

¹ Characterizing the Transcriptome of Turkey Hemorrhagic
² Enteritis Virus in a Turkey B-cell Line

³

⁴ **Running Title:** Novel Insights into Turkey Hemorrhagic Enteritis Virus Transcriptome

⁵ Abraham Quaye^{1*}, Brett E. Pickett^{*}, Joel S. Griffitts^{*}, Bradford K. Berges^{*}, Brian D. Poole^{†**}

⁶ *Department of Microbiology and Molecular Biology, Brigham Young University

⁷ ¹First-author

⁸ [†] Corresponding Author

⁹ **Corresponding Author Information**

¹⁰ brian_poole@byu.edu

¹¹ Department of Microbiology and Molecular Biology,

¹² 4007 Life Sciences Building (LSB),

¹³ Brigham Young University,

¹⁴ Provo, Utah

¹⁵

16 **ABSTRACT**

17 Characterizing the splice map of turkey hemorrhagic enteritis virus (THEV) is an essential step that would
18 allow studies of individual genes mediating its immunosuppressive functions. We used an RNA-sequencing
19 experiment to characterize the transcriptome of THEV for the first time, providing key insight into the THEV
20 gene expression and mRNA structures. Researchers previously annotated the genome of THEV as encoding
21 23 open reading frames (ORFs). In this work we identified 29 spliced transcripts all of which consisted of
22 novel exons although some exons matched some previously annotated ORFs. The three annotated splice
23 junctions were also corroborated by our data. We performed PCR amplification of THEV cDNA, cloned the
24 PCR products, and used Sanger sequencing to validate all identified splice junctions. During validation we
25 identified five additional unique transcripts, a subset of which were further validated by 3' rapid amplification
26 of cDNA ends (3' RACE) experiments. Thus, we report that the genome of THEV contains 34 transcripts
27 with the coding capacity for all annotated ORFs. However, we found six of the previously annotated ORFs
28 (ORF1, E3, 33K, ORF8, IVa2, and protease) to be truncated ORFs on the basis of the identification of an
29 in-frame upstream start codon or the detection of additional coding exons. We also identified three of the
30 annotated ORFs with longer or shorter isoforms, and seven novel unannotated ORFs that could potentially
31 be translated; although it is beyond the scope of this manuscript to investigate whether they are translated.
32 Similar to other adenoviruses (AdVs), THEV also produces multiple distinctly spliced transcripts that code for
33 the same proteins across its genome. Our data show that all THEV transcripts are spliced and organized
34 into five transcription units under the control of their cognate promoters like other AdVs.

35 INTRODUCTION

36 Adenoviruses (AdVs) are non-enveloped icosahedral-shaped DNA viruses, causing infection in virtually all
37 types of vertebrates studied to date. Their double-stranded linear DNA genomes range between 26 and
38 45kb in size, producing a broad repertoire of transcripts via highly complex alternative splicing patterns (1,
39 2). The AdV genome is one of the most optimally economized; both the forward and reverse DNA strands
40 harbor protein-coding genes, making it highly gene-dense. There are 16 genes termed “genus-common”
41 that are homologous in all AdVs, presumably inherited from a common ancestor. All other genes are termed
42 “genus-specific”. The genus-specific genes tend to be located at the termini of the genome while genus-
43 common genes are usually towards the center of the genome (1). This pattern is also observed in *Poxviridae*
44 and *Herpesviridae*, which also have linear DNA genomes (1, 3, 4). The family *Adenoviridae* consists of five
45 genera: *Mastadenovirus* (MAdV), *Aviadenovirus*, *Atadenovirus*, *Ichtadenovirus*, and *Siadenovirus* (SiAdV) to
46 which turkey adenovirus 3 also called turkey hemorrhagic enteritis virus (THEV) belongs (5–10). Members of
47 SiAdV have the smallest genome size (~26 kb) and gene content of all known AdVs (see **Figure 1**) (1, 2, 6).

48 Virulent THEV strains (THEV-V) and avirulent strains (THEV-A) of THEV both infect turkeys, with THEV-
49 V causing hemorrhagic enteritis (HE), a debilitating acute disease predominantly affecting turkey pouls
50 characterized by immunosuppression, intestinal lesions leading to bloody diarrhea, and up to 80% mortality
51 (2, 11–13). While the current vaccine strain (a THEV-A called Virginia Avirulent Strain [VAS]) has proven
52 effective at preventing HE in turkey pouls, it still retains its immunosuppressive ability. Thus, vaccinated
53 birds are rendered more susceptible to opportunistic infections and death than unvaccinated birds leading to
54 substantial economic losses (11, 14–16). To eliminate the immunosuppressive effect of the vaccine strain,
55 a thorough investigation of the culprit viral genes mediating this phenomenon is essential. However, the
56 transcriptome (splicing and gene expression patterns) of THEV has not been characterized, making an
57 investigation of specific immunosuppressive viral genes impractical.

58 A myriad of studies have elucidated the AdV transcriptome in fine detail (17, 18). However, a large
59 preponderance of studies focus on MAdVs – specifically human AdVs. Thus, most of the current AdV gene
60 expression and replication knowledge is based on MAdV studies, which is generalized for all other AdVs (10,
61 19). MAdV transcription is temporally regulated; their genes are categorized into five early transcription units
62 (E1A, E1B, E2, E3, and E4), two intermediate (IM) units (pIX and IVa2), and one major late transcription unit
63 (MLTU or major late promoter [MLP] region), which generates five families of late mRNAs (L1-L5) based on
64 the polyadenylation site. An additional gene (UXP or U exon) is located on the reverse strand. The early
65 genes encode non-structural proteins such as enzymes or host-cell modulating proteins, primarily involved
66 in DNA replication, or providing the necessary intracellular niche for optimal replication while late genes

67 encode structural proteins that act as capsid proteins, promote virion assembly, or direct genome packaging.
68 The immediate early genes E1A are expressed first, followed by the delayed early genes, E1B, E2, E3 and
69 E4. Then the intermediate early genes, IVa2 and pIX are expressed followed by the late genes (10, 17,
70 18). It is noteworthy that the MLP shows basal transcriptional activity during early infection (before DNA
71 replication), with a comparable efficiency to other early viral promoters, but it reaches its maximal activity
72 during late infection (after DNA replication). However, during early infection only a subset of the MLP-derived
73 transcripts are expressed (10). MAdV makes an extensive use of alternative RNA splicing to produce a very
74 complex array of mRNAs. All but the pIX mRNA undergo at least one splicing event. For instance, the MLTU
75 produces over 20 distinct splice variants all containing three non-coding exons at the 5'-end (collectively
76 known as the tripartite leader; TPL) (17, 18). There is also an alternate three-exon 5' non-coding leader
77 sequence present in varying amounts on a subset of MLTU mRNAs (known as the x-, y-, and z-leaders).
78 Lastly, there is the i-leader exon, which is infrequently included between the second and third TPL exons,
79 and codes for the i-leader protein (20). Thus, the MLTU produces a complex repertoire of mRNA with diverse
80 5' untranslated regions (UTRs) spliced onto different 3' coding exons which are grouped into five different
81 3'-end classes (L1-L5) based on polyadenylation site. Each transcription unit (TU) contains its own promoter
82 driving the expression of the array of mRNA transcripts produced via alternative splicing in the unit (10, 17,
83 18). The promoters are activated at different phases of the infection by proteins from previously activated
84 TUs. Paradoxically, the early-to-late phase transition during infection requires the L4 gene products, 22K and
85 33K, which should only be available after the transition. However, a promoter in the L4 region (L4P) that
86 directs the expression of these two proteins independent of the MLP was found, resolving the paradox (10,
87 17, 21). During translation of AdV mRNA, recent studies using long-read direct RNA sequencing strongly
88 suggest the potential usage of secondary start codons; adding to what was already a highly complex system
89 for gene expression (17, 22).

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

97 **RESULTS**

98 **Overview of sequencing data and analysis pipeline outputs**

99 A prior study by Aboeazz *et al.* demonstrated that nearly all THEV transcripts became detectable starting
100 at 4 hours post-infection (hpi), with one replication cycle concluding around 18 hpi (24). Consequently,
101 we harvested infected MDTC-RP19 cells (MOI of 100 genome copy numbers/cell) at 4-, 12-, 24-, and
102 72-hpi to capture all transcripts within a broad time window. Our paired-end RNA sequencing (RNA-seq)
103 experiment generated an average of 107.1 million total reads of 149 bp length per time-point. These reads
104 were concurrently mapped to both the virus (THEV) and host (*Meleagris gallopavo*) genomes using the
105 Hisat2 (25) reference-based aligner. A total of 18.1 million reads from all time-points mapped to the virus
106 genome, providing comprehensive coverage and leaving no regions unmapped. The mapped reads to the
107 virus genome increased significantly from a scant 432 reads at 4 hpi to 16.9 million reads at 72 hpi (**Table**
108 **1, Figure 2A**). From these mapped reads, we identified 2,457 unique THEV splice junctions across all
109 time-points, with later time-points exhibiting significantly more sequence reads supporting the splice junctions
110 than earlier time-points. For instance, all 13 unique junctions at 4 hpi had fewer than 10 supporting reads
111 each, averaging only 2.8 reads per junction. In contrast, the 2374 unique junctions at 72 hpi averaged 898.4
112 reads per junction, with some junctions reaching as high as 322,677 reads. The marked increase in splice
113 junction and mapping reads to the THEV genome over time indicates an active infection and successful viral
114 replication, which is corroborated by our quantitative PCR (qPCR) assay that quantified the total number of
115 viral genome copies over time (**Figure 2B**).

116 Using StringTie (25), we assembled the data into potential transcripts, guided by the genomic locations of
117 the previously predicted THEV ORFs. In the consolidated transcriptome, a composite of all non-redundant
118 transcripts across all time points, we identified a total of 29 novel transcripts. We found that a subset of
119 exons in the viral transcripts match some predicted ORFs exactly, with the majority of the exons being longer
120 and spanning multiple predicted ORFs (**Figure 3**).

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

127 **Changes in THEV splicing profile over time**

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

135 Examining individual mRNAs, TRXPT_21 – from the E2 region – was the most significantly expressed at
136 12 hpi, constituting 33.58% of the total expression of all transcripts. Transcripts in the E3 and E4 regions
137 also contributed substantial proportions, along with some MLP region transcripts. The later time points were
138 dominated by the MLP region transcripts — TRXPT_10 and TRXPT_14 were the most abundantly expressed
139 at 24 and 72 hpi, respectively (**Figure 4A**). Our analysis of the FPKM values of transcripts per region/TU
140 revealed a similar pattern: the E2 region was the most abundantly expressed at 12 hpi, after which the MLP
141 region assumed dominance (**Figure 4B**).

142 Next, we estimated the relative abundances of all splice junctions at each time point using the raw reads.
143 Only junctions with a read coverage of at least 1% of the total splice junction reads at the given time point
144 were considered significant and included in **Tables 2a-2c**. At 12 hpi, 18 junctions met the 1% threshold,
145 predominantly from early regions (E1, E2, E3, and E4), although the MLTU was the single most predominant
146 region overall, constituting 38.8% of all the junction reads (**Table 2a** and **Supplementary Table S1a**). The
147 most abundant junctions at 12 hpi remained the most significantly expressed at 24 hpi. However, here, the
148 MLP-derived junctions unsurprisingly became even more predominant overall, accounting for 45.7% of all
149 the junction reads counted (**Table 2b** and **Supplementary Table S1b**). At 72 hpi, the trend of increased
150 activity of the MLP continued as expected; at this time, the MLP region junctions were not only the most
151 abundant overall — accounting for 67.3% of all junction reads, — but also contained the most significantly
152 expressed individual junctions (**Table 2c**, **Supplementary Table S1c** and **Figure 4C**). When we limited this
153 analysis to only junctions in the final transcriptome, we observed the relative abundances of the junctions for
154 each region over time to be similar to the pattern seen with all the junctions included (**Figure 4D**).

155 Finally, we analyzed splice donor and acceptor site nucleotide usage over time to investigate any peculiarities
156 that THEV may exhibit, generally or over the course of the infection. We found that most splice donor-acceptor
157 sequences were, unsurprisingly, the canonical GU-AG nucleotides. However, the splice acceptor-donor
158 pairing became less specific over time, such that all combinations of nucleotide pairs were eventually detected
159 (**Figure 5**).

160 **Early Region 1 (E1) transcripts**

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

166 We identified four novel transcripts in this region, containing 3 unique splice junctions (**Figure 6**), and
167 encoding four distinct novel ORFs in addition to Hyd. All transcripts have coding potential (CP) for the Hyd
168 protein as the 3'-most coding sequence if secondary start codon usage is considered (17, 18). Also, all
169 the transcripts have a common transcription termination site (TTS; at position 2325 bp), but TRXPT_1 and
170 TRXPT_2 have an upstream transcription start sites (TSS) to TRXPT_3 and TRXPT_4. Given that E1
171 mRNAs in MAdVs share a common TTS and TSS, differing only in the internal splicing (18), we consider
172 the upstream TSS (position 54 bp) as the actual TSS for all E1 region transcripts. We also identified the
173 canonical polyadenylation signal (PAS; AAUAAA) in the immediate context of the TTS at position 2323 bp
174 (location of the “U” in the PAS sequence); see **Supplementary Table S2**.

175 From the 5'-most start codon (SC), TRXPT_1 encodes a multi-exonic novel 17.9 kDa, 160 residue protein
176 (ORF9). TRXPT_2 encodes a two-exon 66.4 kDa, 597 residue novel protein (ORF10), spanning almost the
177 entire predicted ORF1 and Hyd. The intron of TRXPT_2 excludes the C-terminus of ORF1 (including its
178 stop codon) from ORF10 but the SC of ORF10 is 102 bp upstream and in-frame with the predicted SC of
179 ORF1. TRXPT_3, similar to TRXPT_1 but lacking the second exon, encodes a 13.1 kDa, 115 residue protein
180 (ORF4), previously predicted (27) but excluded in later annotations (1, 12). Our data suggest it is genuinely
181 expressed. Lastly, TRXPT_4 encodes a distinct novel 15.9 kDa, 143 residue protein (ORF11).

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

192 **Early Region 2 (E2) and Intermediate Region (IM) transcripts**

193 The E2 TU expressed on the anti-sense strand, is subdivided into E2A and E2B and encodes three classical
194 AdV proteins – pTP and Ad-pol (E2B proteins), and DBP (E2A protein) – essential for genome replication (17,
195 18). Unlike MAdV where two promoters are known (17), we discovered only a single TSS (E2 TSS; 18,751
196 bp) for both E2A and E2B transcripts in THEV. However, similar to MAdVs, E2A and E2B transcripts have
197 distinct TTSs, with E2B transcripts sharing the TTS of the IVa2 transcript of the IM region (17, 18) (**Figure 7**).

198 The E2A ORF, DBP, is one of three THEV ORFs predicted to be spliced from two exons. The corresponding
199 transcript (TRXPT_21) in our data matches the predicted splice junction but includes an additional non-coding
200 exon at the 5'-end (E2-5'UTR). Thus, TRXPT_21 is a three-exon transcript encoding DBP (380 residues,
201 43.3 kDa) precisely. TRXPT_21 was also corroborated in a 3' RACE experiment. Additionally, from the 3'
202 RACE data, we found a splice variant of TRXPT_21 which retains the second intron, leading to a 2-exon
203 transcript (TRXPT_21B). Although longer, TRXPT_21B encodes a truncated isoform of DBP (tDBP; a 346
204 residue, 39.3 kDa product) using a downstream in-frame SC but the same STC as DBP. Both TRXPT_21 and
205 TRXPT_21B share a common TTS, seen in our 3' RACE data, located 39 bp downstream of the CDS in an
206 adenine/thymine (A/T)-rich sequence followed by the poly-A tail sequence, suggesting this position (16,934
207 bp) as the true E2A TTS. There are two canonical PASs (AAUAAA; 16,964 and 16950 bp) immediately after
208 the CDS any of which can serve as the PAS without affecting the encoded proteins (**Supplementary Table**
209 **S2**).

210 The E2B region transcripts also start with the E2-5'UTR but extend downstream to reach the TTS at
211 2334 bp in the IM region, which is in the immediate context of a canonical PAS (position 2333 bp) where
212 polyadenylation likely occurs. The E2B transcripts, TRXPT_6 and TRXPT_7, are almost identical except
213 for an extra splice junction at the 3'-end of TRXPT_6 (**Figure 7**). TRXPT_7 has the CP for both classical
214 proteins (pTP and Ad-pol) encoded in this region, with the pTP ORF predicted to be spliced from two exons.
215 The predicted splice junction of pTP is corroborated by our data but the full transcript is markedly longer
216 than the predicted ORF, although, the encoded product (pTP) remains unchanged. Ad-pol (polymerase) is
217 encoded downstream of pTP with secondary SC (secSC) usage. The CP of TRXPT_6 slightly differs from
218 TRXPT_7 because a new STC resulting from the extra splice site forms a minimal truncation of the Ad-pol
219 encoded from its secSC.

220 While both TRXPT_6 and TRXPT_7 have the CP for Ad-pol with secSC usage, in all AdVs studied, the
221 two proteins (pTP and Ad-pol) are encoded by separate mRNAs with identical first three 5' exons and TTS,
222 but different splice junctions to the terminal coding exons. Hence, we checked for a longer splice junction
223 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 (**Figure 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 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 transcripts in this TU. The 12.5K 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 (100K, 22K, 33K, and pVIII) overlap the E3 region transcripts entirely and share similar TSS and TTS locations (**Figure 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, 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 PAs (28) near the TTS of the transcripts (see **Supplementary Table S2**). These E3 transcripts collectively have the CP for several predicted THEV ORFs: 100K, 22K, 33K, 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 33K, 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 33K (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, but also contained the second exon of TRXPT_24 (**Figure 8**). TRXPT_30 encodes a novel 140 residue, 15.7kDa protein (e22K), spanning all three exons. Interestingly, the last 81 C-terminus residues of e22K are identical to 22K (89 residues), a single-exon ORF predicted to use the same SC as 33K. Just as seen for 33K, the first 20 bp of 22K is intronic, excluding the first 7 residues of 22K from e22K. We consider e22K as a long variant of the predicted 22K 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 100K ORF. secSC usage on this transcript yields the predicted 22K 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 100K protein due to its splice junction, this band (which we cloned and validated by Sanger sequencing) suggests that the predicted 100K 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

287 pVIII as the 5'-most ORF and has the CP for eE3 and Fiber in that order. TRXPT_27, a two-exon transcript,
288 encodes Fiber as the 5'-most ORF, and ORF7 downstream with secSC usage. TRXPT_13 is an L4P
289 transcript that uses the MLP TSS; it is discussed under the MLTU transcripts.

290 **Early Region 4 (E4) transcripts** This TU is found at the 3'-end of the genome and expressed on the
291 anti-sense strand. Based on nucleotide position, ORF7 and ORF8 were predicted in this region (1); however,
292 as ORF7 is neither on the anti-sense strand nor transcribed from a promoter in the E4 region; hence, we
293 only classify ORF8 in this TU. This is corroborated by our RNA-seq data, showing only one transcript in this
294 region on the anti-sense strand (**Figure 9**). The transcript (TRXPT_28) spans 25192-26247 bp and is spliced
295 at 25701-26055 bp, forming a two-exon transcript. The second exon fully matches the predicted ORF8 with
296 12 extra base pairs at the 3'-end. However, we identified a SC 192 bp upstream of the predicted SC in the
297 first exon from which an in-frame protein is encoded. We consider this longer isoform (eORF8 – 26.4 kDa,
298 229 residues) as the genuinely expressed ORF. We also identified a canonical PAS 11 bp upstream of TTS
299 (**Supplementary Table S2**).

300 **Major Late Transcription Unit (MLTU) or MLP Region transcripts**

301 The MLTU transcripts, dominant in the late phase of the AdV infectious cycle, are produced by alternative
302 polyadenylation and splicing of a primary transcript and grouped into five transcript classes (L1-L5). About
303 13 out of the 23 predicted ORFs in THEV fall within this TU, some of which we have categorized under the
304 E3 TU instead. Our RNA-seq data revealed 12 transcripts (TRXPT_8, TRXPT_9, TRXPT_10, TRXPT_11,
305 TRXPT_12, TRXPT_13, TRXPT_14, TRXPT_16, TRXPT_17, TRXPT_18, TRXPT_19, TRXPT_20) in this
306 TU, most of which have the 5' TPL sequence as in all AdVs. However, three transcripts (TRXPT_16,
307 TRXPT_17, TRXPT_18) use a different leader sequence (sTPL), where a different first exon is used instead
308 of the first TPL exon, and TRXPT_20 uses only the third TPL exon (TPL3); see **Figure 10**.

309 We identified five TTSs (10,549, 12,709, 16,870, 22,116, 25,168 bp) in this TU, which we consider as
310 corresponding to the five late mRNA classes (L1-L5), respectively. L1 mRNAs include TRXPT_8, encoding
311 the 52K ORF as predicted. L2 mRNAs include TRXPT_16, TRXPT_17, and TRXPT_18, all containing the
312 sTPL with their respective coding exons. They encode pIIa, III (penton), and pVII, respectively. The L3
313 mRNAs, TRXPT_14 and TRXPT_20, both encode the hexon (II) ORF but hexon is the only ORF encoded on
314 TRXPT_14, whereas TRXPT_20 encodes pX (pre-Mu), pVI, and hexon in that order. L4 mRNAs, TRXPT_9,
315 TRXPT_10, TRXPT_11, and TRXPT_13 are the largest transcripts in the transcriptome and encode several
316 similar late proteins. TRXPT_9 and TRXPT_10 are very similar but not identical. The last exon of TRXPT_9
317 seems to be truncated and likely shares the same TTS as TRXPT_10. They both encode pVII as the 5'-most
318 ORF and also have the CP for pX, pVI, hexon, a longer variant of protease (eProt) from an upstream in-frame

319 SC, and ORF12 (a novel 120 residue protein). Additionally they have the CP for pVIII and eE3. TRXPT_11
320 encodes hexon as its 5'-most ORF and also has the CP for eProt, ORF12, e33K, pVIII and eE3. Typically,
321 MLTU transcripts splice the TPL onto a splice site just upstream of the ORF to be expressed (17). While
322 this holds true for most MLTU ORFs, several late ORFs (pVI, protease, and ORF7) do not have such close
323 proximity splicing but are contained in larger transcripts such as these L4 mRNAs, strongly suggesting the
324 use of non-standard ribosomal initiation mechanisms such as secSC usage or ribosome shunting described
325 in other AdVs for their translation (17, 29). TRXPT_13, an E3 ORF utilizing the MLP TSS, encodes the
326 classical L4P genes, pVIII and eE3. Lastly, the L5 class transcript, TRXPT_12, encodes Fiber as its 5'-most
327 ORF but also has the CP for ORF7. Interestingly, the CP of TRXPT_12 and TRXPT_27 of the E3 TU are
328 identical but are initiated from different TSSs.

329 **DISCUSSION/CONCLUSIONS**

330 While the advent of next-generation sequencing has rendered easier the study of large and complex eukary-
331 otic transcriptomes, studying the smaller, compact viral transcriptomes remains unintuitively challenging,
332 as the transcripts typically have significant overlaps due to genome economization. AdV transcriptomes
333 escalate the difficulty due to the wide array of mRNAs produced via very complex alternative splicing and
334 polyadenylation, all initiated from relatively few promoters. Standard RNA-seq analysis programs, not primar-
335 ily designed for such compact, gene-dense, and complex transcriptomes, further compound this challenge.
336 Furthermore, in our case, no prior transcriptomic studies for THEV exist. Our approach combines standard
337 RNA-seq analysis programs with custom analyses and experimentally validating all splice junctions with
338 independent methods. The transcript map for THEV produced from our analysis is strikingly similar to that of
339 the MAdVs.

340 Our work provides the first insights into THEV splicing, revealing 34 transcripts grouped into five transcription
341 units (TUs). The general temporal gene expression regulation observed in MAdVs, with early regions peaking
342 at earlier time points followed by MLTU predominance at later time points, seems to also apply to THEV.
343 An unexpected observation is that the pileup of mapped reads to THEV seems consistently skewed over
344 similar regions of the genome at all time points. Given the temporal regulation of AdVs gene expression, we
345 anticipated distinct differences in read pileups over the genome at different time points, indicating the different
346 stages of infection. This probably due to the infection not being well-synchronized, leading to transcripts
347 overlapping the time points. Further research is required to determine the precise temporal regulation of
348 THEV.

349 Short read deep sequencing effectively reconstructs full AdV mRNA structures, particularly mapping splice
350 sites (18). However, the substantial overlapping nature of AdV mRNAs and fragmentation during library
351 preparation make it challenging to map the exact TSS, TTS, and PAs of assembled transcripts. Also,
352 transcripts with significant overlaps may be assembled as a single, longer mRNA, since the short reads
353 alone do not provide enough context for the transcript assembler (StringTie) to distinguish them. Our results
354 show transcripts in the same TU initiating or terminating in similar areas, but not at the exact same position.
355 We consider the most upstream TSS or most downstream TTS for the transcripts involved but we present
356 them unchanged in all the figures shown (see **Supplementary Table S2**). Also, comparing our results to the
357 better-studied MAdV transcriptomes, we believe some long transcripts in the MLTU (TRXPT_9, TRXPT_10,
358 and TRXPT_11) are likely due to fusing some E3 transcripts to the terminal exons of the MLTU transcripts by
359 StringTie, making them significantly longer. These mRNAs have unusually many exons and their last few
360 exons are identical to some E3 mRNAs. Future studies using long read sequencing technologies will provide

more precise mapping of the TSS and TTS 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.

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 start codon 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 (**Figure 5**). Recent studies show

393 that a virtually unlimited number of combinatorial alternative splicing events occur in an AdV lytic infection,
394 resulting in a variety of novel transcripts (17, 22). It is unlikely that the entire repertoire of mRNA produced via
395 this mechanism will actually be translated. However, it has been speculated that the plasticity in alternative
396 RNA splicing enables AdVs to fine-tune protein synthesis by providing different alternatively spliced variants
397 encoding the same protein under changing conditions, conferring an evolutionary advantage (17, 22).

398 In summary, the THEV transcriptome bears remarkable similarity to the better-studied MAdVs. The tran-
399 scriptome is organized into five TUs, with temporal regulation divided into early and late genes, and a
400 broad repertoire of transcripts are produced via virtually unlimited alternative splicing. However, the THEV
401 transcriptome appears less sophisticated (i.e, it encodes fewer genes) than MAdVs, primarily because the
402 MAdV genomes are close to twice as long as that of THEV. The lack of subdivision of the E1 region into E1a
403 and E1b is one of the most obvious examples. Also, the MAdV E4 region encodes several proteins unlike
404 in THEV where only one transcript encoding one protein was found. The complexity of the MLTU leader
405 sequences is another example. While the majority of the THEV MLTU transcripts begin with the TPL just like
406 MAdVs with a small subset using a variant leader sequence (sTPL), significantly more diverse 5'-UTRs are
407 employed for MAdV MLTU transcripts. Namely, the TPL, the so-called x, y, and z leaders, and the i-leader
408 are 5' leaders utilized by MAdV MLTU mRNAs. The absence of these non-TPL leaders in our data could
409 mean that the 5'-UTR diversity of THEV's MLTU mRNAs is more limited due to its smaller genome size or
410 future studies using long read sequencing technologies could uncover more variety not seen in our results.

411 **MATERIALS AND METHODS**

412 **Cell culture and THEV Infection**

413 The Turkey B-cell line (MDTC-RP19, ATCC CRL-8135) was grown as suspension cultures in 1:1 complete
414 Leibovitz's L-15/McCoy's 5A medium with 10% fetal bovine serum (FBS), 20% chicken serum (ChS), 5%
415 tryptose phosphate broth (TPB), and 1% antibiotic solution (100 U/mL Penicillin and 100 μ g/mL Streptomycin),
416 at 41°C in a humidified atmosphere with 5% CO₂. Infected cells were maintained in 1:1 serum-reduced
417 Leibovitz's L15/McCoy's 5A media (SRLM) with 2.5% FBS, 5% ChS, 1.2% TPB, and 1% antibiotic solution. A
418 commercially available THEV vaccine was purchased from Hygieia Biological Labs as a source of THEV-A
419 (VAS strain). The stock virus was titrated using an in-house qPCR assay with titer expressed as genome
420 copy number (GCN)/mL, similar to Mahshoub *et al* (31) with modifications. Cells were infected in triplicate
421 at a multiplicity of infection (MOI) of 100 GCN/cell, incubated at 41°C for 1 hour, and washed three times
422 with phosphate buffered saline (PBS) to get rid of free virus particles. Triplicate samples were harvested at
423 4-, 12-, 24-, and 72-hpi for total RNA extraction. The infection was repeated but samples in triplicate were
424 harvested at 12-, 24-, 36-, 48-, and 72-hpi for PCR validation of novel splice sites. Still one more independent
425 infection was done at time points ranging from 12 to 168-hpi for qPCR quantification of virus titers.

426 **RNA extraction and Sequencing**

427 Total RNA was extracted from infected cells using the Thermo Fisher RNAqueous™-4PCR Total RNA Isolation
428 Kit (which includes a DNase I digestion step) per manufacturer's instructions. An agarose gel electrophoresis
429 was performed to check RNA integrity. The RNA quantity and purity was initially assessed using nanodrop,
430 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.
431 Extracted total RNA samples were sent to LC Sciences, Houston TX for poly-A-tailed mRNA sequencing
432 where RNA integrity was checked with Agilent Technologies 2100 Bioanalyzer High Sensitivity DNA Chip
433 and poly(A) RNA-seq library was prepared following Illumina's TruSeq-stranded-mRNA sample preparation
434 protocol. Paired-end sequencing to generate 150 bp reads was performed on the Illumina NovaSeq 6000
435 sequencing system.

436 **Validation of Novel Splice Junctions**

437 All splice junctions identified in this work are novel except one predicted splice site each for pTP, DBP,
438 and 33K, which were corroborated in our work. However, these predicted splice junctions had not been
439 experimentally validated hitherto, and we identified additional novel exons, giving the complete picture of

440 these transcripts. The novel splice junctions discovered in this work using the StringTie transcript assembler
441 were validated by PCR, cloning, and Sanger Sequencing (**Supplementary PCR methods**). Briefly, primers
442 spanning a range of novel exon-exon boundaries for each specific transcript in a transcription unit (TU) were
443 designed. Universal forward or reverse primers for each respective TU were designed and paired with primers
444 binding specific positions in each transcript. Each forward primer contained a KpnI restriction site and each
445 reverse primer, an XbaI site in the primer 5' ends. After first-strand cDNA synthesis of total RNA obtained
446 from THEV infected MDTC-RP19 cells was done using SuperScript™ IV First-Strand Synthesis System, the
447 primers were used in a targeted PCR amplification, the products analyzed with agarose gel electrophoresis
448 to confirm expected band sizes, cloned by traditional restriction enzyme method, and Sanger sequenced
449 to validate these splice junctions at the sequence level. The total RNA was extracted as described above,
450 including the DNase I digestion step. We included infected total RNA controls with no reverse transcriptase
451 (no RT) during the cDNA synthesis step and the parent RNA were digested using RNase H after cDNA
452 synthesis was complete to ensure that the bands obtained from the targeted PCR amplifications did not
453 originate from the viral genomic DNA. As seen in the agarose gel images in **Supplementary PCR methods**,
454 DNA bands were not found in the “no RT” controls, indicating that the DNA bands seen are of cDNA origin.

455 **3' Rapid Amplification of cDNA Ends (3' RACE)**

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

465 **Computational Analysis of RNA Sequencing Data: Mapping and Transcript characterization**

466 Sequencing reads were analyzed following a well-established protocol described by Pertea *et al* (25), using
467 Snakemake - version 7.24.0 (33), a popular workflow management system to drive the pipeline. Briefly,
468 sequencing reads were trimmed with the Trim-galore - version 0.6.6 (34) program to achieve an overall Mean
469 Sequence Quality (Phred Score) of 36. Trimmed reads were mapped simultaneously to the complete genomic

470 sequence of avirulent turkey hemorrhagic enteritis virus (<https://www.ncbi.nlm.nih.gov/nuccore/AY849321.1/>)
471 and *Meleagris gallopavo* (<https://www.ncbi.nlm.nih.gov/genome/?term=Meleagris+gallopavo>) using Hisat2
472 - version 2.2.1 (25) with default settings. The generated binary alignment (BAM) files from each infection
473 time point were filtered for reads mapping to the THEV genome using Samtools - version 1.16.1 and fed into
474 StringTie - version 2.2.1 (25) to assemble the transcripts, using a gene transfer format (GTF) annotation
475 file derived from a gene feature format 3 (GFF3) annotation file obtained from NCBI, which contains the
476 predicted ORFs of THEV as a guide. GFFCOMPARE - version 0.12.6 was used to merge all transcripts
477 from all time points without redundancy and using a custom R script, adenovirus transcripts units (regions)
478 were assigned to each transcript, generating the transcriptome of THEV. StringTie was set to expression
479 estimation mode to calculate FPKM scores for all transcripts after which Ballgown - version 2.33.0 in R was
480 used to perform the statistical analysis on the transcript expression levels. Samtools was also used to count
481 the total sequencing reads for all replicates at each time point and Regtools - version 1.0.0 was used to
482 count all junctions, the reads supporting them, and extract all other information related to the junction. See
483 **Supplementary Computational Analysis** for the details of transcript expression level estimations and splice
484 junction read counts.

485 DATA AVAILABILITY

486 The raw sequencing read data (FastQ), transcript expression counts, and total unique junctions have
487 been deposited at the National Center for Biotechnology Information Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo>) under accession number GSE254416.
488
489 Data is also available on request by contacting the designated corresponding author.

490 CODE AVAILABILITY

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

493 ACKNOWLEDGMENTS

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

496 **REFERENCES**

- 497 1. Davison A, Benko M, Harrach B. 2003. Genetic content and evolution of adenoviruses. *The Journal of general virology* 84:2895–908.
- 498 2. Harrach B. 2008. Adenoviruses: General features, p. 1–9. *In* Mahy, BWJ, Van Regenmortel, MHV (eds.), *Encyclopedia of virology* (third edition). Book Section. Academic Press, Oxford.
- 499 3. Upton C, Slack S, Hunter AL, Ehlers A, Roper RL. 2003. Poxvirus orthologous clusters: Toward defining the minimum essential poxvirus genome. *Journal of virology* 77:7590–7600.
- 500 4. McGeoch D, Davison AJ. 1999. Chapter 17 - the molecular evolutionary history of the herpesviruses, p. 441–465. *In* Domingo, E, Webster, R, Holland, J (eds.), *Origin and evolution of viruses*. Book Section. Academic Press, London.
- 501 5. Harrach B, Benko M, Both GW, Brown M, Davison AJ, Echavarría M, Hess M, Jones M, Kajon A, Lehmkühl HD, Mautner V, Mittal S, Wadell G. 2011. Family adenoviridae. *Virus Taxonomy: 9th Report of the International Committee on Taxonomy of Viruses* 125–141.
- 502 6. Kovács ER, Benkő M. 2011. Complete sequence of raptor adenovirus 1 confirms the characteristic genome organization of siadenoviruses. *Infection, Genetics and Evolution* 11:1058–1065.
- 503 7. Davison AJ, Wright KM, Harrach B. 2000. DNA sequence of frog adenovirus. *J Gen Virol* 81:2431–2439.
- 504 8. Kovács ER, Jánoska M, Dán Á, Harrach B, Benkő M. 2010. 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. *Journal of Virological Methods* 163:262–268.
- 505 9. Katoh H, Ohya K, Kubo M, Murata K, Yanai T, Fukushi H. 2009. A novel budgerigar-adenovirus belonging to group II avian adenovirus of siadenovirus. *Virus Research* 144:294–297.

- 506 10. Guimet D, Hearing P. 2016. 3 - adenovirus replication, p. 59–84. In Curiel, DT (ed.), Adenoviral
vectors for gene therapy (second edition). Book Section. Academic Press, San Diego.
- 507 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.
- 508 12. Beach NM, Duncan RB, Larsen CT, Meng XJ, Sriranganathan N, Pierson FW. 2009. Comparison of
12 turkey hemorrhagic enteritis virus isolates allows prediction of genetic factors affecting virulence. J
Gen Virol 90:1978–85.
- 509 13. Gross WB, Moore WE. 1967. Hemorrhagic enteritis of turkeys. Avian Dis 11:296–307.
- 510 14. Rautenschlein S, Sharma JM. 2000. Immunopathogenesis of haemorrhagic enteritis virus (HEV) in
turkeys. Dev Comp Immunol 24:237–46.
- 511 15. Larsen CT, Domermuth CH, Sponenberg DP, Gross WB. 1985. Colibacillosis of turkeys exacerbated
by hemorrhagic enteritis virus. Laboratory studies. Avian Dis 29:729–32.
- 512 16. Dhami K, Gowthaman V, Karthik K, Tiwari R, Sachan S, Kumar MA, Palanivelu M, Malik YS, Singh
RK, Munir M. 2017. Haemorrhagic enteritis of turkeys – current knowledge. Veterinary Quarterly
37:31–42.
- 513 17. Donovan-Banfield I, Turnell AS, Hiscox JA, Leppard KN, Matthews DA. 2020. Deep splicing plasticity
of the human adenovirus type 5 transcriptome drives virus evolution. Communications Biology 3:124.
- 514 18. Zhao H, Chen M, Pettersson U. 2014. A new look at adenovirus splicing. Virology 456-457:329–341.
- 515 19. Wolfrum N, Greber UF. 2013. Adenovirus signalling in entry. Cell Microbiol 15:53–62.
- 516 20. Falvey E, Ziff E. 1983. Sequence arrangement and protein coding capacity of the adenovirus type 2
"i" leader. Journal of Virology 45:185–191.

- 517 21. Morris SJ, Scott GE, Leppard KN. 2010. Adenovirus late-phase infection is controlled by a novel L4 promoter. *Journal of Virology* 84:7096–7104.
- 518 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. *Journal of Virology* 95.
- 519 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. 2012. Landscape of transcription in human cells. *Nature* 489:101–108.
- 520 24. Aboezz Z, Mahsoub H, El-Bagoury G, Pierson F. 2019. In vitro growth kinetics and gene expression analysis of the turkey adenovirus 3, a siadenovirus. *Virus Research* 263:47–54.
- 521 25. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and ballgown. *Nature Protocols* 11:1650–1667.
- 522 26. Jack Fu [Aut], Alyssa C. Frazee [Aut, Cre], Leonardo Collado-Torres [Aut], Andrew E. Jaffe [Aut], Jeffrey T. Leek [Aut, Ths]. 2017. Ballgown. Bioconductor.
- 523 27. Pitcovski J, Mualem M, Rei-Koren Z, Krispel S, Shmueli E, Peretz Y, Gutter B, Gallili GE, Michael A, Goldberg D. 1998. The complete DNA sequence and genome organization of the avian adenovirus, hemorrhagic enteritis virus. *Virology* 249:307–315.

- 524 28. Beaudoin E, Freier S, Wyatt JR, Claverie J-M, Gautheret D. 2000. Patterns of variant polyadenylation
signal usage in human genes. *Genome Research* 10:1001–1010.
- 525 29. Yueh A, Schneider RJ. 1996. Selective translation initiation by ribosome jumping in adenovirus-infected
and heat-shocked cells. *Genes & Development* 10:1557–1567.
- 526 30. Akusjarvi G. 2008. Temporal regulation of adenovirus major late alternative RNA splicing. *Frontiers in
Bioscience Volume:5006*.
- 527 31. Mabsoub HM, Evans NP, Beach NM, Yuan L, Zimmerman K, Pierson FW. 2017. Real-time PCR-based
infectivity assay for the titration of turkey hemorrhagic enteritis virus, an adenovirus, in live vaccines.
Journal of Virological Methods 239:42–49.
- 528 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.
- 529 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. 2021. Sustainable data
analysis with snakemake. *F1000Research* 10:33.
- 530 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.

531 **TABLES AND FIGURES**

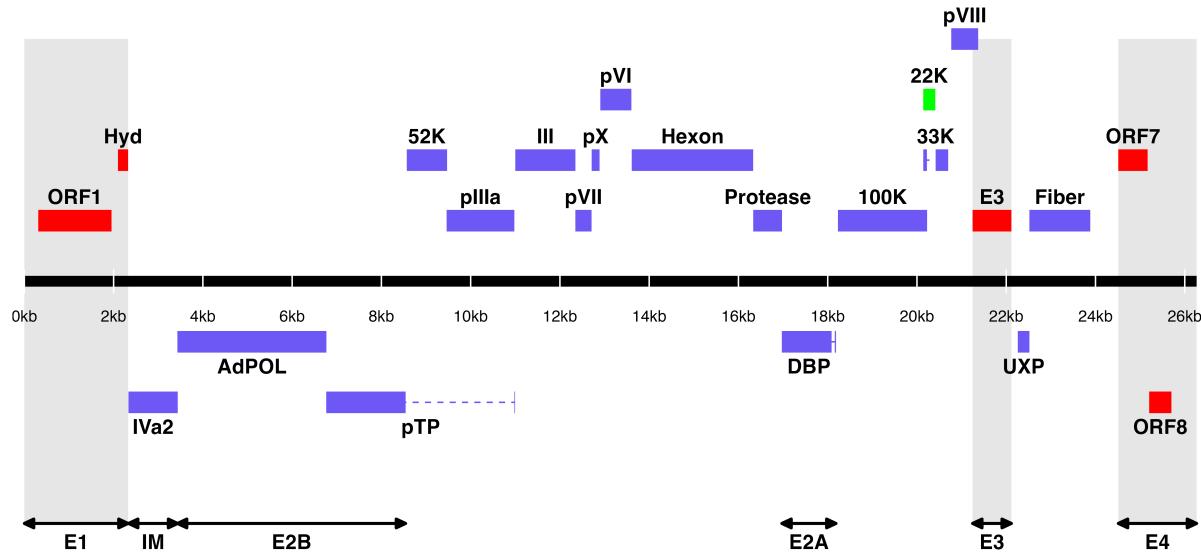
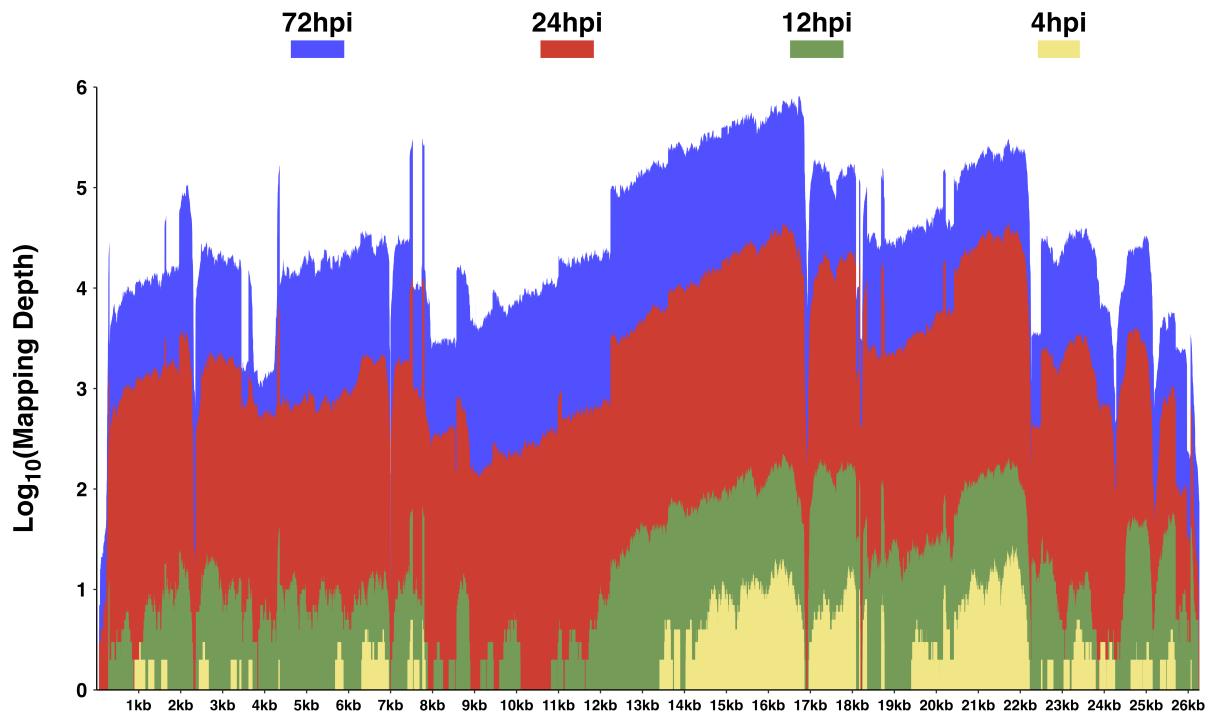


Figure 1. Predicted ORF map of THEV avirulent strain. The central horizontal line represents the double-stranded DNA marked at 2kb intervals as white line breaks. Colored blocks represent viral genes. Blocks above the DNA line are transcribed on the sense DNA strand and those below, on the anti-sense strand. pTP, DBP and 33K are predicted to be spliced and are shown as two exons connected with dashed lines. Shaded regions indicate regions containing “genus-specific” genes (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 transcription units are labelled at the bottom (E1, E2A, E2B, E3, E4, and IM); the unlabeled regions comprise the MLTU.

A

Mapping Depth of RNA-seq Reads Over THEV Genome

**B**

THEV Growth Curve in RP-19 Turkey B-Cells

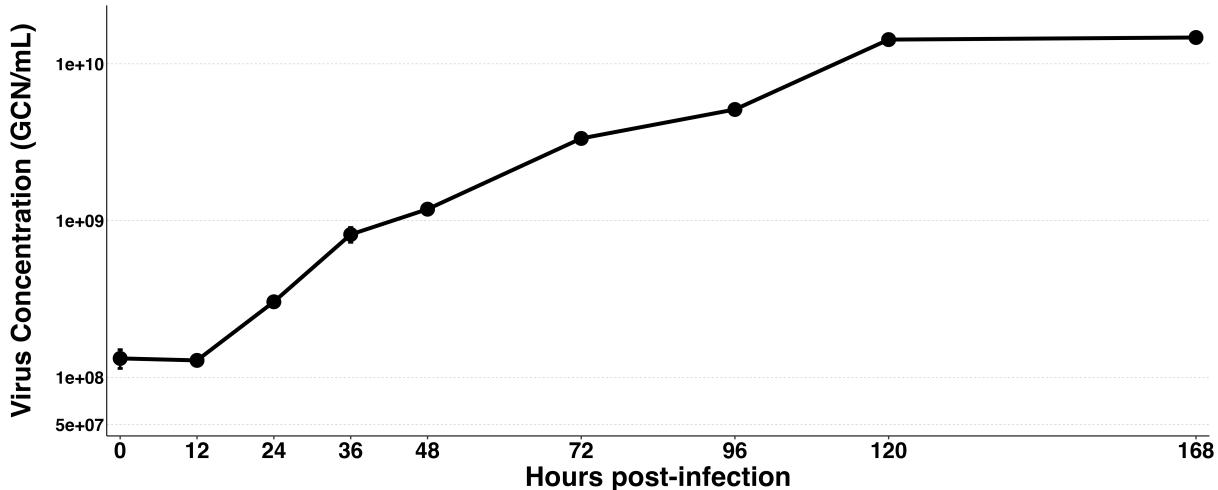


Figure 2: Increasing levels of THEV over time. A) Per base coverage of sequence reads mapping to THEV genome by time point. The pileup of mRNA reads mapping to THEV genome at the base-pair level for each indicated time point. B) Growth curve of THEV (VAS vaccine strain) in MDTC-RP19 cell line. Virus quantities in the freeze-thawed supernatant from infected cells were 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.

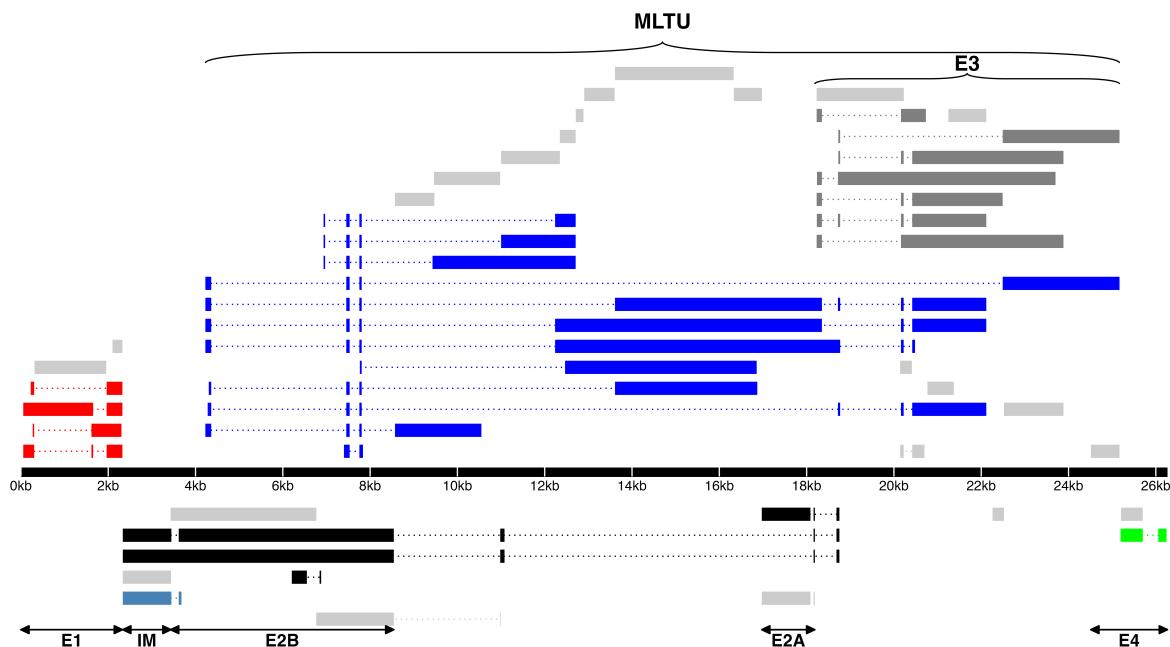
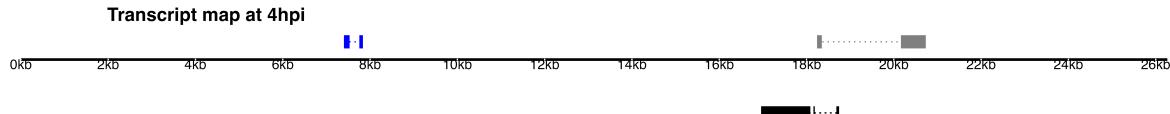
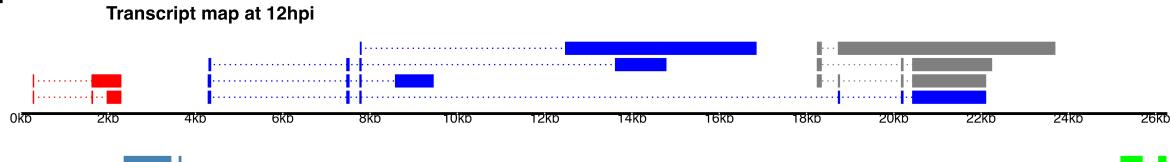
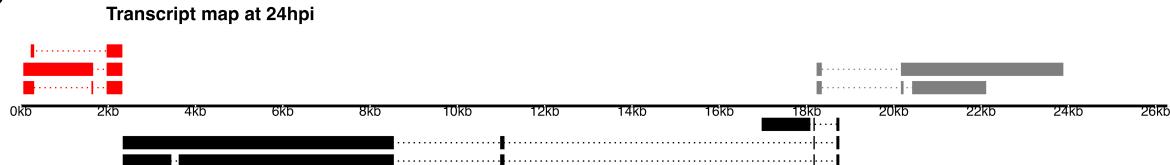
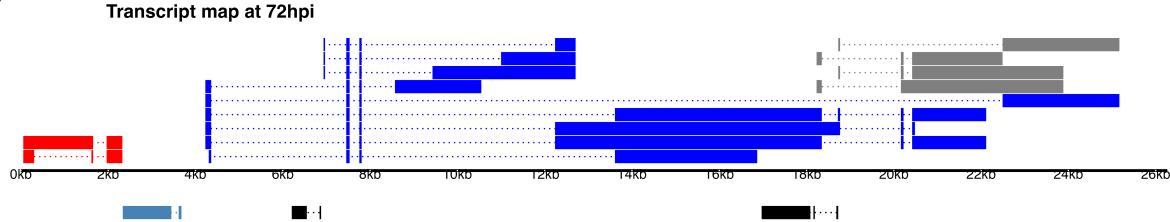
A**1****2****3****4**

Figure 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 transcription unit (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).

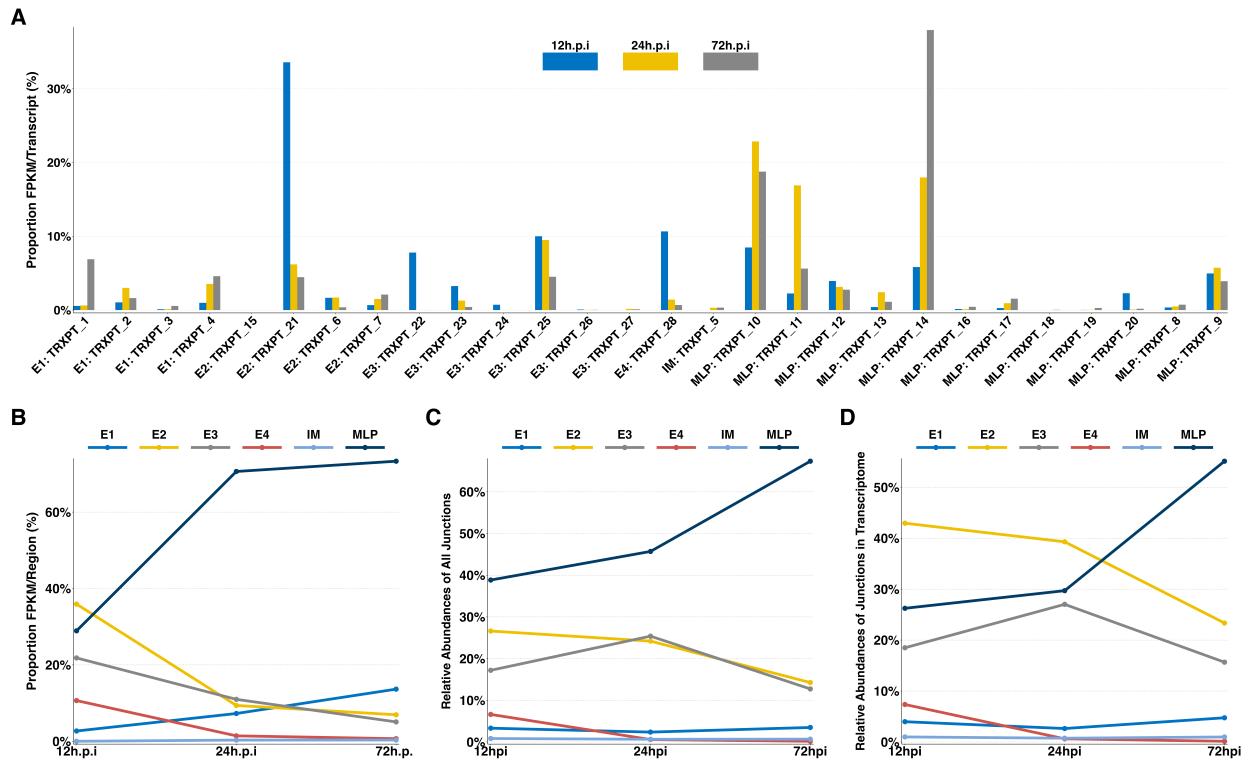


Figure 4: Changes in splicing and expression profile of THEV over time. **A) Normalized (FPKM) expression levels of transcripts over time.** The expression levels (FPKM) of individual transcripts as a percentage of the total expression of all transcripts at each time point are indicated. Only transcripts 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.

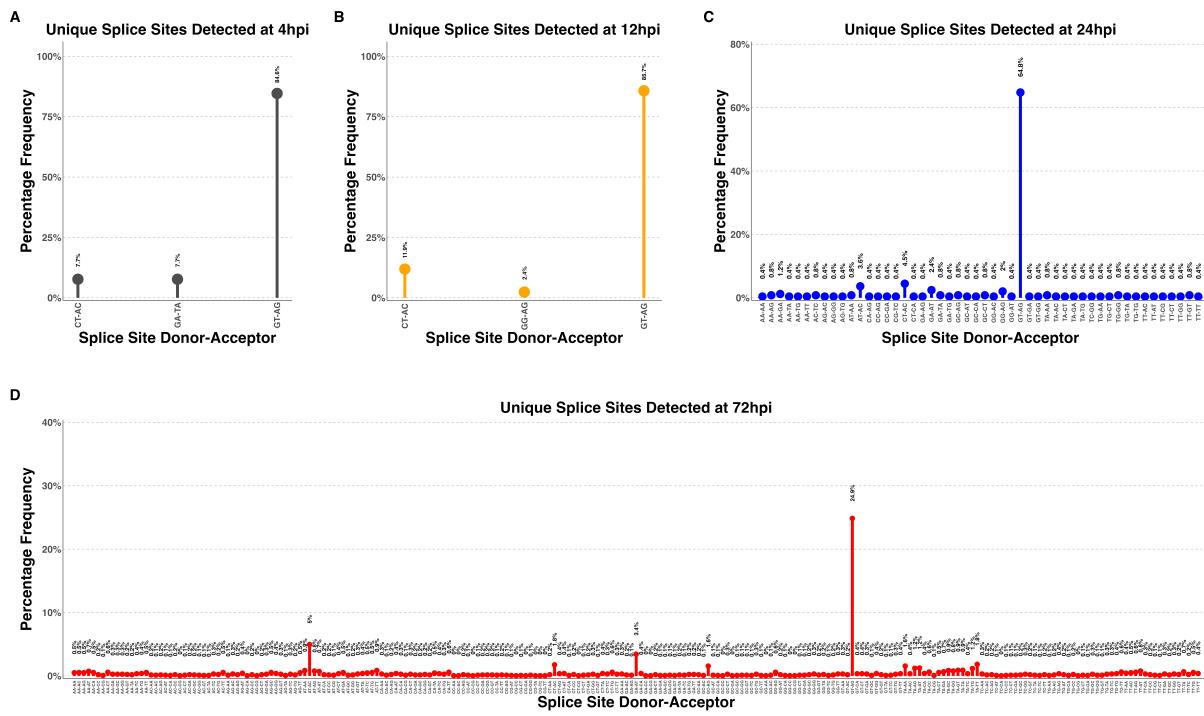
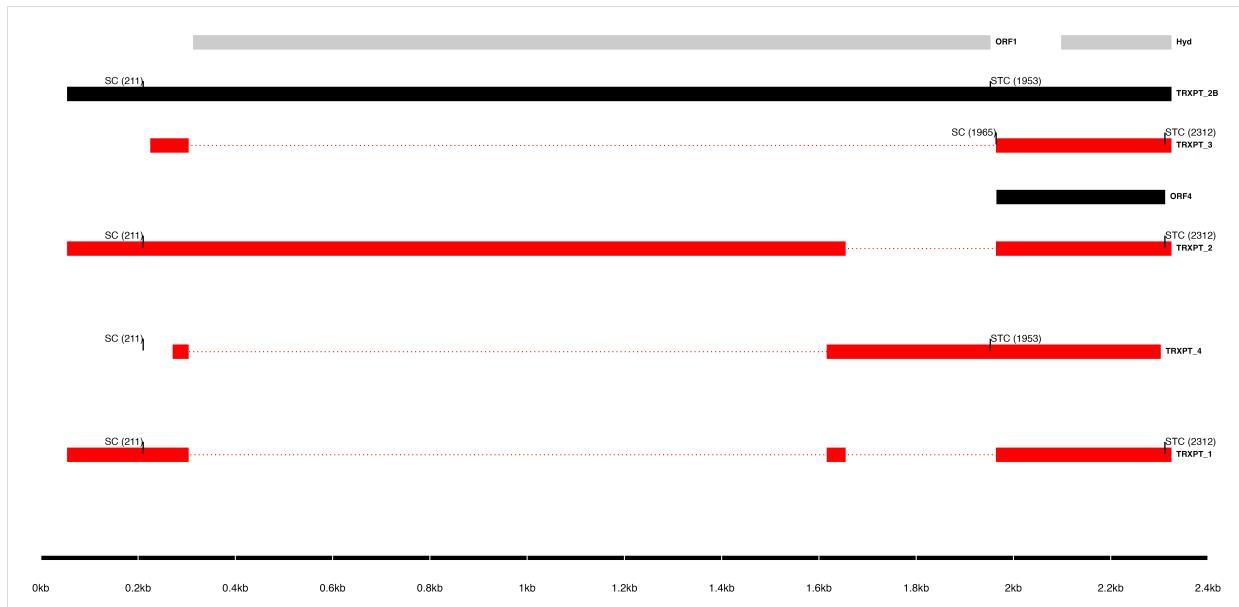


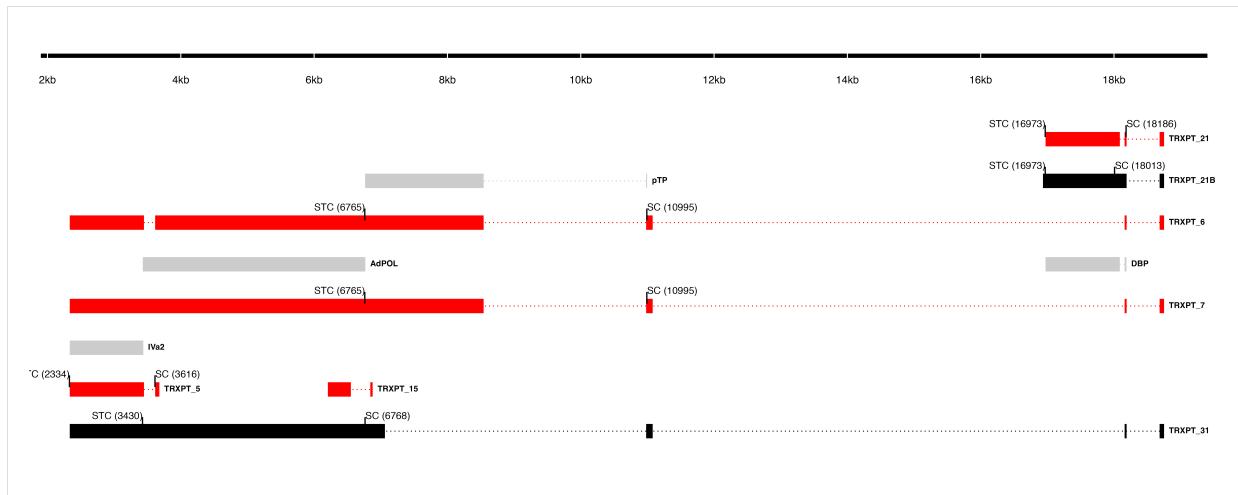
Figure 5: Changes in splice donor-acceptor nucleotides over time. The splice donor-acceptor nucleotides of THEV just like other AdVs is mostly the canonical GU-AG. At early time points (4h.p.i and 12h.p.i **(A)** and **(B)**, respectively) the junction nucleotides used appear to be well scrutinized or restricted, utilizing mostly the canonical splice nucleotides. However, as the infection progresses to the late stages (24h.p.i and 72h.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.



Transcript ID	Splice Junction				Junction Reads					Junction Status
	Start	End	Intron Length	Strand	4h.p.i	12h.p.i	24h.p.i	72h.p.i		
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	

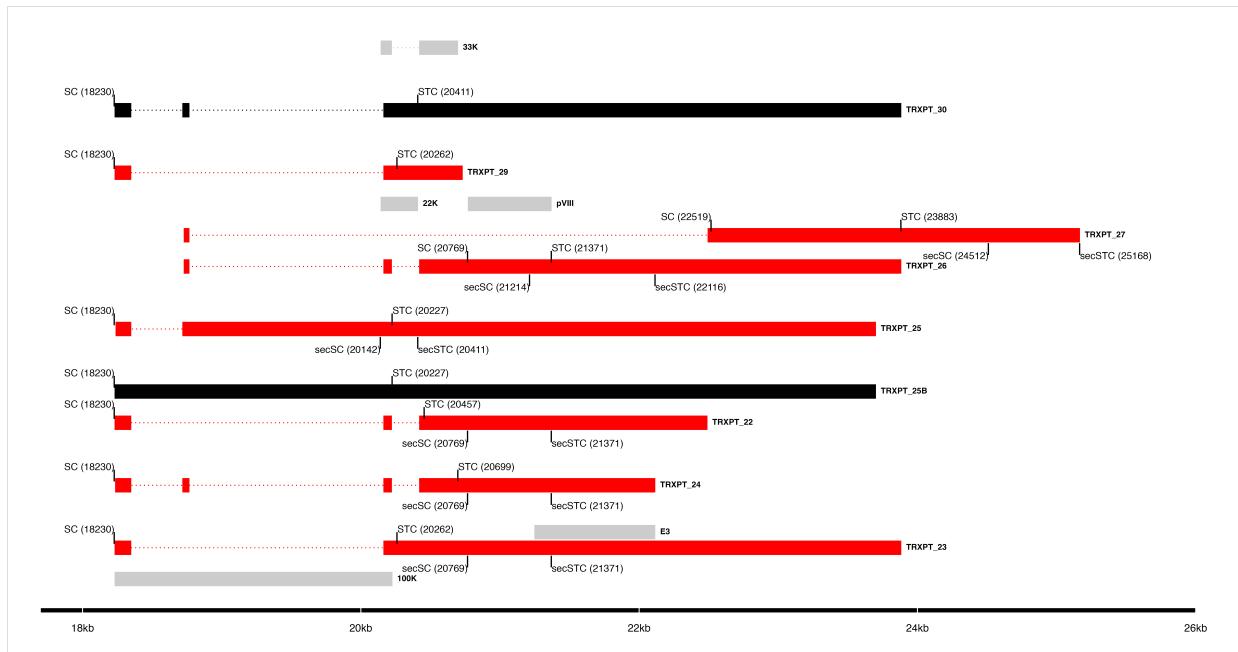
^{*}Not validated for TRXPT_4

Figure 6: The splice map of the E1 transcription unit (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 start codon (SC) and stop codon (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.



Transcript ID	Splice Junction				Junction Reads					Junction Status
	Start	End	Intron Length	Strand	4h.p.i	12h.p.i	24h.p.i	72h.p.i		
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	

Figure 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.



Transcript ID	Splice Junction				Junction Reads					Junction Status
	Start	End	Intron Length	Strand	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	

Figure 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 start codon (SC) and stop codon (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.

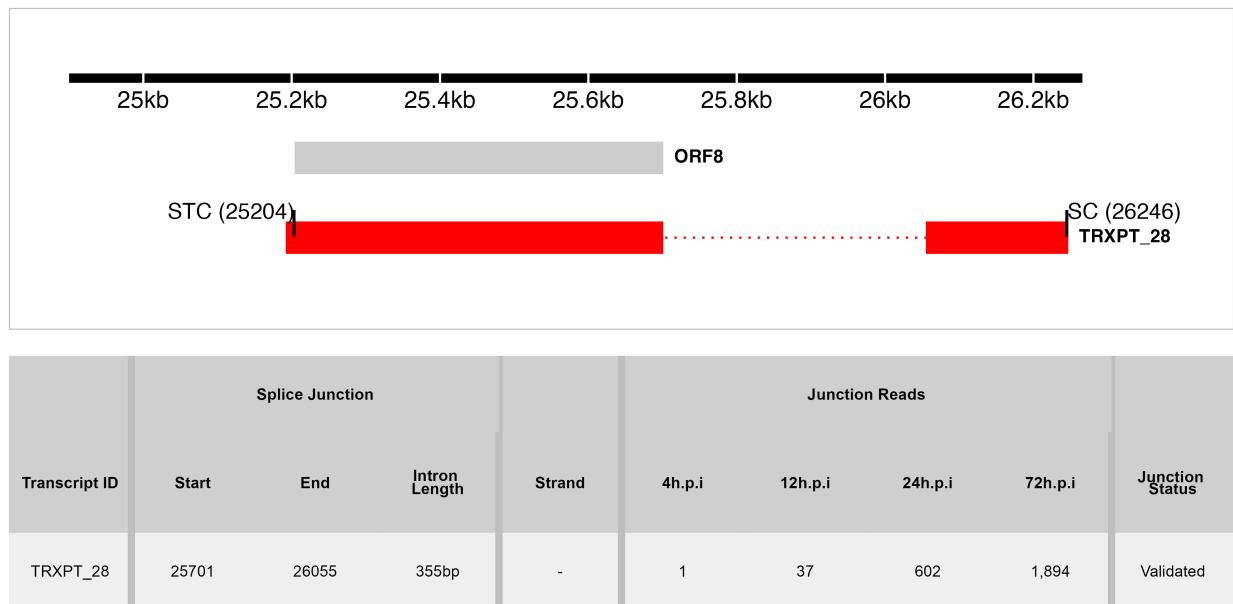
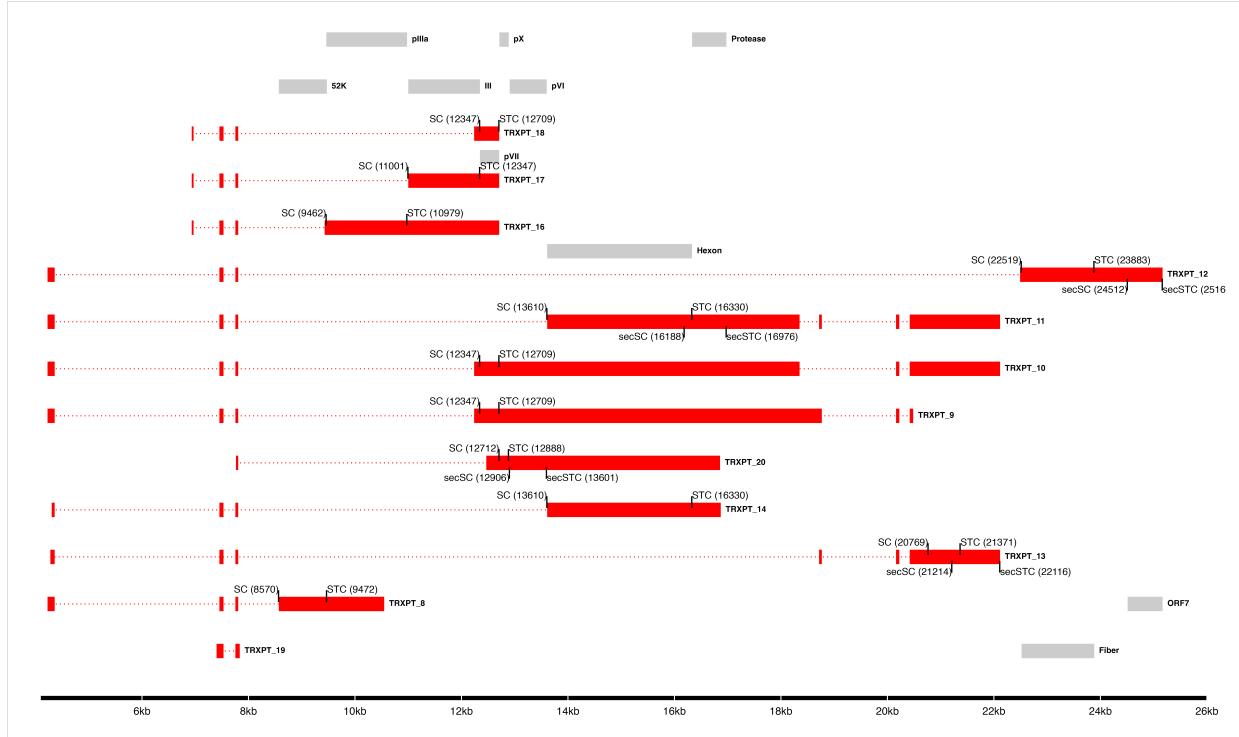


Figure 9: The splice map of the E4 TU. Exons are depicted as boxes connected by introns (dotted lines). The transcript from RNA-seq data is colored red and the predicted ORF, grey. The transcript and ORF are labelled with their names to the right. The start codon (SC) and stop codon (STC) of the 5'-most 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.



Transcript ID	Splice Junctions			Junction Results					Junction Status
	Start	End	Intron Length	Strand	45 bp	12h bp	24h bp	72h bp	
TRXPT_12,TRXPT_13,TRXPT_14,TRXPT_15,TRXPT_16,TRXPT_17,TRXPT_18,TRXPT_19	4303	7654	3265bp	-	1	12	7126	17607	Validated
TRXPT_16,TRXPT_17,TRXPT_18,TRXPT_19,TRXPT_20,TRXPT_21,TRXPT_22,TRXPT_23,TRXPT_24,TRXPT_25,TRXPT_26,TRXPT_27,TRXPT_28,TRXPT_29,TRXPT_30	7651	7764	224bp	-	4	18	12119	32977	Validated
TRXPT_8	7667	8070	3643bp	-	0	5	717	0	Validated
TRXPT_16	7667	11021	3195bp	+	0	1	218	10234	Validated
TRXPT_18,TRXPT_19	7667	12238	4429bp	-	0	7	2446	0	Validated
TRXPT_20	7667	13466	4669bp	-	0	3	457	0	Validated
TRXPT_26,TRXPT_27	7667	13918	5564bp	-	0	18	2935	0	Validated
TRXPT_13	7667	14717	1041bp	-	0	2	1341	0	Validated
TRXPT_22	7667	22002	14595bp	-	1	6	1642	21891	Validated
TRXPT_25,TRXPT_26,TRXPT_19	16350	18117	3667bp	-	4	21	1890	31490	Validated
TRXPT_23,TRXPT_24,TRXPT_11	16350	20162	1615bp	-	3	18	6619	35841	Validated
TRXPT_26,TRXPT_27,TRXPT_23,TRXPT_8,TRXPT_10	16708	20742	13934bp	-	2	21	1027	40962	Validated
TRXPT_11,TRXPT_12,TRXPT_13,TRXPT_14,TRXPT_15,TRXPT_16	20223	25118	4975bp	-	3	33	10283	32238	Validated
TRXPT_16,TRXPT_17,TRXPT_18	6999	7434	455bp	-	0	0	143	15492	Validated
TRXPT_17	7667	9430	1624bp	-	0	0	47	0	Validated

Figure 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 start codon (SC) and stop codon (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.

Table 1: Overview of sequencing results

Metric	4h.p.i	12h.p.i	24h.p.i	72h.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	6.79e+07	1.06e+08	8.38e+07	3.62e+08
Mapped (THEV)	4.32e+02	6.70e+03	1.18e+06	1.69e+07	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	2374	2,457
Junction coverage Total (at least 1 read)	37	605	115075	2132806	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	1791	1,936
Junction coverage (at least 100 reads)	0	1	53	805	859
Junction coverage (at least 1000 reads)	0	0	18	168	186

Table 2a: Most abundant splice junctions at 12h.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 24h.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 72h.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%)

532 **SUPPLEMENTARY MATERIALS**

533 **Supplementary Table S1A**

Table S1a: Most Transcriptionally Active Regions of THEV at 12h.p.i

Time	Region	Strand	Total Reads	Percentage
12hpi	MLP	+	235	38.8%
12hpi	E2	-	161	26.6%
12hpi	E3	+	104	17.2%
12hpi	E4	-	40	6.6%
12hpi	Unassigned	-,+/-	40	6.6%
12hpi	E1	+	20	3.3%
12hpi	IM	-	5	0.8%

534 **Supplementary Table S1B**

Table S1b: Most Transcriptionally Active Regions of THEV at 24h.p.i

Time	Region	Strand	Total Reads	Percentage
24hpi	MLP	+	52,589	45.7%
24hpi	E3	+	29,208	25.4%
24hpi	E2	-	27,833	24.2%
24hpi	E1	+	2,724	2.4%
24hpi	Unassigned	-,+/-	1,313	1.1%
24hpi	IM	-	744	0.6%
24hpi	E4	-	664	0.6%

535 **Supplementary Table S1C**

Table S1c: Most Transcriptionally Active Regions of THEV at 72h.p.i

Time	Region	Strand	Total Reads	Percentage
72hpi	MLP	+	1,436,199	67.3%
72hpi	E2	-	304,191	14.3%
72hpi	E3	+	271,310	12.7%
72hpi	E1	+	74,135	3.5%
72hpi	Unassigned	-,+/-	28,921	1.4%
72hpi	IM	-	14,482	0.7%
72hpi	E4	-	3,568	0.2%

536 **Supplementary Table S2**

Table S2: Mapping Transcript TTS with Closest Polyadenylation Signal

Region	Sub Region	Transcripts	Nearest 'AAUAAA' TTS	
			TTS	signal (location refers to 'U' in the sequence)
E1	E1	TRXPT_1, TRXPT_2, TRXPT_3, TRXPT_4	2,325	2,323
E2	E2A	TRXPT_21, TRXPT_21B	16,934	16,950
E2/IM	E2B/IM	TRXPT_5, TRXPT_6, TRXPT_7, TRXPT_31	2,334	2,333
E3	E3A	TRXPT_22, TRXPT_24	22,491	22,458 (AGUAAA)

Table S2: Mapping Transcript TTS with Closest Polyadenylation Signal

Region	Sub Region	Transcripts	TTS	Nearest 'AAUAAA' TTS signal (location refers to 'U' in the sequence)
E3	E3B	TRXPT_23, TRXPT_25, TRXPT_25B, TRXPT_26, TRXPT_30	23,884	23,889
E3	E3C	TRXPT_27	25,168	25,166
E4	E4	TRXPT_28	25,192	25,203
MLTU	L1	TRXPT_8	10,549	10,537
MLTU	L2	TRXPT_16, TRXPT_17	12,709	-
MLTU	L3	TRXPT_14, TRXPT_20	16,870	16,903
MLTU	L4	TRXPT_9, TRXPT_10, TRXPT_11, TRXPT_13	22,116	22,098 (AAGAAA)
MLTU	L5	TRXPT_12	25,168	25,166

⁵³⁷ **Supplementary Table S3**

Table S3: The Encoded Proteins and Coding Potential of THEV Transcripts

mRNA	Primary Product	Size (kDa)	Number of Residues	Start Codon Position	Stop Codon Position	Downstream Coding Potential
TRXPT_1	ORF9 ^{new}	17.9	160	211	2312	Hyd, ORF4
TRXPT_2	ORF10 ^{new}	66.4	597	211	2312	Hyd, ORF4
TRXPT_2B	eORF1 ^L	64.3	580	1965	1953	Hyd, ORF4
TRXPT_3	ORF4	13.1	115	211	2312	Hyd

Table S3: The Encoded Proteins and Coding Potential of THEV Transcripts

mRNA	Primary Product	Size (kDa)	Number of Residues	Start Codon Position	Stop Codon Position	Downstream Coding Potential
TRXPT_4	ORF11 ^{new}	15.9	143	211	1953	Hyd, ORF4
TRXPT_5	IVa2	42.3	371	3616	2334	-
TRXPT_6	pTP	70.5	597	10995	6765	Ad-pol
TRXPT_7	pTP	70.5	597	10995	6765	Ad-pol
TRXPT_8	52K	33.8	300	8570	9472	-
TRXPT_9	pVII	13.2	120	12347	12709	pX, pVI, hexon, eProt, ORF12, pVIII, eE3
TRXPT_10	pVII	13.2	120	12347	12709	pX, pVI, hexon, eProt, ORF12, pVIII, eE3
TRXPT_11	Hexon (II)	101.1	906	13610	16330	eProt, ORF12, e33K, pVIII, eE3
TRXPT_12	Fiber (IV)	49	454	22519	23883	ORF7
TRXPT_13	pVIII	21.8	200	20769	21371	eE3
TRXPT_14	Hexon (II)	101.1	906	13610	16330	-
TRXPT_15	-	-	-	-	-	-
TRXPT_16	pIIIa	57.6	505	9462	10979	Penton (III), pVII
TRXPT_17	Penton (III)	50.8	448	11001	12347	-
TRXPT_18	pVII	13.2	120	12347	12709	-
TRXPT_19	-	-	-	-	-	-
TRXPT_20	pX	6.1	58	12712	12888	pVI, hexon

Table S3: The Encoded Proteins and Coding Potential of THEV Transcripts

mRNA	Primary Product	Size (kDa)	Number of Residues	Start Codon Position	Stop Codon Position	Downstream Coding Potential
TRXPT_21	DBP	43.3	380	18186	16973	-
TRXPT_21B	tDBP ^S	39.3	346	18013	16973	-
TRXPT_22	8.3KII ^{new}	8.6	73	18230	20457	pVIII, eE3
TRXPT_23	8.3KI ^{new}	8.3	73	18230	20262	pVIII, eE3, Fiber
TRXPT_24	e33K ^L	19.8	171	18230	20699	pVIII, E3
TRXPT_25	t100K	62.4	543	18230	20227	22K, pVIII, eE3
TRXPT_25B	100K	76.5	665	18230	20227	22K, pVIII, eE3
TRXPT_26	pVIII	21.8	200	20769	21371	eE3, Fiber
TRXPT_27	Fiber (IV)	49	454	22519	23883	ORF7
TRXPT_28	eORF8 ^L	26.4	229	26246	25204	-
TRXPT_29	8.3KI	8.3	73	18230	20262	-
TRXPT_30	e22K ^L	15.7	140	18230	20411	pVIII, eE3, Fiber
TRXPT_31	Ad-pol	129.2	1112	6768	3430	-

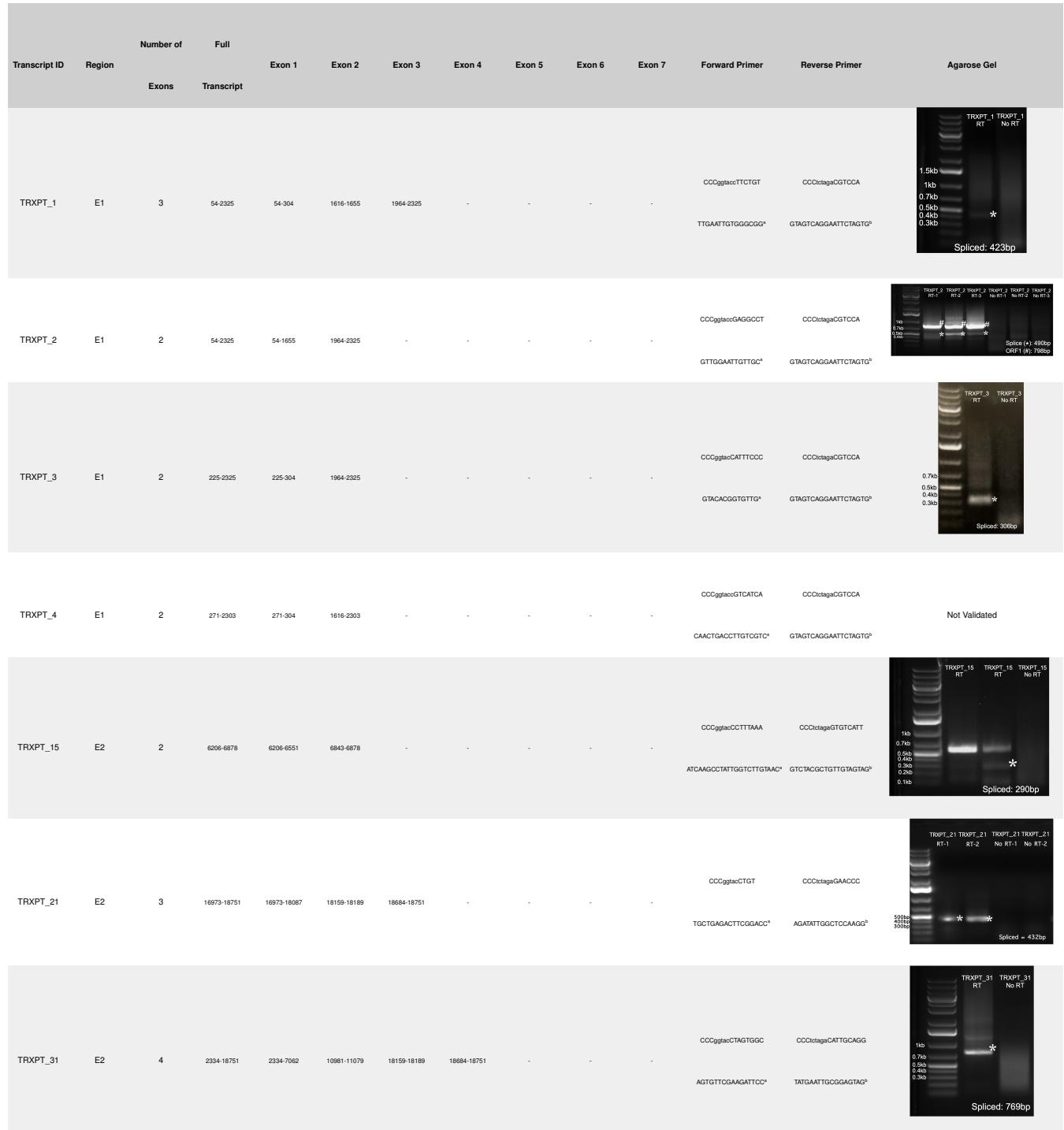
^{new}Novel product identified in this work;

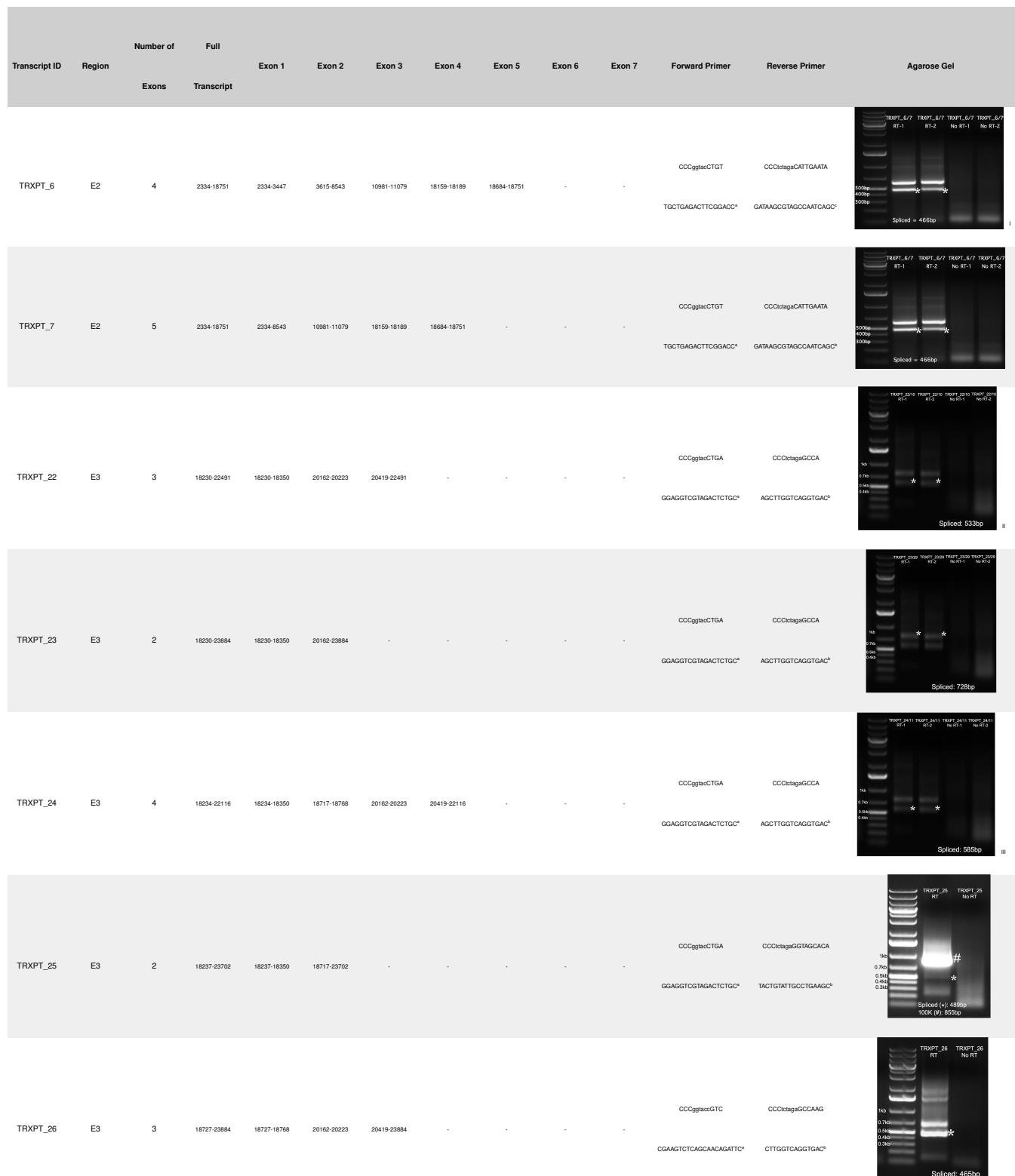
^LLonger isoform identified in this work

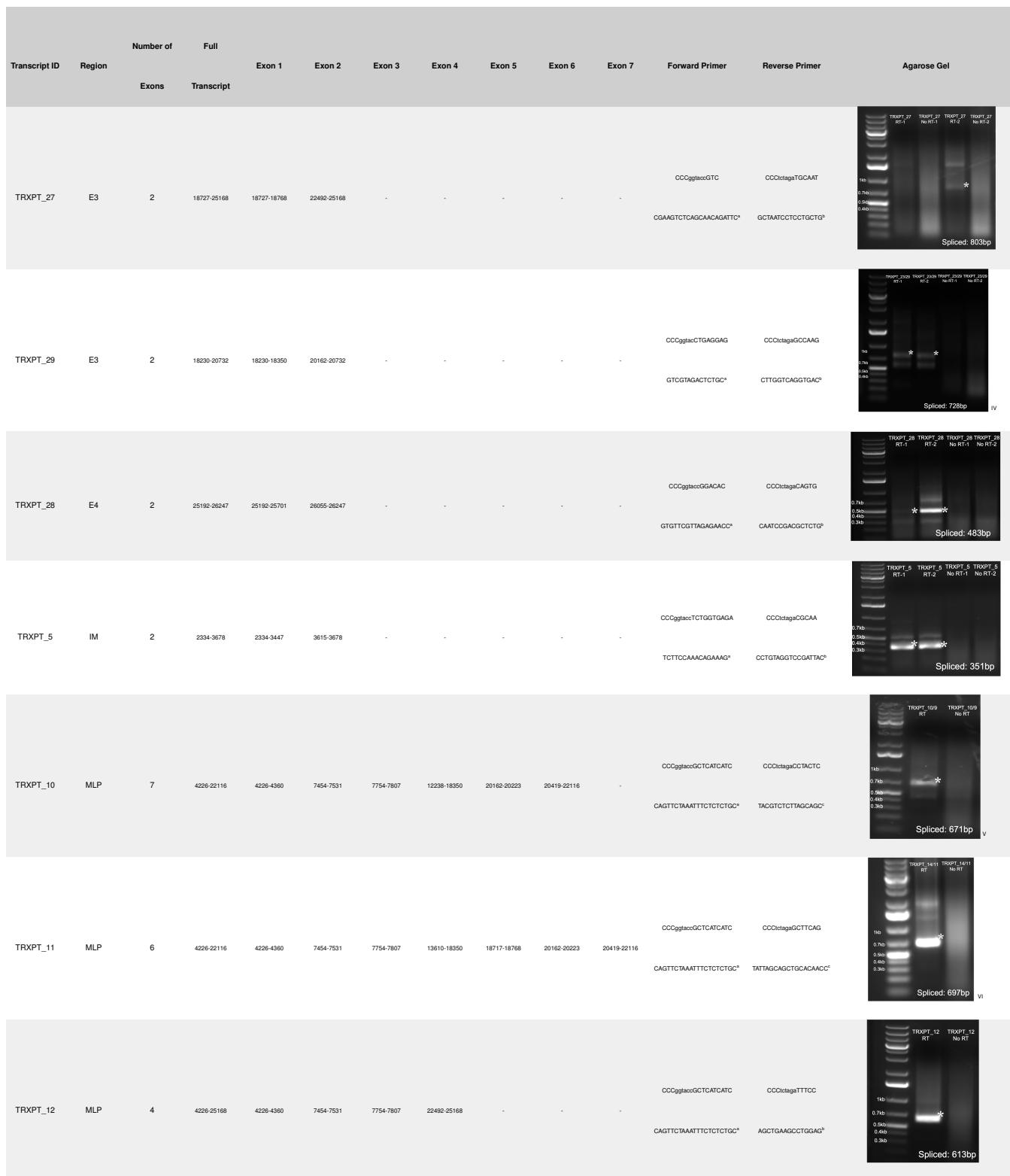
^SShorter isoform identified in this work

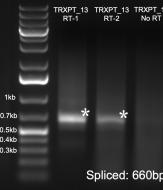
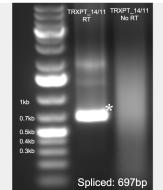
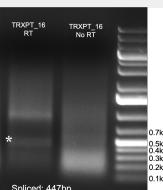
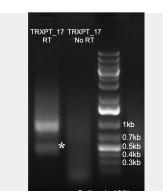
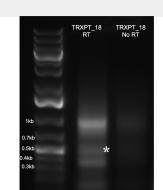
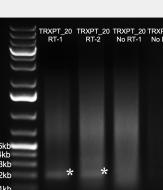
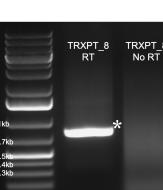
538 **Supplementary PCR Methods**

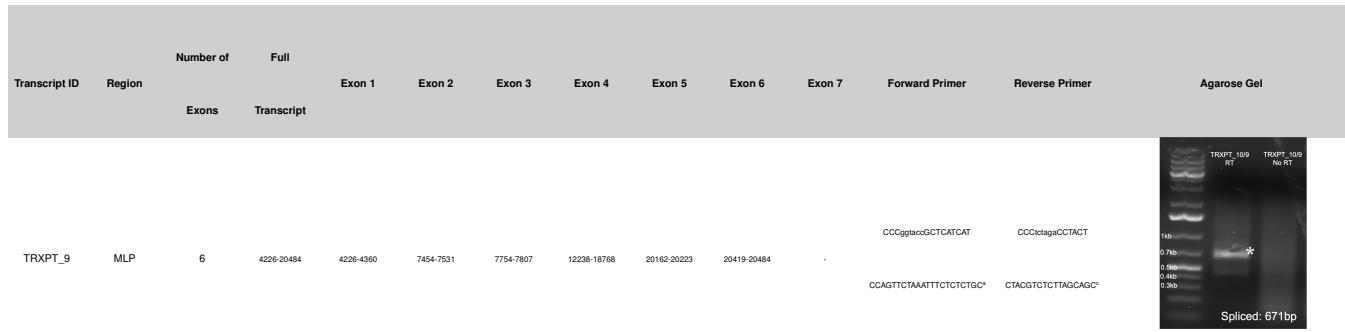
Agarose Gels Showing PCR Amplification of THEV cDNA With Gene-Specific Primers







Transcript ID	Region	Number of Exons	Full Transcript							Forward Primer	Reverse Primer	Agarose Gel
			Exon 1	Exon 2	Exon 3	Exon 4	Exon 5	Exon 6	Exon 7			
TRXPT_13	MLP	6	4279-22116	4279-4360	7454-7531	7754-7807	18717-18768	20162-20223	20419-22116	CCGgttaccGCTCATCATC	CCCttagaGCCAAG	
TRXPT_14	MLP	4	4304-16870	4304-4360	7454-7531	7754-7807	13610-16870	-	-	CCGgttaccGCTCATCATC	CCCttagaGCTTCAGT	
TRXPT_16	MLP	4	6934-12709	6934-6969	7454-7531	7754-7807	9430-12709	-	-	CCGgttaccGGATCTC	CCCttagaGCCT	
TRXPT_17	MLP	4	6934-12709	6934-6969	7454-7531	7754-7807	11001-12709	-	-	CCGgttaccGGATCTC	CTCCCCATCTAGAC	
TRXPT_18	MLP	4	6934-12709	6934-6969	7454-7531	7754-7807	12238-12709	-	-	CCGgttaccGGATCTC	CCCttagaGTTCTC	
TRXPT_19	MLP	2	7401-7836	7401-7531	7754-7836	-	-	-	-	-	-	N/A
TRXPT_20	MLP	2	7765-16856	7765-7807	12466-16856	-	-	-	-	CCGgttaccGAGGATTGA	CCCttagaCTGAA	
TRXPT_8	MLP	4	4226-10549	4226-4360	7454-7531	7754-7807	8570-10549	-	-	CCGgttaccGCTCATCAT	CCCttagaCCTATC	



^aPrimer binds inside first exon; ^bPrimer binds inside terminal exon; ^cPrimer binds inside fourth exon; ^dAgarose gel identical to TRXPT_7 due to identical splicing; ^eAgarose gel identical to last 3 exons of TRXPT_10 due to identical splicing; ^fAgarose gel identical to last 4 exons of TRXPT_11 due to identical splicing; ^gAgarose gel identical to TRXPT_23 due to identical splicing; ^hAgarose gel identical to TRXPT_9 due to identical splicing; ⁱAgarose gel identical to TRXPT_14 due to identical splicing;

539 In the table above, the restriction sites in the primer tails are shown in lowercase letters. All the primer melting
 540 temperatures (TMs) are 58-60°C using a hot start Taq DNA polymerase. The PCR reaction mix was made
 541 per manufacturer's instructions. The PCR cycling conditions were as follows: Initial denaturation – 95°C for
 542 1 minute; cyclical denaturation – 95°C for 30 seconds, annealing – variable temperature (53°C-56°C) for
 543 30 seconds, primer extension – 68°C for variable time, and final elongation – 68°C for 5 minutes. We used
 544 35 cycles of amplification. The PCR products of THEV cDNA shown in the gels in the table above were
 545 cloned and Sanger sequenced to confirm the splice junctions. We included "No reverse transcriptase (No
 546 RT)" template controls to ensure the PCR products are of RNA origin.

547 **Supplementary Computational Analysis**

548 Snakemake v7.24.0 was used to manage our entire workflow. A graph of the main steps in our pipeline
 549 generated with Snakemake is shown below. Our trimmed RNA-seq reads were mapped to the genome of
 550 *M. gallopano* (with the THEV genome as one of its chromosomes) using Hisat2, to generate the alignment
 551 (BAM) files and StringTie used to assemble the transcriptome with a GTF annotation file containing the
 552 predicted THEV ORFs as a guide. The GTF annotation file was derived from a GFF3 annotation file obtained
 553 from NCBI using Agat - version 1.0.0, a program for converting between many different file formats used in
 554 bioinformatics. However, the NCBI GFF3 annotation file itself was first modified to remove all unimportant
 555 features, leaving only the ORFs.

556 StringTie was also used to estimate the normalized expression levels (FPKM) of all the transcripts and
 557 Ballgown in R was used to perform statistical analysis and comparisons of the transcript expression levels,
 558 which was instructive in understanding the temporal regulation THEV gene expression.

559 In these steps above, each sample (replicate of each time point) was processed independently and merged
 560 only in the final transcriptome assembly or during analysis with Ballgown. In the subsequent steps described

561 below, all samples for each time point were processed together.

562 We used RegTools to extract and analyze the splice junctions in the BAM files. The command “regtools
563 junctions extract” provides a wealth of information about all the splice sites in the BAM file provided such as:
564 the start and end positions, the strand, and number of reads supporting the splice junctions. The command
565 “regtools junctions annotate” gives even more information such as: the splice site donor-acceptor sequences
566 and transcripts/genes that overlap the junction. This information was the basis for estimating and comparing
567 the splicing activity of different regions (TUs) of THEV over time. Also, Samtools was also used to count the
568 total sequencing reads for all replicates at each time point.



A flowchart of the major steps in the computational analysis pipeline (*generated with Snakemake*)