genes. Blocks above the DNA line are transcribed on the sense DNA strand and those below, on the anti-sense strand. pTP, DBP and 33 K are predicted to be spliced and are shown as two exons connected with dashed lines. Shaded regions indicate regions containing the five genus-specific genes of undefined functions (colored red). Genes colored in blue are "genus-common". The gene colored in light green is conserved in all but Atadenoviruses. Regions comprising the different TUs are labelled at the bottom (E1, E2A, E2B, E3, E4, and IM); the unlabeled regions comprise the MLTU

transition. However, a promoter in the L4 region (L4P) that directs the expression of these two proteins independent of the MLP was found, resolving the paradox [10, 17, 21]. During translation of AdV mRNA, recent studies using long-read direct RNA sequencing strongly suggest the potential usage of secondary start codons; adding to what was already a highly complex system for gene expression [17, 22].

High throughput sequencing methods have facilitated the discovery of many novel transcribed regions and splicing isoforms. It is also a very powerful tool to study alternative splicing under different conditions at an unparalleled depth [18, 22, 23]. In this paper, we use a paired-end deep sequencing experiment to characterize, for the first time, the transcriptome and splicing of THEV (VAS vaccine strain) during different phases of the infection. Our paired-end sequencing allowed for reading 149 bp long high quality (mean Phred Score of 36) sequences from each end of cDNA fragments, which were mapped to the genome of THEV.

Results

Overview of sequencing data and analysis pipeline outputs

A prior study by Aboezz et al. demonstrated that nearly all THEV transcripts became detectable starting at 4 h post-infection (hpi), with one replication cycle concluding around 18 hpi [24]. Consequently, we harvested infected MDTC-RP19 cells (multiplicity of infection (MOI) of 100 genome copy numbers/cell) at 4-, 12-, 24-, and 72-hpi to capture all transcripts within a broad time window. Our paired-end RNA sequencing (RNA-seq) experiment generated an average of 107.1 million total reads of 149 bp length per time-point. These reads were concurrently mapped to both the virus (THEV) and host (*Meleagris gallopavo*) genomes using the Hisat2 [25] reference-based aligner. A total of 18.1 million reads from all time-points mapped to the virus genome, providing comprehensive coverage and leaving no regions unmapped. The mapped reads to the virus genome increased significantly from a scant 432 reads at 4 hpi to 16.9 million reads at 72 hpi (Table 1; Fig. 2A). From these mapped reads, we identified 2,457 unique THEV splice junctions across all time-points, with later time-points exhibiting significantly more sequence reads supporting the splice junctions than earlier time-points. For instance, all 13 unique junctions at 4 hpi had fewer than 10 supporting reads each, averaging only 2.8 reads per junction. In contrast, the 2,374 unique junctions at 72 hpi averaged 898.4 reads per junction, with some junctions reaching as high as 322,677 reads. The marked increase in splice junction and mapping reads to the THEV genome over time indicates an active infection and successful viral replication, which is corroborated by our quantitative PCR (qPCR) assay that quantified the total number of viral genome copies over time (Fig. 2B).

Using StringTie [25], we assembled the data into potential transcripts, guided by the genomic locations of the previously predicted THEV ORFs. In the consolidated transcriptome, a composite of all non-redundant transcripts across all time points, we identified a total of 29 novel transcripts. We found that a subset of exons in the viral transcripts match some predicted ORFs exactly,

Table 1 Overview of sequencing results

Metric	4 h.*p*.i	12 h.*p*.i	24 h.*p*.i	72 h.*p*.i	Total
Total reads	1.17e + 08	7.63e + 07	1.20e + 08	1.15e + 08	4.28e + 08
Mapped (Host)	1.04e + 08 (89.06%)	6.79e + 07 (89.0393%)	1.06e + 08 (88.2719%)	8.38e + 07 (72.9802%)	3.62e + 08
Mapped (THEV)	4.32e + 02 (0.0004%)	6.70e + 03 (0.0088%)	1.18e + 06 (0.9841%)	1.69e + 07 (14.6904%)	1.81e + 07
Mean Per Base Coverage/Depth	2.42	37.71	6,666.96	95,041.7	101,749
Total unique splice junctions	13	37	236	2,374	2,457
Junction coverage Total (at least 1 read)	37	605	115,075	2.13e + 06	2.25e + 06
Junction coverage Mean reads	2.8	16.4	487.6	898.4	351.3
Junction coverage (at least 10 reads)	0	13	132	1,791	1,936
Junction coverage (at least 100 reads)	0	1	53	805	859
Junction coverage (at least 1000 reads)	0	0	18	168	186

with the majority of the exons being longer and spanning multiple predicted ORFs (Fig. 3).

We then validated the splice junctions in all transcripts by PCR amplification of viral cDNA, cloning, and Sanger sequencing (Supplementary PCR methods). During validation, we identified five additional transcripts, some of which were further validated by 3' Rapid Amplification of cDNA Ends (3' RACE) data. The complete list of unique splice junctions mapped to the THEV genome has been submitted to the National Center for Biotechnology Information Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) under accession number GSE254416.

Changes in THEV splicing profile over time

AdV gene expression is subject to meticulous temporal regulation, with each promoter typically generating one or a few pre-mRNAs. These pre-mRNAs undergo alternative splicing to produce a diverse array of mature mRNAs. To assess the temporal activity of each promoter, we utilized StringTie and Ballgown (a tool for statistical analysis of assembled transcriptomes) [26]. These tools estimated the normalized expression levels of all transcripts at each time point, measured in Fragments Per Kilobase of transcript per Million mapped reads (FPKM) units. At 4 hpi, we counted very few unique splice junctions, reads, and transcripts; hence, this timepoint was excluded from this analysis.

Examining individual mRNAs, TRXPT_21 – from the E2 region – was the most significantly expressed at 12 hpi, constituting 33.58% of the total expression of all transcripts. Transcripts in the E3 and E4 regions also contributed substantial proportions, along with some MLP region transcripts. The later time points were dominated by the MLP region transcripts –- TRXPT_10 and TRXPT_14 were the most abundantly expressed at 24 and 72 hpi, respectively (Fig. 4A). Our analysis of the FPKM values of transcripts per region/TU revealed a similar pattern: the E2 region was the most abundantly expressed at 12 hpi, after which the MLP region assumed dominance (Fig. 4B).

Next, we estimated the relative abundances of all splice junctions at each time point using the raw reads. Only junctions with a read coverage of at least 1% of the total splice junction reads at the given time point were considered significant and included in Tables 2a-2c. At 12 hpi, 18 junctions met the 1% threshold, predominantly from early regions (E1, E2, E3, and E4), although the MLTU was the single most predominant region overall, constituting 38.8% of all the junction reads (Table 2a and Supplementary Table S1a). The most abundant junctions at 12 hpi remained the most significantly expressed at 24 hpi. However, here, the MLP-derived junctions unsurprisingly became even more predominant overall, accounting for 45.7% of all the junction reads counted (Table 2b and Supplementary Table S1b). At 72 hpi, the trend of increased activity of the MLP continued as expected; at this time, the MLP region junctions were not only the most abundant overall –- accounting for 67.4% of all junction reads, -– but also contained the most significantly expressed individual junctions (Table 2c, Supplementary Table S1c and Fig. 4C). When we limited this analysis to only junctions in the final transcriptome, we observed the relative abundances of the junctions for each region over time to be similar to the pattern seen with all the junctions included (Fig. 4D).

Finally, we analyzed splice donor and acceptor site nucleotide usage over time to investigate any peculiarities

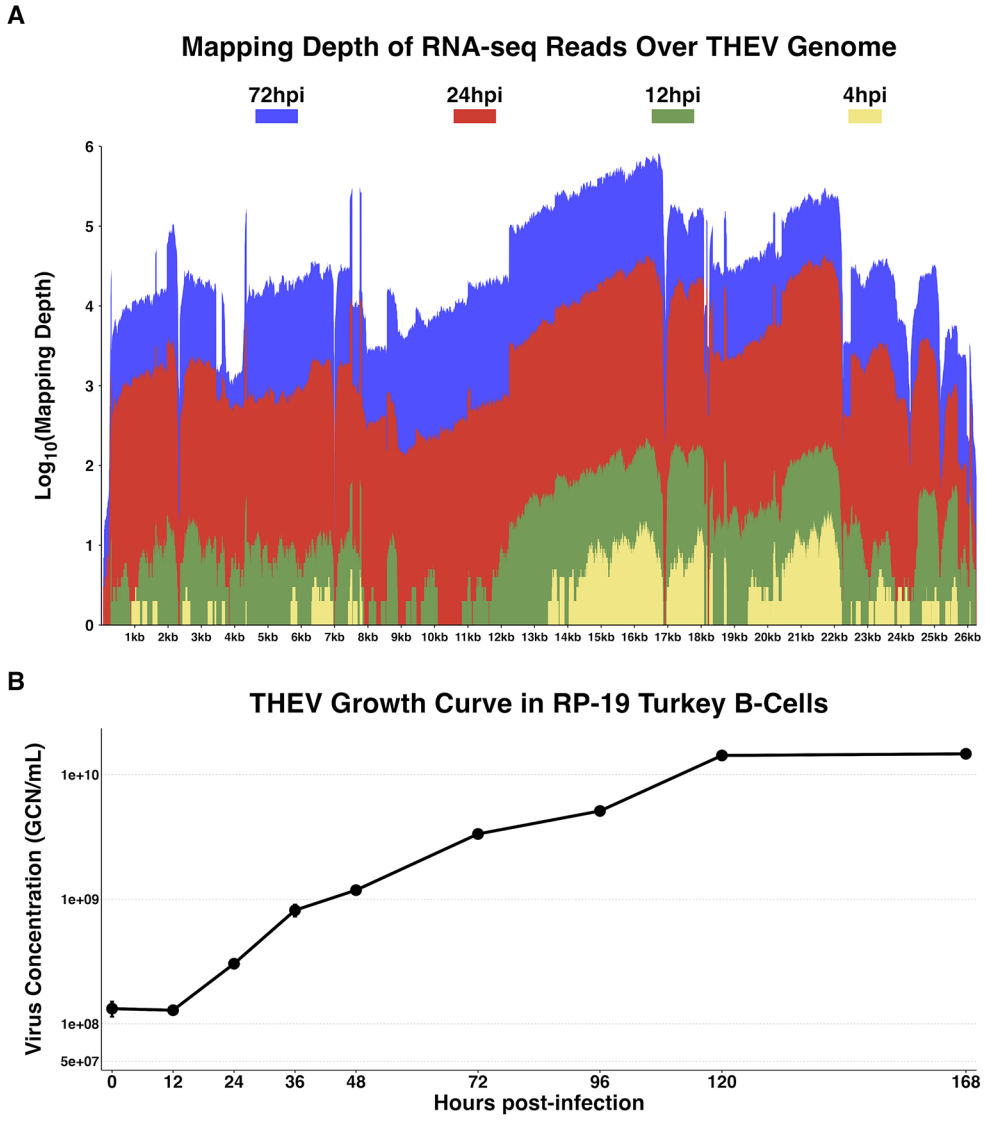


Fig. 2 Increasing levels of THEV over time. (**A**) Per base coverage of sequence reads mapping to THEV genome by time point. The pileup of mRNA reads mapping to THEV genome at the base-pair level for each indicated time point. (**B**) Growth curve of THEV (VAS vaccine strain) in MDTC-RP19 cell line. Virus quantities in the freeze-thawed supernatant from infected cells were a quantified with a qPCR assay. There is no discernible increase in virus titer up 12 hpi, after which a steady increase in virus titer is measured. The virus titer expands exponentially beginning from 48 hpi, increasing by orders of magnitude before reaching a plateau at 120 hpi. GCN: genome copy number

that THEV may exhibit, generally or over the course of the infection. We found that most splice donor-acceptor sequences were, unsurprisingly, the canonical GU-AG nucleotides. However, the splice acceptor-donor pairing became less specific over time, such that all combinations of nucleotide pairs were eventually detected (Fig. 5).

Early region 1 (E1) transcripts

In MAdVs, E1 is the first region transcribed post-viral DNA entry into the host cell nucleus, mediated solely by host transcription machinery [18]. Translated E1 proteins subsequently activate other viral promoters in conjunction with host transcription factors [10]. Despite subdivision into E1a and E1b units in MAdVs, our THEV data does not reflect this. This region is predicted to encode only two ORFs: ORF1 (sialidase) and Hyd (a hydrophobic product with unknown function) in THEV.

We identified four novel transcripts in this region, containing 3 unique splice junctions (Fig. 6), and encoding four distinct novel ORFs in addition to Hyd. All transcripts have coding potential (CP) for the Hyd protein as the 3'-most coding sequence if secondary start codon (SC) usage is considered [17, 18]. Also, all the transcripts have a common transcription termination site (TTS; at position 2325 bp), but TRXPT_1 and TRXPT_2 have an upstream transcription start sites (TSS) to TRXPT_3

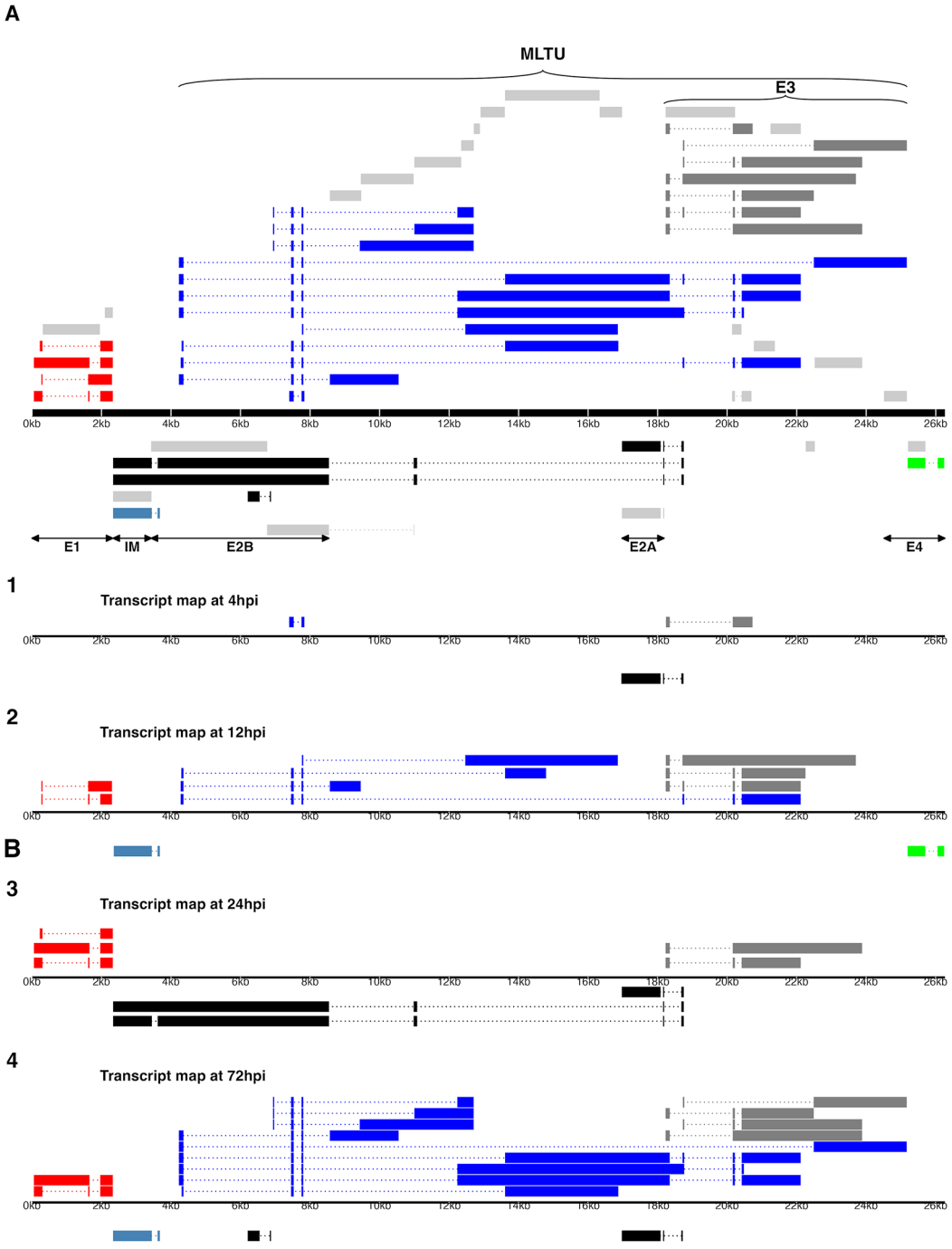


Fig. 3 (A) Transcriptome of THEV from RNA-seq. THEV transcripts assembled from all time points by StringTie are unified forming this transcriptome (splicing map). Transcripts belonging to the same TU are located in close proximity on the genome and are color coded and labeled in this figure as such. The organization of TUs in the THEV genome is unsurprisingly similar to MAdVs; however, the MAdV genome shows significantly more transcripts. The TUs are color coded: E1 transcripts - red, E2 - black, E3 - dark grey, E4 - green, MLTU - blue. Predicted ORFs are also indicated here, colored light grey. (**B**) THEV transcripts identified at given time points. Transcripts are color coded as explained in (**A**)

and TRXPT_4. Given that E1 mRNAs in MAdVs share a common TTS and TSS, differing only in the internal splicing [18], we consider the upstream TSS (position 54 bp) as the actual TSS for all E1 region transcripts. We also identified the canonical polyadenylation signal (PAS; AAUAAA) in the immediate context of the TTS at position 2323 bp (location of the "U" in the PAS sequence); see Supplementary Table S2.

From the 5'-most SC, TRXPT_1 encodes a multi-exonic novel 17.9 kDa, 160 residue protein (ORF9). TRXPT_2

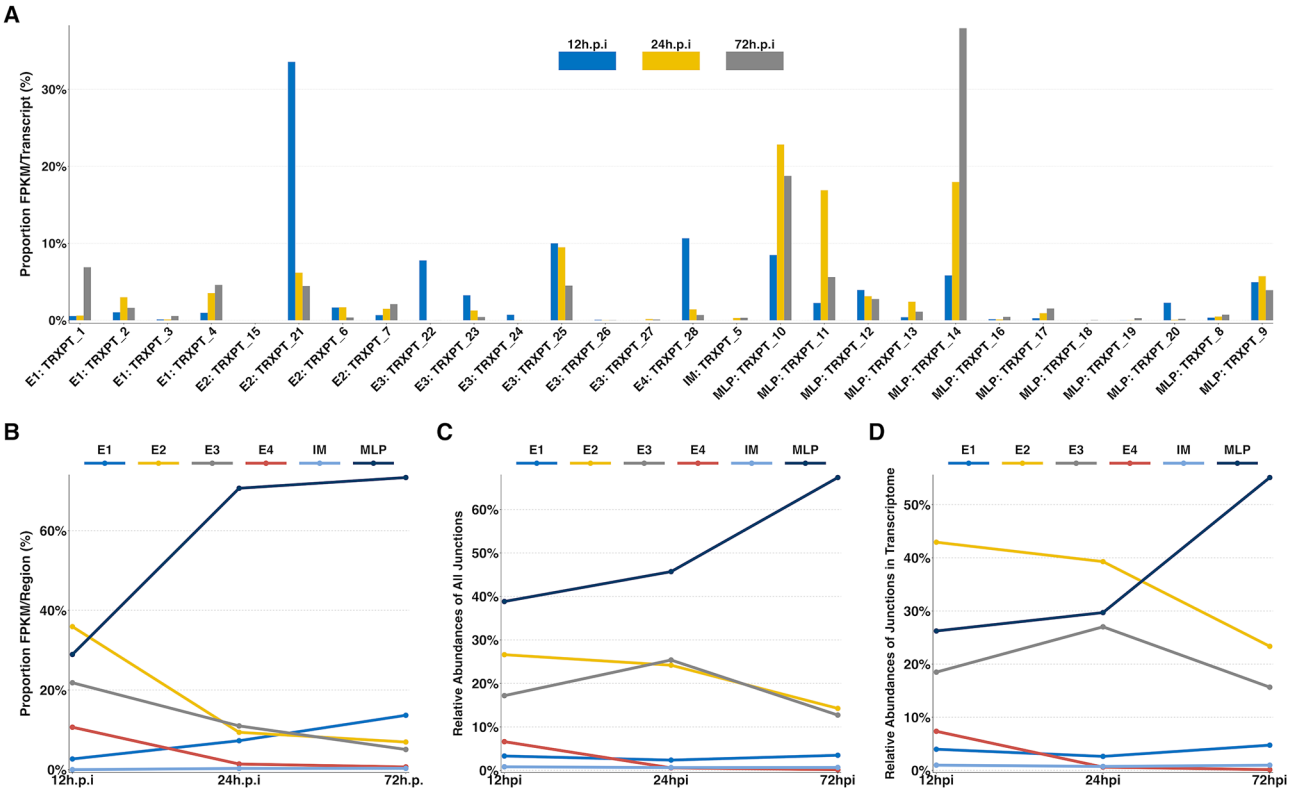


Fig. 4 Changes in splicing and expression profile of THEV over time. (**A**) Normalized (FPKM) expression levels of transcripts over time. The expression levels (FPKM) of individual transcripts as a percentage of the total expression of all transcripts at each time point are indicated. Only transcripts from our RNA-seq data are included here. (**B**) Normalized (FPKM) expression levels of transcripts by region over time. The expression levels of each region/TU as a percentage of the total expression of all transcripts at each time point are indicated. Region expression levels were calculated by summing up the FPKMs of all transcripts categorized in that region. **(C)** Relative abundances of all splice junctions grouped by region/TU over time. After assigning all 2,457 unique junctions to a TU and the total junction reads counted at each time point for each region, the total junction reads for each TU were plotted as percentages of all junction reads at each time point. Note that the junction read counts are not normalized. **(D)** Relative abundances of junctions in transcriptome grouped by region/TU over time. This is identical to **(C)**, except that only the junctions found in the full transcriptome obtained from the RNA-seq data were included

Table 2A Most abundant splice junctions at 12 h.p.i

Timepoint	Strand	Start	End	Region	Intron Length	Reads (Percentage)
12hpi	-	18,087	18,159	E2	72 bp	103 (17%)
12hpi	+	18,189	18,684	MLP	495 bp	97 (16%)
12hpi	+	7,531	7,754	MLP	223 bp	58 (9.6%)
12hpi	-	25,701	26,055	E4	354 bp	37 (6.1%)
12hpi	+	20,223	20,419	E3	196 bp	33 (5.5%)
12hpi	+	4,360	7,454	MLP	3,094 bp	32 (5.3%)
12hpi	-	18,751	20,668	E2	1,917 bp	22 (3.6%)
12hpi	+	18,350	18,717	E3	367 bp	21 (3.5%)
12hpi	+	18,768	20,162	E3	1,394 bp	21 (3.5%)
12hpi	+	7,807	13,610	MLP	5,803 bp	18 (3%)
12hpi	+	18,350	20,162	E3	1,812 bp	18 (3%)
12hpi	-	18,189	18,684	E2	495 bp	14 (2.3%)
12hpi	-	18,751	21,682	E2	2,931 bp	10 (1.7%)
12hpi	+	304	1,616	E1	1,312 bp	9 (1.5%)
12hpi	+	1,655	1,964	E1	309 bp	9 (1.5%)
12hpi	-	18,087	18,163	E2	76 bp	8 (1.3%)
12hpi	+	7,807	12,238	MLP	4,431 bp	7 (1.2%)
12hpi	+	7,807	22,492	MLP	14,685 bp	6 (1%)

Table 2B Most abundant splice junctions at 24 h.p.i

Timepoint	Strand	Start	End	Region	Intron Length	Reads (Percentage)
24hpi	-	18,087	18,159	E2	72 bp	18,825 (16.4%)
24hpi	+	18,189	18,684	MLP	495 bp	17,670 (15.4%)
24hpi	+	7,531	7,754	MLP	223 bp	12,319 (10.7%)
24hpi	+	20,223	20,419	E3	196 bp	10,583 (9.2%)
24hpi	+	4,360	7,454	MLP	3,094 bp	7,128 (6.2%)
24hpi	+	18,350	20,162	E3	1,812 bp	6,619 (5.8%)
24hpi	+	18,768	20,162	E3	1,394 bp	5,207 (4.5%)
24hpi	+	18,350	18,717	E3	367 bp	3,930 (3.4%)
24hpi	-	18,751	20,668	E2	1,917 bp	3,870 (3.4%)
24hpi	+	7,807	13,610	MLP	5,803 bp	2,553 (2.2%)
24hpi	+	7,807	12,238	MLP	4,431 bp	2,446 (2.1%)
24hpi	+	7,807	22,492	MLP	14,685 bp	1,642 (1.4%)
24hpi	+	1,655	1,964	E1	309 bp	1,395 (1.2%)
24hpi	+	7,807	18,717	MLP	10,910 bp	1,391 (1.2%)
24hpi	-	18,189	18,684	E2	495 bp	1,124 (1%)
24hpi	-	18,751	21,128	E2	2,377 bp	1,124 (1%)
24hpi	+	20,223	20,894	E3	671 bp	1,208 (1%)

Table 2C Most abundant splice junctions at 72 h.p.i

Timepoint	Strand	Start	End	Region	Intron Length	Reads (Percentage)
72hpi	+	7,531	7,754	MLP	223 bp	322,677 (15.1%)
72hpi	+	4,360	7,454	MLP	3,094 bp	179,607 (8.4%)
72hpi	-	18,087	18,159	E2	72 bp	161,336 (7.6%)
72hpi	+	18,189	18,684	MLP	495 bp	146,425 (6.9%)
72hpi	+	20,223	20,419	E3	196 bp	93,238 (4.4%)
72hpi	+	7,807	13,610	MLP	5,803 bp	81,420 (3.8%)
72hpi	+	7,807	12,238	MLP	4,431 bp	77,616 (3.6%)
72hpi	+	18,768	20,162	E3	1,394 bp	45,062 (2.1%)
72hpi	+	1,655	1,964	E1	309 bp	38,491 (1.8%)
72hpi	+	18,350	20,162	E3	1,812 bp	38,841 (1.8%)
72hpi	+	18,350	18,717	E3	367 bp	35,490 (1.7%)
72hpi	+	304	1,616	E1	1,312 bp	25,041 (1.2%)
72hpi	-	18,751	20,668	E2	1,917 bp	26,338 (1.2%)
72hpi	+	7,807	12,904	MLP	5,097 bp	21,946 (1%)
72hpi	+	7,807	22,492	MLP	14,685 bp	21,891 (1%)

encodes a two-exon 66.4 kDa, 597 residue novel protein (ORF10), spanning almost the entire predicted ORF1 and Hyd. The intron of TRXPT_2 excludes the C-terminus of ORF1 (including its stop codon) from ORF10 but the SC of ORF10 is 102 bp upstream and in-frame with the predicted SC of ORF1. TRXPT_3, similar to TRXPT_1 but lacking the second exon, encodes a 13.1 kDa, 115 residue protein (ORF4), previously predicted [27] but excluded in later annotations [1, 12]. Our data suggest it is genuinely expressed. Lastly, TRXPT_4 encodes a distinct novel 15.9 kDa, 143 residue protein (ORF11).

The splice junctions of all transcripts in this region, except for TRXPT_4, were validated by cloning of viral cDNA and Sanger sequencing (see Supplementary PCR methods). During TRXPT_2 validation, ORF1 was found on the agarose gel (an unspliced band size) and Sanger sequencing results showed it to be a transcribed mRNA (Supplementary PCR methods). This was corroborated by our 3' RACE experiment, which showed a transcript (TRXPT_2B) spanning the entire ORF1 and Hyd ORFs without splicing, with a poly-A tail immediately after the E1 TTS. The 5'-most coding sequence (CDS) of this transcript (TRXPT_2B) encodes ORF1. However, TRXPT_2B has an upstream and in-frame SC to the predicted SC of ORF1, suggesting that the predicted ORF1 CDS is truncated – the expressed ORF1 (eORF1) shares the same SC as ORF10, but has a unique stop codon (STC). See Supplementary Table S3 for all transcripts and their encoded proteins.

Early region 2 (E2) and intermediate region (IM) transcripts

The E2 TU expressed on the anti-sense strand, is subdivided into E2A and E2B and encodes three classical AdV proteins – pTP and Ad-pol (E2B proteins), and DBP

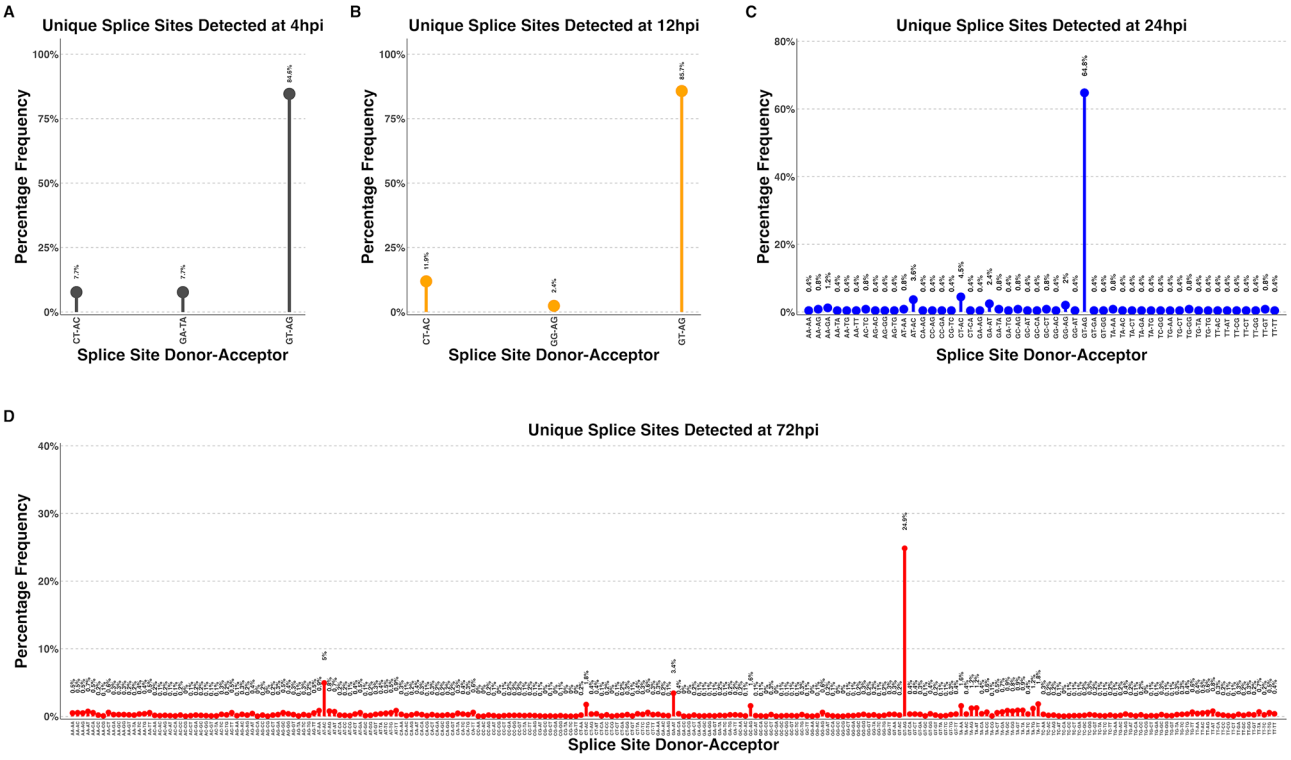


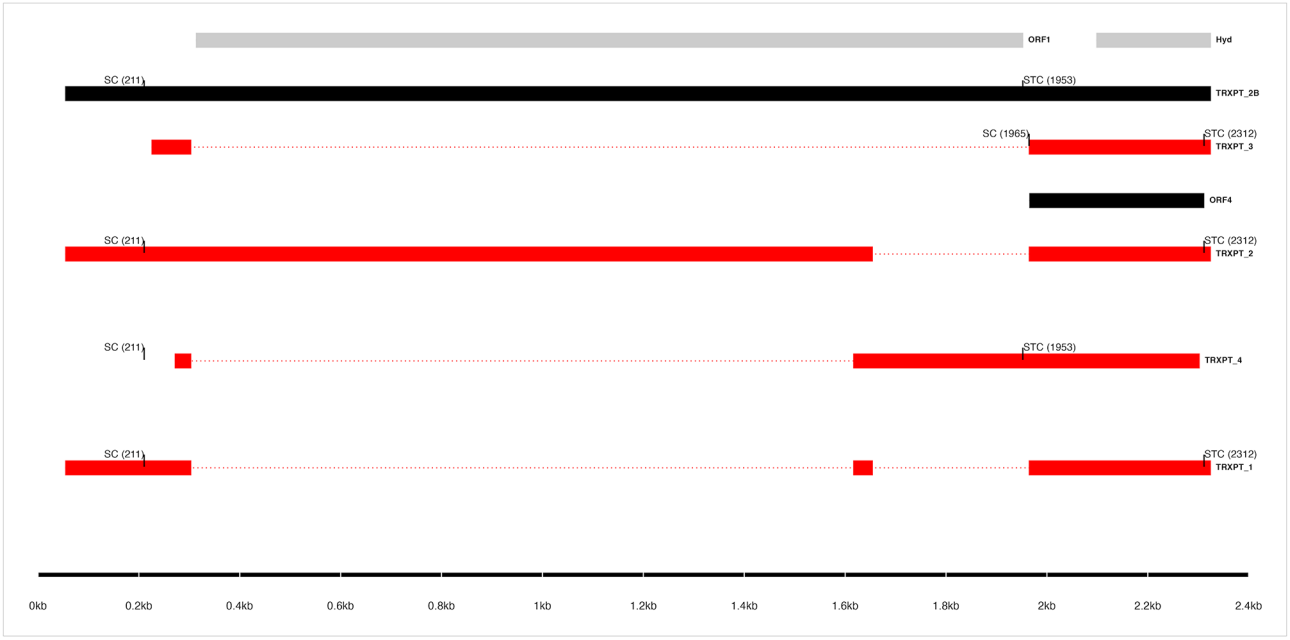
Fig. 5 Changes in splice donor-acceptor nucleotides over time. The splice donor-acceptor nucleotides of THEV just like other AdVs is mostly the canonical GU-AG. At early time points (4 h.p.i and 12 h.p.i **(A)** and **(B)**, respectively) the junction nucleotides used appear to be well scrutinized or restricted, utilizing mostly the canonical splice nucleotides. However, as the infection progresses to the late stages (24 h.p.i and 72 h.p.i **(C)** and **(D)**, respectively), the selectivity of specific splice acceptor-donor pairs seems to degenerate significantly, such that all combinations of nucleotides are utilized

(E2A protein) – essential for genome replication [17, 18]. Unlike MAdV where two promoters are known [17], we discovered only a single TSS (E2 TSS; 18,751 bp) for both E2A and E2B transcripts in THEV. However, E2A and E2B transcripts have distinct TTSs, with E2B transcripts sharing the TTS of the IVa2 transcript of the IM region similar to MAdVs [17, 18] (Fig. 7).

The E2A ORF, DBP, is one of three THEV ORFs predicted to be spliced from two exons. The corresponding transcript (TRXPT_21) in our data matches the predicted splice junction but includes an additional non-coding exon at the 5'-end (E2-5'UTR). Thus, TRXPT_21 is a three-exon transcript encoding DBP (380 residues, 43.3 kDa) precisely. TRXPT_21 was also corroborated in a 3' RACE experiment. Additionally, from the 3' RACE data, we found a splice variant of TRXPT_21 which retains the second intron, leading to a 2-exon transcript (TRXPT_21B). Although longer, TRXPT_21B encodes a truncated isoform of DBP (tDBP; a 346 residue, 39.3 kDa product) using a downstream in-frame SC but the same STC as DBP. Both TRXPT_21 and TRXPT_21B share a common TTS, seen in our 3' RACE data, located 39 bp downstream of the CDS in an adenine/thymine (A/T)-rich sequence followed by the poly-A tail sequence, suggesting this position (16,934 bp) as the true E2A TTS. There are two canonical PASs (AAUAAA; 16,964 and 16950 bp) immediately after the CDS any of which can serve as the PAS without affecting the encoded proteins (Supplementary Table S2).

The E2B region transcripts also start with the E2-5'UTR but extend downstream to reach the TTS at 2334 bp in the IM region, which is in the immediate context of a canonical PAS (position 2333 bp) where polyadenylation likely occurs. The E2B transcripts, TRXPT_6 and TRXPT_7, are almost identical except for an extra splice junction at the 3'-end of TRXPT_6 (Fig. 7). TRXPT_7 has the CP for both classical proteins (pTP and Ad-pol) encoded in this region, with the pTP ORF predicted to be spliced from two exons. The predicted splice junction of pTP is corroborated by our data but the full transcript is markedly longer than the predicted ORF, although the encoded product (pTP) remains unchanged. Ad-pol (polymerase) is encoded downstream of pTP with secondary SC (secSC) usage. The CP of TRXPT_6 slightly differs from TRXPT_7 because a new STC resulting from the extra splice site forms a minimal truncation of the Ad-pol encoded from its secSC.

While both TRXPT_6 and TRXPT_7 have the CP for Ad-pol with secSC usage, in all AdVs studied, the two proteins (pTP and Ad-pol) are encoded by separate



	Splice Junction				Junction Reads				
Transcript ID	Start	End	Intron Length	Strand	4h.p.i	12h.p.i	24h.p.i	72h.p.i	Junction Status
TRXPT_1, TRXPT_4	304	1616	1313bp	+	0	9	1019	25041	Validated*
TRXPT_3	304	1964	1661bp	+	0	2	168	1588	Validated
TRXPT_2, TRXPT_1	1655	1964	310bp	+	0	9	1395	38491	Validated

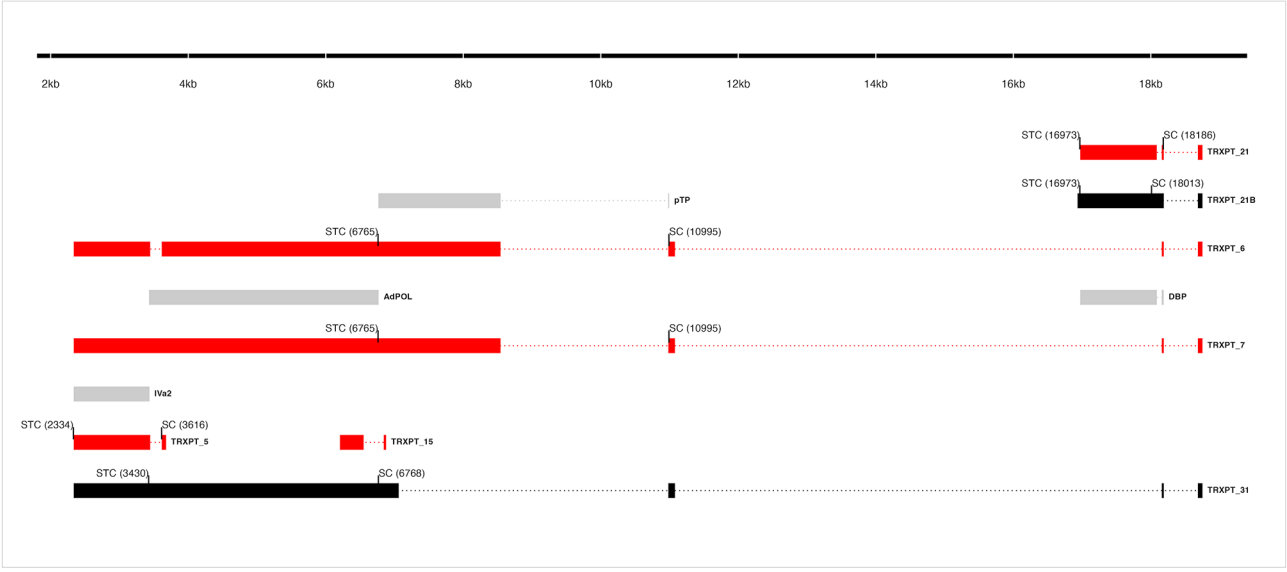
*Not validated for TRXPT_4

Fig. 6 The splice map of the E1 TU. Exons are depicted as boxes connected by introns (dotted lines). Transcripts from RNA-seq data are colored red, predicted ORFs are colored grey, and transcripts or ORFs discovered by other means are colored black. Each transcript or ORF is labelled with its name to the right. The SC and STC of the 5′-most CDS of each transcript is indicated with the nucleotide position in brackets. The region of the virus is depicted at the bottom as a black line with labels of the nucleotide positions for reference. The table shows sequence reads covering the splice junctions with information about their validation status using cloning and Sanger sequencing

mRNAs with identical first three 5’ exons and TTS, but different splice junctions to the terminal coding exons. Hence, we checked for a longer splice junction between the third and fourth (terminal) exons of TRXPT_7 with our junction validation method (targeted PCR, cloning, and Sanger sequencing). We discovered a unique splice junction (10,981–7062 bp) not present in our RNA-seq data. If initiated from the E2 TSS and terminated at the E2B TTS, this transcript (TRXPT_31) would encode Ad-pol in its 5’-most CDS (Fig. 7).

Our RNA-seq data also showed a novel short transcript (TRXPT_15) entirely nested within the terminal exon of TRXPT_7 but with a unique splice site. This transcript is an incomplete construction from the mapped reads as it contains a truncated CDS. However, we validated this splice junction to be genuine (Supplementary PCR methods).

The IM region is a single-transcript TU, encoding a single classical protein, IVa2. The promoter expressing this single transcript (TRXPT_5) is embedded in the E2B region and shares a TTS with E2B transcripts [17, 18]. TRXPT_5 is a two-exon transcript with a non-coding first exon, except the last 2 nucleotides, which connect with the first nucleotide of the second exon to form the 5’-most SC. This new SC is four codons upstream and in-frame of the predicted IVa2 SC. Beside the four



	Splice Junction				Junction Reads				
Transcript ID	Start	End	Intron Length	Strand	4h.p.i	12h.p.i	24h.p.i	72h.p.i	Junction Status
TRXPT_5, TRXPT_7	3447	3615	169bp	-	1	5	720	13422	Validated
TRXPT_6, TRXPT_7	11079	18159	7081bp	-	0	2	0	0	Validated
TRXPT_21	18087	18159	73bp	-	9	103	0	0	Validated
TRXPT_21, TRXPT_6, TRXPT_7	18189	18684	496bp	-	0	111	18794	156037	Validated
TRXPT_6, TRXPT_7	8543	10981	2439bp	-	0	0	298	850	Validated
TRXPT_15	6551	6843	293bp	-	0	0	0	6	Validated

Fig. 7 The splice map of the E2 and IM TUs. Exons are depicted as boxes connected by introns (dotted lines). Red transcripts are generated from RNA-seq data and predicted ORFs are colored grey. TRXPT_21B discovered by 3'RACE is colored black. Each transcript or ORF is labelled with its name to the right. The SC and STC of the 5'-most CDS of each transcript are indicated with the nucleotide position in brackets. The region of the virus is depicted at the bottom as a black line with labels of the nucleotide positions for reference. The table shows sequence reads covering the splice junctions with information about their validation status using cloning and Sanger sequencing

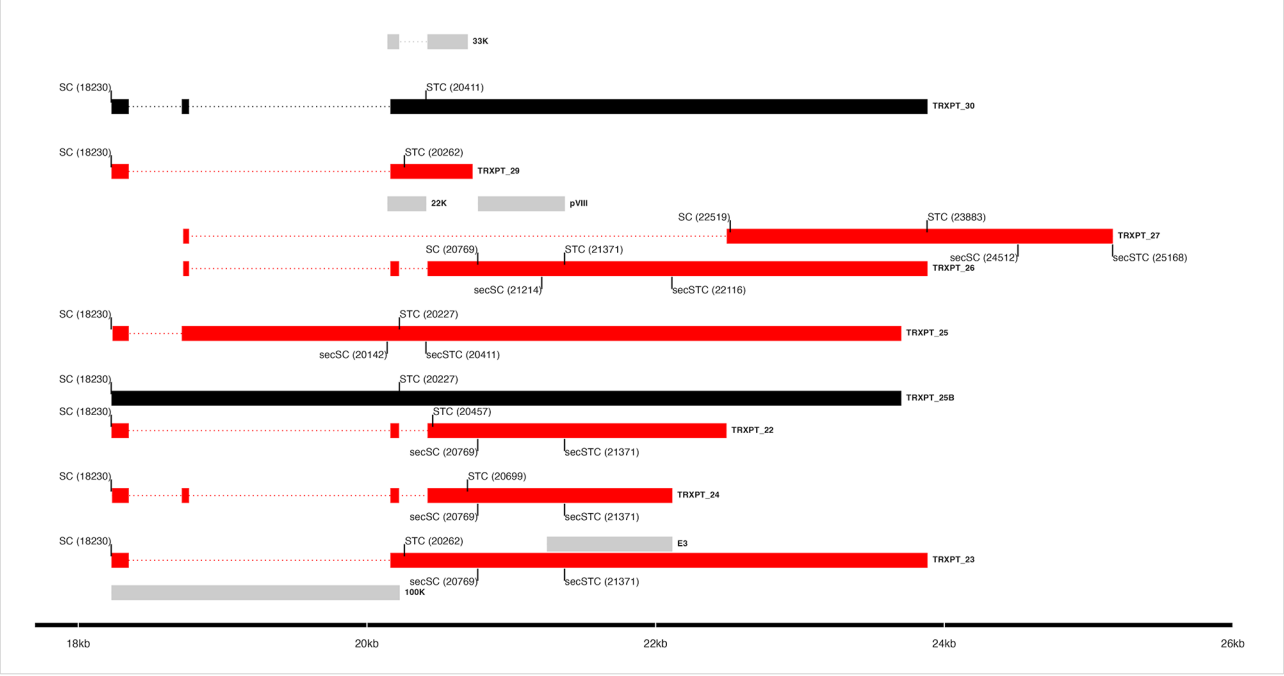
additional N-terminus residues, the protein sequence is unchanged.

Early region 3 (E3) transcripts

The E3 region, nested within the MLTU, encodes proteins that modulate and evade host immune defenses. In MAdVs, this region contains seven ORFs expressed from multiple transcripts sharing the same TSS (from the E3 promoter) but having different TTSs [10, 17, 18]. However, some E3 transcripts use the TSS of the MLP. Due to sharing the same TSS, in MAdVs, secSC usage is heavily relied on for gene expression in this region as utilizing only the first SC cannot produce the downstream proteins in this TU. The 12.5 K ORF and transcripts using the MLP TSS are exceptions [17].

In THEV, only one ORF (E3) was predicted in this region. However, as the E3 TU is nested in the MLTU, transcripts from the L4 promoter (100 K, 22 K, 33 K, and pVIII) overlap the E3 region transcripts entirely and share similar TSS and TTS locations (Fig. 8). Therefore, we have categorized these two groups together as E3 transcripts.

We identified seven novel transcripts (TRXPT_22, TRXPT_23, TRXPT_24, TRXPT_25, TRXPT_26,



Transcript ID	Splice Junction			Strand	Junction Reads				Junction Status
	Start	End	Intron Length		4h.p.i	12h.p.i	24h.p.i	72h.p.i	
TRXPT_25, TRXPT_24, TRXPT_10	18350	18717	368bp	+	4	21	3930	35490	Validated
TRXPT_23, TRXPT_22, TRXPT_11	18350	20162	1813bp	+	3	18	6619	38841	Validated
TRXPT_26, TRXPT_24, TRXPT_13, TRXPT_9, TRXPT_10	18768	20162	1395bp	+	2	21	5207	45062	Validated
TRXPT_26, TRXPT_22, TRXPT_24, TRXPT_13, TRXPT_11, TRXPT_9, TRXPT_10	20223	20419	197bp	+	3	33	10583	93238	Validated
TRXPT_27	18768	22492	3725bp	+	0	0	101	1950	Validated

Fig. 8 The splice map of the E3 TU. Exons are depicted as boxes connected by introns (dotted lines). Red transcripts are generated from RNA-seq data and predicted ORFs are colored grey. Transcripts discovered by other means are colored black. Each transcript or ORF is labelled with its name to the right. The SC and STC of the 5′-most CDS of each transcript are indicated with the nucleotide position in brackets. Similarly, the secondary SC (secSC) and secondary STC (secSTC) are shown. The region of the virus is depicted at the bottom as a black line with labels of the nucleotide positions for reference. The table shows sequence reads covering the splice junctions with information about their validation status using cloning and Sanger sequencing

TRXPT_27, TRXPT_29) from our RNA-seq data, all originating from two distinct TSSs. We consider the first TSS (position 18,230 bp) as corresponding to the L4 promoter (L4P) and the other at 18,727 bp as corresponding to the E3 promoter (E3P). We also identified the canonical or other known PASs [28] near the TTS of the transcripts (see Supplementary Table S2). These E3 transcripts collectively have the CP for several predicted THEV ORFs: 100 K, 22 K, 33 K, pVIII, and E3, as well as Fiber (IV) and ORF7 of the MLTU (see Supplementary Table S3). However, some of these CDSs differ from the predictions due to either unknown exons or the presence of an in-frame upstream SC. For instance, we discovered that 33 K, predicted to be spliced from two exons, is actually a significantly longer four-exon ORF (e33K; 19.8 kDa, 171 residues) encoded on TRXPT_24. Its first two exons were unknown but the last two match the predicted exons and the CDS is in-frame, albeit the first 20 bp of the predicted 33 K (including the SC) is spliced out as part of the second intron of TRXPT_24. TRXPT_24 also has the CP for pVIII and E3 if we consider downstream SC usage. However, we found an upstream in-frame SC for the predicted E3; thus, this longer version of E3 (eE3) is likely the genuinely expressed ORF. TRXPT_29,

the shortest transcript in this TU, encodes a novel 73 residue protein (8.3KI) across its two exons using the SC of e33K with a unique STC. TRXPT_23, spliced identically as TRXPT_29, also encodes 8.3KI from its first SC. Similarly, TRXPT_22 encodes a 73 residue novel protein (8.3KII) from its first SC that shares over 80% similarity with 8.3KI, but they differ at the C-terminus. Considering downstream SC usage, both TRXPT_22 and TRXPT_23 can encode pVIII and eE3 in that order, but TRXPT_23 being longer, also has the CP for the Fiber ORF.

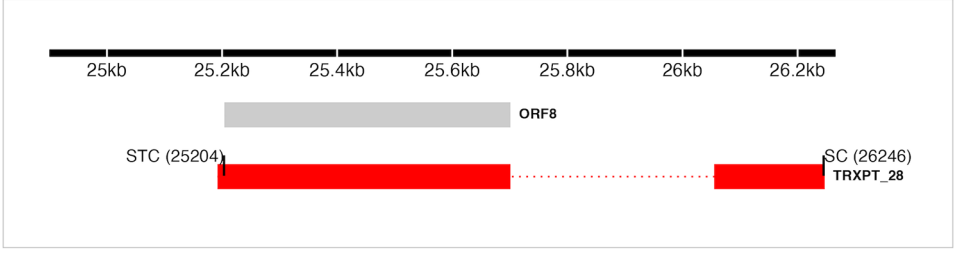
As the splice junctions of TRXPT_22, TRXPT_23, TRXPT_24, and TRXPT_29 share the same genomic space, their validation was done with a single primer pair, and they were differentiated from each other by cloning the cDNA and Sanger sequencing (Supplementary PCR methods). In addition to corroborating the splice junctions for the aforementioned transcripts, the Sanger sequencing results also showed a distinct splice variant undetected in our RNA-seq transcriptome. This was a three-exon transcript (TRXPT_30) with identical first and last exons as TRXPT_23, which also contained the second exon of TRXPT_24 (Fig. 8). TRXPT_30 encodes a novel 140 residue, 15.7 kDa protein (e22K), spanning all three exons. Interestingly, the last 81 C-terminus residues of e22K are identical to 22 K (89 residues), a single-exon ORF predicted to use the same SC as 33 K. Just as seen for 33 K, the first 20 bp of 22 K is intronic, excluding the first 7 residues of 22 K from e22K. We consider e22K as a long variant of the predicted 22 K ORF. Assuming TRXPT_30 shares the same TSS and TTS as TRXPT_23, it would also have the downstream CP of TRXPT_23.

TRXPT_25, the largest transcript in the TU, is a two-exon transcript, encoding a novel protein (t100K; 543 residues), which is a shorter isoform of the predicted 100 K ORF. secSC usage on this transcript yields the predicted 22 K ORF. It also has the CP for pVIII and eE3 downstream. Furthermore, during the validation of the TRXPT_25 splice junction using primers that span its junction (18,350−18,717 bp), we noticed a DNA band corresponding to the full unspliced sequence (Supplementary PCR methods). As TRXPT_25 only falls short of encoding the complete predicted 100 K protein due to its splice junction, this band (which we cloned and validated by Sanger sequencing) suggests that the predicted 100 K is indeed expressed. We assume that this transcript (TRXPT_25B) shares the same TSS and TTS as TRXPT_25.

Lastly, TRXPT_26 and TRXPT_27, both originate from the E3P but have distinct TTSs. TRXPT_26 encodes pVIII as the 5'-most ORF and has the CP for eE3 and Fiber in that order. TRXPT_27, a two-exon transcript, encodes Fiber as the 5'-most ORF, and ORF7 downstream with secSC usage. TRXPT_13 is an L4P transcript that uses the MLP TSS; it is discussed under the MLTU transcripts.

Early region 4 (E4) transcripts

This TU is found at the 3'-end of the genome and expressed on the anti-sense strand. Based on nucleotide position, ORF7 and ORF8 were predicted in this region [1]; however, as ORF7 is neither on the anti-sense strand nor transcribed from a promoter in the E4 region, we only classify ORF8 in this TU. This is corroborated by our RNA-seq data, showing only one transcript in this region on the anti-sense strand (Fig. 9). The transcript (TRXPT_28) spans 25,192–26,247 bp and is spliced at 25,701–26,055 bp, forming a two-exon transcript. The second exon fully matches the predicted ORF8 with 12



	Splice Junction				Junction Reads				
Transcript ID	Start	End	Intron Length	Strand	4h.p.i	12h.p.i	24h.p.i	72h.p.i	Junction Status
TRXPT_28	25701	26055	355bp	-	1	37	602	1,894	Validated

Fig. 9 The splice map of the E4 TU. Exons are depicted as boxes connected by introns (dotted lines). The transcript from RNA-seq data is colored red and the predicted ORF, grey. The transcript and ORF are labelled with their names to the right. The SC and STC of the 5'-most CDS are indicated with the nucleotide position in brackets. The region of the virus is depicted at the bottom as a black line with labels of the nucleotide positions for reference. The table shows sequence reads covering the splice junction with its validation status using cloning and Sanger sequencing

extra base pairs at the 3'-end. However, we identified a SC 192 bp upstream of the predicted SC in the first exon from which an in-frame protein is encoded. We consider this longer isoform (eORF8–26.4 kDa, 229 residues) as the genuinely expressed ORF. We also identified a canonical PAS 11 bp upstream of TTS (Supplementary Table S2).

MLTU or MLP region transcripts

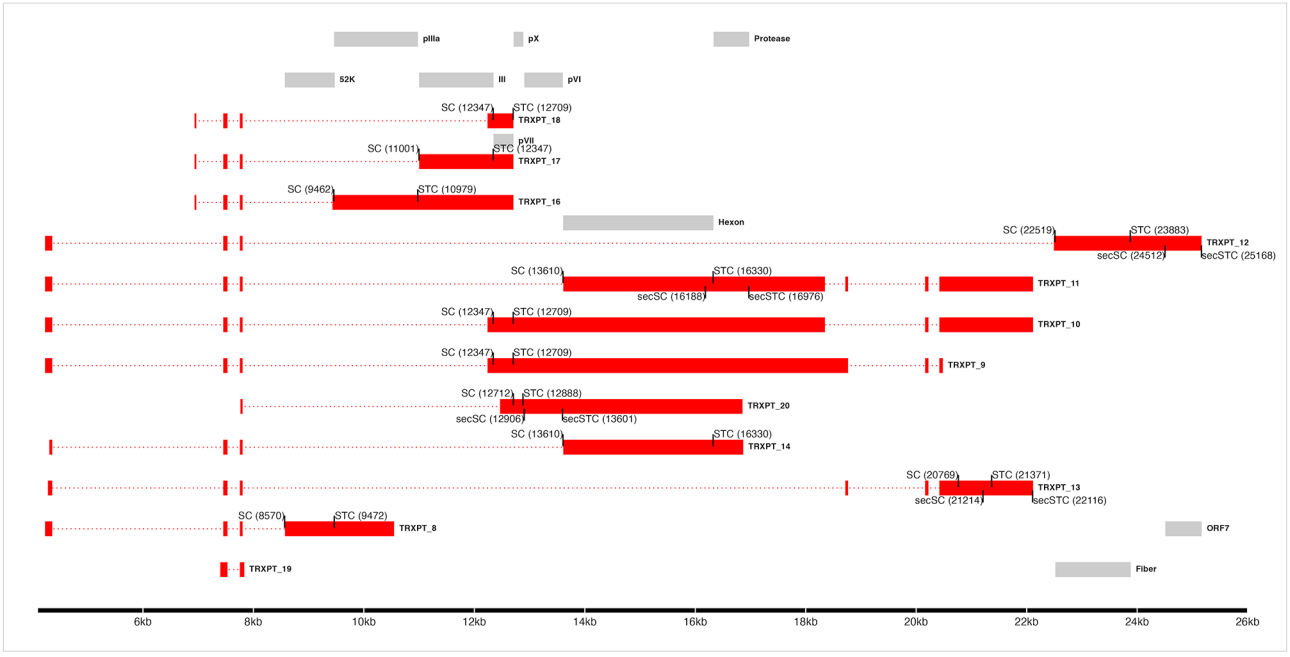
The MLTU transcripts, dominant in the late phase of the AdV infectious cycle, are produced by alternative polyadenylation and splicing of a primary transcript and grouped into five transcript classes (L1-L5). About 13 out of the 23 predicted ORFs in THEV fall within this TU, some of which we have categorized under the E3 TU instead. Our RNA-seq data revealed 12 transcripts (TRXPT_8, TRXPT_9, TRXPT_10, TRXPT_11, TRXPT_12, TRXPT_13, TRXPT_14, TRXPT_16, TRXPT_17, TRXPT_18, TRXPT_19, TRXPT_20) in this TU, most of which have the 5' TPL sequence as in all AdVs. However, three transcripts (TRXPT_16, TRXPT_17, TRXPT_18) use a different leader sequence (sTPL), where a different first exon is used instead of the first TPL exon, and TRXPT_20 uses only the third TPL exon (TPL3); see Fig. 10.

We identified five TTSs (10,549, 12,709, 16,870, 22,116, 25,168 bp) in this TU, which we consider as corresponding to the five late mRNA classes (L1-L5), respectively. L1 mRNAs include TRXPT_8, encoding the 52K ORF as predicted. L2 mRNAs include TRXPT_16, TRXPT_17, and TRXPT_18, all containing the sTPL with their respective coding exons. They encode pIIIa, III (penton), and pVII, respectively. The L3 mRNAs, TRXPT_14 and TRXPT_20, both encode the hexon (II) ORF but hexon is the only ORF encoded on TRXPT_14, whereas TRXPT_20 encodes pX (pre-Mu), pVI, and hexon in that order. L4 mRNAs, TRXPT_9, TRXPT_10, TRXPT_11, and TRXPT_13 are the largest transcripts in the transcriptome and encode several similar late proteins. TRXPT_9 and TRXPT_10 are very similar but not identical. The last exon of TRXPT_9 seems to be truncated and likely shares the same TTS as TRXPT_10. They both encode pVII as the 5'-most ORF and also have the CP for pX, pVI, hexon, a longer variant of protease (eProt) from an upstream in-frame SC, and ORF12 (a novel 120 residue protein). Additionally, they have the CP for pVIII and eE3. TRXPT_11 encodes hexon as its 5'-most ORF and also has the CP for eProt, ORF12, e33K, pVIII and eE3. Typically, MLTU transcripts splice the TPL onto a splice site just upstream of the ORF to be expressed [17]. While this holds true for most MLTU ORFs, several late ORFs (pVI, protease, and ORF7) do not have such close proximity splicing but are contained in larger transcripts such as these L4 mRNAs, strongly suggesting the use of non-standard ribosomal initiation mechanisms such as secSC usage or ribosome shunting described in other AdVs for their translation [17, 29]. TRXPT_13, an E3 ORF utilizing the MLP TSS, encodes the classical L4P genes, pVIII and eE3. Lastly, the L5 class transcript, TRXPT_12, encodes Fiber as its 5'-most ORF but also has the CP for ORF7. Interestingly, the CP of TRXPT_12 and TRXPT_27 of the E3 TU are identical but are initiated from different TSSs.

Discussion

While the advent of next-generation sequencing has rendered easier the study of large and complex eukaryotic transcriptomes, studying the smaller, compact viral transcriptomes is counterintuitively more challenging, as the transcripts typically have significant overlaps due to genome economization. AdV transcriptomes escalate the difficulty due to the wide array of mRNAs produced via very complex alternative splicing and polyadenylation, all initiated from relatively few promoters. Standard RNA-seq analysis programs, not primarily designed for such compact, gene-dense, and complex transcriptomes, further compound this challenge. Furthermore, in our case, no prior transcriptomic studies for THEV exist; hence, assembling the transcripts without any prior experimentally-derived annotation of THEV splicing using only short illumina reads proved difficult. Lastly, we had initially planned to sequence RNA from another time point (8 hpi); however, all the RNA samples from 8 hpi and one replicate sample from 12 hpi got too degraded during the library preparation steps to be yield any useful data. We believe that these would have contributed to better insights into the temporal expression levels and splicing. As these data are the result of one study, replication of our results by others would be useful in ensuring the best characterization of the viral genome. Our approach combines standard RNA-seq analysis programs with custom analyses and experimentally validating all splice junctions with independent methods. The transcript map for THEV produced from our analysis is strikingly similar to that of the MAdVs.

Our work provides the first insights into THEV splicing, revealing 34 transcripts grouped into five TUs. The general temporal gene expression regulation observed in MAdVs, with early regions peaking at earlier time points followed by MLTU predominance at later time points, seems to also apply to THEV. An unexpected observation is that the pileup of mapped reads to THEV seems consistently skewed over similar regions of the genome at all time points. Given the temporal regulation of AdVs gene expression, we anticipated distinct differences in read pileups over the genome at different time points, indicating the different stages of infection. This may be due to an unsynchronized infection, leading to transcripts



Transcript ID	Splice Junction			Strand	Junction Reads				Junction Status
	Start	End	Intron Length		4h.p.i	12h.p.i	24h.p.i	72h.p.i	
TRXPT_13, TRXPT_8, TRXPT_14, TRXPT_10, TRXPT_12, TRXPT_11, TRXPT_9	4360	7454	3095bp	+	1	32	7128	179607	Validated
TRXPT_19, TRXPT_13, TRXPT_8, TRXPT_14, TRXPT_10, TRXPT_16, TRXPT_12, TRXPT_17, TRXPT_18, TRXPT_11, TRXPT_9	7531	7754	224bp	+	4	58	12319	322677	Validated
TRXPT_8	7807	8570	764bp	+	0	5	707	0	Validated
TRXPT_16	7807	11001	3195bp	+	0	1	226	10254	Validated
TRXPT_18, TRXPT_11, TRXPT_9	7807	12238	4432bp	+	0	7	2446	0	Validated
TRXPT_20	7807	12466	4660bp	+	0	3	437	0	Validated
TRXPT_14, TRXPT_10	7807	13610	5804bp	+	0	18	2553	0	Validated
TRXPT_13	7807	18717	10911bp	+	0	2	1391	0	Validated
TRXPT_12	7807	22492	14686bp	+	1	6	1642	21891	Validated
TRXPT_25, TRXPT_24, TRXPT_10	18350	18717	368bp	+	4	21	3930	35490	Validated
TRXPT_23, TRXPT_22, TRXPT_11	18350	20162	1813bp	+	3	18	6619	38841	Validated
TRXPT_26, TRXPT_24, TRXPT_13, TRXPT_9, TRXPT_10	18768	20162	1395bp	+	2	21	5207	45062	Validated
TRXPT_26, TRXPT_22, TRXPT_24, TRXPT_13, TRXPT_11, TRXPT_9, TRXPT_10	20223	20419	197bp	+	3	33	10583	93238	Validated
TRXPT_16, TRXPT_17, TRXPT_18	6969	7454	486bp	+	0	0	143	13482	Validated
TRXPT_17	7807	9430	1624bp	+	0	0	67	0	Validated

Fig. 10 The splice map of the MLTU. Exons are depicted as boxes connected by introns (dotted lines). The transcripts from our RNA-seq data are colored red and the predicted ORFs, grey. The transcripts and ORFs are labelled with their names to the right. The SC and STC of the 5′-most CDS of each transcript is indicated with the nucleotide position in brackets. Similarly, the secondary SC (secSC) and secondary STC (secSTC) are shown. The region of the virus is depicted at the bottom as a black line with labels of the nucleotide positions for reference. The table shows sequence reads covering the splice junctions with information about their validation status using cloning and Sanger sequencing

overlapping the time points. Further research is required to determine the precise temporal regulation of THEV.

Short read deep sequencing effectively reconstructs full AdV mRNA structures, particularly mapping splice sites [18]. However, the substantial overlapping nature of AdV mRNAs and fragmentation during library preparation make it challenging to map the exact TSS, TTS, and PASs of assembled transcripts. As AdVs make heavy use of alternative polyadenylation, short read RNA-seq is ill-equipped to discriminate mRNA variants of the same gene produced via alternative polyadenylation. Thus, shorter variants of alternatively polyadenylated mRNAs may potentially be incorporated into the longer variants during transcript assembly, significantly diminishing the diversity of mRNA in the transcriptome. Also, independent transcripts with significant overlaps may

be assembled as a single, longer mRNA, since the short reads alone do not provide enough context for the transcript assembler (StringTie) to distinguish them. Such fusions may affect the transcript expression level estimations by inflating or deflating the expression levels of the transcripts involved, affecting the proper understanding of the temporal gene expression regulation and also the diversity of the transcriptome. Transcripts that have reads mistakenly fused with them would have inflated expression levels while those whose reads are counted elsewhere would show false lower expression levels. In our case, since we used other independent methods to validate the splice junctions, we believe these drawbacks to be minimized. Our results show transcripts in the same TU initiating or terminating in similar areas, but not at the exact same position. We consider the most upstream TSS or most downstream TTS for for the transcripts involved but we present them unchanged in all the figures shown (see Supplementary Table S2). Also, comparing our results to the better-studied MAdV transcriptomes, we believe some long transcripts in the MLTU (TRXPT_9, TRXPT_10, and TRXPT_11) are likely due to fusing some E3 transcripts to the terminal exons of the MLTU transcripts by StringTie, making them significantly longer. These mRNAs have unusually many exons and their last few exons are identical to some E3 mRNAs. Future studies using long read sequencing technologies will provide more precise mapping of the TSS, TTS, and PASs and clarify the structures of the long MLTU transcripts.

While most predicted ORFs are encoded by the spliced transcripts, we found some that seem to be truncated predictions, as either an upstream in-frame SC or unknown upstream exons were found. Other ORFs were identified that were either shorter or longer isoforms of some predicted ORF. We also found several novel unpredicted ORFs (Supplementary Table S3). On this basis, we anticipate that further studies will likely reveal more unpredicted novel ORFs or new variants of predicted ORFs. Furthermore, it is not unreasonable to presume that several splice variants will likely be found as evidenced firstly by finding unique transcripts using 3' RACE and during our splice junction validation steps. And secondly, recent studies [17, 18, 22] are still discovering novel mRNA variants for even the best studied MAdVs decades later. These new potential proteins and isoforms significantly extend the repertoire of THEV gene products than predicted, adding a hefty number of proteins of undefined function to the previously predicted five (See Fig. 1). These new potential proteins and isoforms of unknown functions may mediate or contribute to viral replication efficiency, or the immunosuppression associated with THEV. Hence, further studies of these potential proteins is urgently needed.

Eukaryotic mRNAs are typically functionally monocistronic, with the 5'-most AUG determining the translation reading frame. However, AdV mRNAs, which span more than one ORF, are functionally polycistronic, employing non-standard mechanisms of translation initiation such as secSC usage and ribosome shunting [10, 22]. AdVs use secondary AUGs as initiation codons for most E1b proteins and some E3 proteins. In fact, recent studies show that secSC usage is found transcriptome-wide. This is thought to occur because translation initiation at the first SC is inefficient, allowing downstream SCs to be employed [17]. Ribosomal shunting or jumping mechanism is utilized for MLTU transcripts that have the TPL. This mechanism allows the ribosome to translocate to a downstream AUG, under the direction of the shunting elements in the TPL, even if a SC in a good Kozak sequence context is bypassed. Thus, predicting the protein(s) expressed from an AdV mRNA is uncertain as any one of the AUGs may be selected [10, 22]. Almost all the THEV transcripts in our data have the CP for several ORFs, some spanning as many as six ORFs. This supports the usage of these special ribosome initiation mechanisms as several predicted and novel ORFs found on mRNA in our data could not be translated using only the typical ribosome scanning mechanism. Interestingly, several distinct THEV mRNAs have identical CPs. This is also observed in human AdVs in a recent study [17]. They proposed that this may permit protein production to be fine-tuned through alteration in the balance between different mRNA groups expressing that ORF.

AdV alternative splicing undergoes a regulated temporal shift in splice site usage, previously thought to be limited to certain TUs. However, recent studies suggest that AdVs routinely produce different combinations of splice acceptor–donor pairs across all TUs [10, 17, 22, 30]. The details of this phenomenon have been best studied for the E1A and L1 units. AdVs modulate the activities of the splicing factor U2AF and the cellular SR family of splicing factors (reviewed here [30]), and encode several proteins that influence the RNA splice site used. This phenomenon appears to occur in the THEV transcriptome, as the stringency of splice acceptor-donor pairs selected decreases from the onset of the late phase (Fig. 5). Recent studies show that a virtually unlimited number of combinatorial alternative splicing events occur in an AdV lytic infection, resulting in a variety of novel transcripts [17, 22]. It is unlikely that the entire repertoire of mRNA produced via this mechanism will actually be translated. However, it has been speculated that the plasticity in alternative RNA splicing enables AdVs to fine-tune protein synthesis by providing different alternatively spliced variants encoding the same protein under changing conditions, conferring an evolutionary advantage [17, 22].

Conclusions

The THEV transcriptome bears remarkable similarity to the better-studied MAdVs. The transcriptome is organized into five TUs, with temporal regulation divided into early and late genes, and a broad repertoire of transcripts are produced via virtually unlimited alternative splicing. However, the THEV transcriptome appears less sophisticated (i.e, it encodes fewer genes) than MAdVs, primarily because the MAdV genomes are close to twice as long as that of THEV. The lack of subdivision of the E1 region into E1a and E1b is one of the most obvious examples. Also, the MAdV E4 region encodes several proteins unlike in THEV where only one transcript encoding one protein was found. The complexity of the MLTU leader sequences is another example. While the majority of the THEV MLTU transcripts begin with the TPL just like MAdVs with a small subset using a variant leader sequence (sTPL), significantly more diverse 5'-UTRs are employed for MAdV MLTU transcripts. Namely, the TPL, the so-called x, y, and z leaders, and the i-leader are 5' leaders utilized by MAdV MLTU mRNAs. The absence of these non-TPL leaders in our data could mean that the 5'-UTR diversity of THEV's MLTU mRNAs is more limited due to its smaller genome size or future studies could uncover more variety not seen in our results. We also note that although THEV genomic sequences show minimal differences between strains [12], the transcriptomes may have significant variations; hence, our results may vary from other THEV strains. Also, performing the study in vivo or with primary turkey cells may show different results. The potential new proteins identified in our work adds to the number of proteins with undefined functions in THEV; these may have roles in viral replication efficiency, or the immunosuppression associated with THEV. Hence, further studies of these proteins and other predicted proteins of unknown function should be useful in elucidating THEV-induced immunosuppression. Being the first transcriptomic characterization of THEV, this work should serve as useful resource to future THEV gene expression and transcriptomic studies, especially, mapping the mRNA splice sites. More importantly, this work characterizing the splicing of THEV mRNAs will allow researchers to accurately clone any THEV gene(s) of interest to study its potential role in inducing immunosuppression or other functions.

Methods

Cell culture and THEV infection

The Turkey B-cell line (MDTC-RP19, ATCC CRL-8135) was grown as suspension cultures in 1:1 complete Leibovitz's L-15/McCoy's 5 A medium with 10% fetal bovine serum (FBS), 20% chicken serum (ChS), 5% tryptose phosphate broth (TPB), and 1% antibiotic solution (100 U/mL Penicillin and 100μg/mL Streptomycin), at 41°C in a humidified atmosphere with 5% CO_2. Infected cells were maintained in 1:1 serum-reduced Leibovitz's L15/McCoy's 5 A media (SRLM) with 2.5% FBS, 5% ChS, 1.2% TPB, and 1% antibiotic solution. A commercially available THEV vaccine was purchased from Hygieia Biological Labs as a source of THEV-A (VAS strain). The stock virus was titrated using an in-house qPCR assay with titer expressed as genome copy number (GCN)/mL, similar to Mahshoub et al. [31] with modifications. Cells were infected in triplicate at MOI of 100 GCN/cell, incubated at 41°C for 1 h, and washed three times with phosphate buffered saline (PBS) to get rid of free virus particles. Triplicate samples were harvested at 4-, 12-, 24-, and 72-hpi for total RNA extraction. The infection was repeated but samples in triplicate were harvested at 12-, 24-, 36-, 48-, and 72-hpi for PCR validation of novel splice sites. Still one more independent infection was done at time points ranging from 12 to 168-hpi for qPCR quantification of virus titers.

RNA extraction and sequencing

Total RNA was extracted from infected cells using the Thermofisher RNAqueous™-4PCR Total RNA Isolation Kit (which includes a DNase I digestion step) per manufacturer's instructions. An agarose gel electrophoresis was performed to check RNA integrity. The RNA quantity and purity was initially assessed using nanodrop, and RNA was used only if the A260/A280 ratio was 2.0 ± 0.05 and the A260/A230 ratio was >2 and <2.2. Extracted total RNA samples were sent to LC Sciences, Houston TX for poly-A-tailed mRNA sequencing where RNA integrity was checked with Agilent Technologies 2100 Bioanalyzer High Sensitivity DNA Chip and poly(A) RNA-seq library was prepared following Illumina's TruSeq-stranded-mRNA sample preparation protocol. Paired-end sequencing to generate 150 bp reads was performed on the Illumina NovaSeq 6000 sequencing system.

Validation of novel splice junctions

All splice junctions identified in this work are novel except one predicted splice site each for pTP, DBP, and 33K, which were corroborated in our work. However, these predicted splice junctions had not been experimentally validated hitherto, and we identified additional novel exons, giving the complete picture of these transcripts. The novel splice junctions discovered in this work using the StringTie transcript assembler were validated by PCR, cloning, and Sanger Sequencing (Supplementary PCR methods). Briefly, primers spanning a range of novel exon-exon boundaries for each specific transcript in a TU were designed. Universal forward or reverse primers for each respective TU were designed and paired with primers binding specific positions in each transcript. Each forward primer contained a KpnI

restriction site and each reverse primer, an XbaI site in the primer 5' ends. After first-strand cDNA synthesis of total RNA obtained from THEV infected MDTC-RP19 cells was done using SuperScript™ IV First-Strand Synthesis System, the primers were used in a targeted PCR amplification, the products analyzed with agarose gel electrophoresis to confirm expected band sizes, cloned by traditional restriction enzyme method, and Sanger sequenced to validate these splice junctions at the sequence level. The total RNA was extracted as described above, including the DNase I digestion step. We included infected total RNA controls with no reverse transcriptase (no RT) during the cDNA synthesis step and the parent RNA were digested using RNase H after cDNA synthesis was complete to ensured that the bands obtained from the targeted PCR amplifications did not originate from the viral genomic DNA. As seen in the agarose gel images in Supplementary PCR methods, DNA bands were not found in the "no RT" controls, indicating that the DNA bands seen are of cDNA origin.

3' rapid amplification of cDNA ends (3' RACE)

A rapid amplification of sequences from the 3' ends of mRNAs (3' RACE) experiment was performed using a portion of the extracted total RNA of infected MDTC-RP19 cells used for the RNA-seq experiment as explained above. We followed the protocol described by Green *et al* [32] with modifications. Briefly, $1\mu g$ of total RNA was reverse transcribed to cDNA using SuperScript™ IV First-Strand Synthesis System following the manufacturing instructions using an adapter-primer with a 3'-end poly(T) and a 5'-end BamHI restriction site. A gene-specific sense primer with a 5'-end KpnI restriction site paired with an anti-sense adapter-primer with a 5'-end BamHI site were used to amplify target sections of the cDNA using Invitrogen's Platinum™ Taq DNA polymerase High Fidelity, following manufacturer's instructions. The PCR amplicons were restriction digested, cloned, and Sanger sequenced.

Computational analysis of RNA sequencing data: mapping and transcript characterization

Sequencing reads were analyzed following a well-established protocol described by Pertea et al. [25], using Snakemake - version 7.24.0 [33], a popular workflow management system to drive the pipeline. Briefly, sequencing reads were trimmed with the Trim-galore - version 0.6.6 [34] program to achieve an overall Mean Sequence Quality (Phred Score) of 36. Trimmed reads were mapped simultaneously to the complete genomic sequence of avirulent THEV (https://www.ncbi.nlm.nih.gov/nuccore/AY849321.1/) and *Meleagris gallopavo* (https://www.ncbi.nlm.nih.gov/genome/?term=Meleagris+gallopavo) using Hisat2 - version 2.2.1 [25] with default settings. The generated binary alignment (BAM) files from each infection time point were filtered for reads mapping to the THEV genome using Samtools - version 1.16.1 and fed into StringTie - version 2.2.1 [25] to assemble the transcripts, using a gene transfer format (GTF) annotation file derived from a gene feature format 3 (GFF3) annotation file obtained from NCBI, which contains the predicted ORFs of THEV as a guide. GFFCOMPARE - version 0.12.6 was used to merge all transcripts from all time points without redundancy and using a custom R script, adenovirus transcripts units (regions) were assigned to each transcript, generating the transcriptome of THEV. StringTie was set to expression estimation mode to calculate FPKM scores for all transcripts after which Ballgown - version 2.33.0 in R was used to perform the statistical analysis on the transcript expression levels. Samtools was also used to count the total sequencing reads for all replicates at each time point and Regtools - version 1.0.0 was used to count all junctions, the reads supporting them, and extract all other information related to the junction. See Supplementary Computational Analysis for the details of transcript expression level estimations and splice junction read counts.

Abbreviations

AdV	Adenovirus
MAdV	Mastadenovirus
SiAdV	Siadenovirus
THEV	Turkey Hemorrhagic Enteritis Virus
HE	Hemorrhagic Enteritis
UTR	Untranslated Region
TU	Transcription Unit
L4P	L4 Promoter
MLP	Major Late Promter
E3P	E3 Promoter
Hpi	Hours Post-infection
qPCR	Quantitative Polymerase Chain Reaction
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
CP	Coding Potential
TSS	Transcription Start Site
TTS	Transcription Termination Site
SC	Start Codon
STC	Stop Codon
secSC	Secondary Start Codon
secSTC	Secondary Stop Codon
ORF	Open Reading Frame
CDS	Coding Sequence
MLTU	Major Late Transcription Unit
TPL	Tripartite Leader
sTPL	Short Tripartite Leader
TPL3	Third exon of Tripartite Leader
GCN	Genome Copy Number

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12985-024-02449-0.

Supplementary Material 1

Acknowledgements

We thank the Office of Research Computing at Brigham Young University for granting us access to the high performance computing systems to perform the memory-intensive steps in the analysis pipeline of this work.

Author contributions
A.Q and B.D.P, conceptualized the work, A.Q., B.D.P, J.S.G. and B.K.B. designed the work; A.Q. acquired the data, A.Q. and B.E.P. analyzed and interpreted the data, A.Q. wrote the manuscript and B.D.P., J.S.G., B.E.P., and B.K.B. substantially revised the work.

Data availability
The raw sequencing read data (FastQ), transcript expression counts, and total unique junctions have been deposited at the National Center for Biotechnology Information Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) under accession number GSE254416.Data is also available on request by contacting the designated corresponding author.CODE AVAILABILITYAll the code/scripts in the entire analysis pipeline are available on github (http://www.ncbi.nlm.nih.gov/geo).

Code availability
All the code/scripts in the entire analysis pipeline are available on github .(https://github.com/Abraham-Quaye/thev_transcriptome).

Declarations

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Microbiology and Molecular Biology, Brigham Young University, 4007 Life Sciences Building (LSB), Provo, UT, USA

Received: 20 May 2024 / Accepted: 27 July 2024
Published online: 06 August 2024

References

1. Davison A, Benko M, Harrach B. Genetic content and evolution of adenoviruses. J Gen Virol. 2003;84:2895–908.
2. Harrach B. Adenoviruses: General features. In: Mahy BWJ, Van Regenmortel MHV, editors. Encyclopedia of virology (third edition). Oxford: Book Section. Academic; 2008. pp. 1–9. *In*.
3. Upton C, Slack S, Hunter AL, Ehlers A, Roper RL. Poxvirus orthologous clusters: toward defining the minimum essential poxvirus genome. J Virol. 2003;77:7590–600.
4. McGeoch D, Davison AJ. Chapter 17 - the molecular evolutionary history of the herpesviruses. In: Domingo E, Webster R, Holland J, editors. Origin and evolution of viruses. London: Book Section. Academic; 1999. pp. 441–65. *In*.
5. Harrach B, Benko M, Both GW, Brown M, Davison AJ, Echavarría M, Hess M, Jones M, Kajon A, Lehmkuhl HD, Mautner V, Mittal S, Wadell G. 2011. Family adenoviridae. Virus Taxonomy: 9th Report of the International Committee on Taxonomy of Viruses 125–141.
6. Kovács ER, Benkő M. Complete sequence of raptor adenovirus 1 confirms the characteristic genome organization of siadenoviruses. Infect Genet Evol. 2011;11:1058–65.
7. Davison AJ, Wright KM, Harrach B. DNA sequence of frog adenovirus. J Gen Virol. 2000;81:2431–9.
8. Kovács ER, Jánoska M, Dán Á, Harrach B, Benkő M. Recognition and partial genome characterization by non-specific DNA amplification and PCR of a new siadenovirus species in a sample originating from parus major, a great tit. J Virol Methods. 2010;163:262–8.
9. Katoh H, Ohya K, Kubo M, Murata K, Yanai T, Fukushi H. A novel budgerigar-adenovirus belonging to group II avian adenovirus of siadenovirus. Virus Res. 2009;144:294–7.
10. Guimet D, Hearing P. 2016. 3 - adenovirus replication, pp. 59–84. *In* Curiel, DT, editor, Adenoviral vectors for gene therapy (second edition). Book Section. Academic Press, San Diego.
11. Beach NM. 2006. Characterization of avirulent turkey hemorrhagic enteritis virus: A study of the molecular basis for variation in virulence and the occurrence of persistent infection. Thesis.
12. Beach NM, Duncan RB, Larsen CT, Meng XJ, Sriranganathan N, Pierson FW. Comparison of 12 Turkey hemorrhagic enteritis virus isolates allows prediction of genetic factors affecting virulence. J Gen Virol. 2009;90:1978–85.
13. Gross WB, Moore WE. Hemorrhagic enteritis of turkeys. Avian Dis. 1967;11:296–307.
14. Rautenschlein S, Sharma JM. Immunopathogenesis of haemorrhagic enteritis virus (HEV) in turkeys. Dev Comp Immunol. 2000;24:237–46.
15. Larsen CT, Domermuth CH, Sponenberg DP, Gross WB. Colibacillosis of turkeys exacerbated by hemorrhagic enteritis virus. Laboratory studies. Avian Dis. 1985;29:729–32.
16. Dhama K, Gowthaman V, Karthik K, Tiwari R, Sachan S, Kumar MA, Palanivelu M, Malik YS, Singh RK, Munir M. Haemorrhagic enteritis of turkeys – current knowledge. Veterinary Q. 2017;37:31–42.
17. Donovan-Banfield I, Turnell AS, Hiscox JA, Leppard KN, Matthews DA. Deep splicing plasticity of the human adenovirus type 5 transcriptome drives virus evolution. Commun Biology. 2020;3:124.
18. Zhao H, Chen M, Pettersson U. A new look at adenovirus splicing. Virology. 2014;456–457:329–41.
19. Wolfrum N, Greber UF. Adenovirus signalling in entry. Cell Microbiol. 2013;15:53–62.
20. Falvey E, Ziff E. Sequence arrangement and protein coding capacity of the adenovirus type 2 i leader. J Virol. 1983;45:185–91.
21. Morris SJ, Scott GE, Leppard KN. Adenovirus late-phase infection is controlled by a novel L4 promoter. J Virol. 2010;84:7096–104.
22. Westergren Jakobsson A, Segerman B, Wallerman O, Bergström Lind S, Zhao H, Rubin C-J, Pettersson U, Akusjärvi G. 2021. The human adenovirus 2 transcriptome: an amazing complexity of alternatively spliced mRNAs. J Virol 95.
23. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, 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 MJ, Gao H, Gonzalez D, Gordon A, Gunawardena H, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Luo OJ, Park E, Persaud K, Preall JB, Ribeca P, Risk B, Robyr D, Sammeth M, Schaffer L, See L-H, Shahab A, Skancke J, Suzuki AM, 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 SE, Hannon G, Giddings MC, Ruan Y, Wold B, Carninci P, Guigó R, Gingeras TR. Landscape of transcription in human cells. Nature. 2012;489:101–8.
24. Aboezz Z, Mahsoub H, El-Bagoury G, Pierson F. In vitro growth kinetics and gene expression analysis of the Turkey adenovirus 3, a siadenovirus. Virus Res. 2019;263:47–54.
25. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and ballgown. Nat Protoc. 2016;11:1650–67.
26. Jack Fu [Aut], Alyssa C. Frazee [Aut, cre], LeonardoCollado-Torres [Aut]. Ballgown. Bioconductor; 2017. Andrew E. Jaffe [Aut], Jeffrey T. Leek[Aut, Ths].
27. Pitcovski J, Mualem M, Rei-Koren Z, Krispel S, Shmueli E, Peretz Y, Gutter B, Gallili GE, Michael A, Goldberg D. The complete DNA sequence and genome organization of the avian adenovirus, hemorrhagic enteritis virus. Virology. 1998;249:307–15.
28. Beaudoing E, Freier S, Wyatt JR, Claverie J-M, Gautheret D. Patterns of variant polyadenylation signal usage in human genes. Genome Res. 2000;10:1001–10.
29. Yueh A, Schneider RJ. Selective translation initiation by ribosome jumping in adenovirus-infected and heat-shocked cells. Genes Dev. 1996;10:1557–67.
30. Akusjarvi G. 2008. Temporal regulation of adenovirus major late alternative RNA splicing. Front Bioscience Volume:5006.
31. Mahsoub HM, Evans NP, Beach NM, Yuan L, Zimmerman K, Pierson FW. Real-time PCR-based infectivity assay for the titration of Turkey hemorrhagic enteritis virus, an adenovirus, in live vaccines. J Virol Methods. 2017;239:42–9.
32. Green MR, Sambrook J. 2019. Rapid amplification of sequences from the 3′ ends of mRNAs: 3′-RACE. Cold Spring Harbor Protocols 2019:pdb.prot095216.
33. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kanitz A, Wilm A, Holtgrewe M, Rahmann S, Nahnsen S, Köster J. Sustainable data analysis with snakemake. F1000Research. 2021;10:33.
34. Krueger F, James F, Ewels P, Afyounian E, Weinstein M, Schuster-Boeckler B, Hulselmans G, Sclamons. 2023. FelixKrueger/TrimGalore: v0.6.10 - add default decompression path. Zenodo.

Publisher's Note