

# Characterizing the Transcriptome of Turkey Hemorrhagic Enteritis Virus

3

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

<sup>5</sup> Abraham Quaye<sup>1\*</sup>, Brett Pickett<sup>\*</sup>, Joel S. Griffitts<sup>\*</sup>, Bradford K. Berges<sup>\*</sup>, Brian D. Poole<sup>†\*</sup>

<sup>6</sup>\*Department of Microbiology and Molecular Biology, Brigham Young University

7 1 First-author

<sup>8</sup> † Corresponding Author

## **9 Corresponding Author Information**

<sup>10</sup> brian\_poole@byu.edu

<sup>11</sup> Department of Microbiology and Molecular Biology,

<sup>12</sup> 4007 Life Sciences Building (LSB),

<sup>13</sup> Brigham Young University,

**14 Provo, Utah**

15

16 **ABSTRACT**

17 We have performed an RNA-sequencing experiment characterizing the transcriptome of turkey hemorrhagic  
18 enteritis virus (THEV) for the first time, yielding the only insight into THEV's gene expression patterns.  
19 Previously, THEV's genome had been predicted to encode 23 open reading frames (ORFs). In this work  
20 we identified 29 transcripts from our RNA-seq data all of which consisted of novel exons albeit some exons  
21 matched the predicted ORFs. The three predicted splice junctions were also corroborated by our data. We  
22 performed PCR amplification of THEV cDNA and cloned the PCR products, and Sanger sequencing was  
23 used to validate all identified splice junctions. During validation, we identified 5 additional transcripts some  
24 of which were further validated by 3'RACE data. Thus, the transcriptome of THEV consists of 34 unique  
25 transcripts with the coding capacity for all predicted ORFs. However, we found 6 predicted ORFs (ORF1, E3,  
26 33K, ORF8, IVa2, and protease) to be truncated predictions as either an in-frame upstream start codon was  
27 identified or additional coding exons were found. We also identified 3 predicted ORFs with longer or shorter  
28 isoforms, and 7 novel unpredicted ORFs that could be encoded by some transcripts; albeit it is beyond  
29 the scope of this manuscript to investigate whether they get translated. Similar to other adenoviruses,  
30 THEV also produce multiple distinctly spliced transcripts that code for the same protein across its genome.  
31 Also, our data shows that all THEV transcripts are spliced, and organized in five transcription units under  
32 the control of their cognate promoter. However, our data suggests that THEV's temporal regulation may  
33 be different from other adenoviruses. Over 2,300 unique splice junctions were found across the genome,  
34 mostly at low levels. This low-level use of broad alternative splicing patterns is thought to enable the virus  
35 to maximize its coding potential in an evolving environment.

36 **INTRODUCTION**

37 Adenoviruses (AdVs) are non-enveloped icosahedral-shaped DNA viruses, causing infection in virtually all  
38 vertebrates. Their double-stranded linear DNA genomes range between 26 and 45kb in size, producing a  
39 broad repertoire of transcripts via highly complex alternative splicing patterns (1, 2). The AdV genome is  
40 one of the most optimally economized; both the forward and reverse DNA strands harbor protein-coding  
41 genes, making it highly gene-dense. There are 16 genes termed “genus-common” that are homologous in  
42 all AdVs; these are thought to be inherited from a common ancestor. All other genes are termed “genus-  
43 specific”. “Genus-specific” genes tend to be located at the termini of the genome while “genus-common”  
44 genes are usually central (1). This pattern is observed in *Adenoviridae*, *Poxviridae*, and *Herpesviridae* (1,  
45 3, 4). The family *Adenoviridae* consists of five genera: *Mastadenovirus* (MAdV), *Aviadenovirus*, *Ataden-  
46 ovirus*, *Ichtadenovirus*, and *Siadenovirus* (SiAdV) (5, 6). Currently, there are three recognized members  
47 of the genus SiAdV: frog adenovirus 1, raptor adenovirus 1, and turkey adenovirus 3 also called turkey  
48 hemorrhagic enteritis virus (THEV) (5, 7–10). Members of SiAdV have the smallest genome size (~26 kb)  
49 and gene content (~23 genes) of all known AdVs, and many “genus-specific” putative genes of unknown  
50 functions have been annotated (see **Figure 1**) (1, 2, 7).

51 Virulent THEV strains (THEV-V) and avirulent strains (THEV-A) of THEV are serologically indistinguishable,  
52 infecting turkeys, chickens, and pheasants, with the THEV-V causing different clinical diseases in these  
53 birds (2, 11). In turkeys, the THEV-V cause hemorrhagic enteritis (HE), a debilitating acute disease affect-  
54 ing predominantly 6-12-week-old turkeys characterized by immunosuppression (IS), weight loss, intestinal  
55 lesions leading to bloody diarrhea, splenomegaly, and up to 80% mortality (11–13). HE is the most econom-  
56 ically significant disease caused by any strain of THEV (11). While the current vaccine strain (a THEV-A  
57 isolated from a pheasant, Virginia Avirulent Strain [VAS]) has proven effective at preventing HE in young  
58 turkey pouls, it still retains the immunosuppressive ability. Thus, vaccinated birds are rendered more sus-  
59 ceptible to opportunistic infections and death than unvaccinated cohorts leading to substantial economic  
60 losses (11, 14–16). To eliminate this immunosuppressive side-effect of the vaccine, a thorough investiga-  
61 tion of the culprit viral factors (genes) mediating this phenomenon is essential. However, the transcriptome  
62 (splicing and gene expression patterns) of THEV has not been characterized, making the investigation of  
63 specific viral genes for possible roles in causing IS impractical. A well-characterized transcriptome of THEV  
64 is required to enable experimentation with specific viral genes that may mediate IS.

65 Myriads of studies have elucidated the AdV transcriptome in fine detail (17, 18). However, a large pre-  
66 ponderance of studies focus on MAdVs – specifically human AdVs. Thus, most of the current knowledge

67 regarding AdV gene expression and replication is based on MAdV studies, which is generalized for all other  
68 AdVs (6, 19). MAdV genes are transcribed in a temporal manner; therefore, genes are categorized into five  
69 early transcription units (E1A, E1B, E2, E3, and E4), two intermediate (IM) units (pIX and IVa2), and one  
70 major late unit (MLTU or major late promoter [MLP] region), which generates five families of late mRNAs  
71 (L1-L5) based on the polyadenylation site. An additional gene (UXP or U exon) is located on the reverse  
72 strand. The early genes encode non-structural proteins such as enzymes or host cell modulating proteins,  
73 primarily involved in DNA replication, or providing the necessary intracellular niche for optimal replication  
74 while late genes encode structural proteins that act as capsid proteins, promote virion assembly, and direct  
75 genome packaging. The immediate early gene E1A is expressed first, followed by the delayed early  
76 genes, E1B, E2, E3 and E4. Then the intermediate early genes, IVa2 and pIX are expressed followed by  
77 the late genes (6, 17, 18). Noteworthily, the MLP shows basal transcriptional activity during early infection  
78 (before DNA replication), with a comparable efficiency to other early viral promoters, but reaches its max-  
79 imal activity during late infection (after DNA replication). However, during early infection the repertoire of  
80 late transcripts from the MLP is restricted until late infection (6). MAdV makes an extensive use of alterna-  
81 tive RNA splicing to produce a very complex array of mRNAs. All but the pIX mRNA undergo at least one  
82 splicing event. For instance, the MLTU produces over 20 distinct splice variants all of which contain three  
83 non-coding exons at the 5'-end (collectively known as the tripartite leader, TPL) (17, 18). There is also  
84 an alternate 5' three non-coding exons present in varying amounts on a subset of MLTU mRNAs (known  
85 as the x-, y- and z-leaders). Lastly, there is the i-leader exon, which is infrequently included between the  
86 second and third TPL exons, and codes for the i-leader protein (20). Thus, the MLTU produces a complex  
87 repertoire of mRNA with diverse 5' untranslated regions (UTRs) spliced onto different 3' coding exons which  
88 are grouped into five different 3'-end classes (L1-L5) based on polyadenylation site. Each transcription unit  
89 (TU) contains its own promoter driving the expression of all the array of mRNA transcripts produced via  
90 alternative splicing in the unit (6, 17, 18). The promoters are activated at different phases of the infection by  
91 proteins from previously activated TUs. Paradoxically, the early-to-late phase transition during infection re-  
92 quires the L4 genes, 22K and 33K, which should only be available after the transition. However, a promoter  
93 in the L4 region (L4P) that directs the expression of these two proteins independent of the MLP was found,  
94 resolving the paradox (6, 17, 21). During translation of AdV mRNA, recent studies strongly suggest the  
95 potential usage of secondary start codons; adding to what was already a highly complex system for gene  
96 expression (17, 22).

97 High throughput sequencing methods have facilitated the discovery of many novel transcribed regions and  
98 splicing isoforms. It is also a very powerful tool to study alternative splicing under different conditions at an

99 unparalleled depth [(23); (18); Westergren2021]. In this paper, a paired-end deep sequencing experiment  
100 was performed to characterize for the first time the transcriptome of THEV (VAS vaccine strain) during  
101 different phases of the infection, yielding the first THEV splicing map. Our paired-end sequencing allowed  
102 for reading **149** bp long high quality (mean Phred Score of 36) sequences from each end of cDNA fragments,  
103 which were mapped to the genome of THEV.

104 **RESULTS**

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

106 A previous study by Aboeza *et al* showed that almost all THEV transcripts were detectable beginning at  
107 4 hours (24). Therefore, infected MDTC-RP19 cells were harvested at 4-, 12-, 24-, and 72-hours post-  
108 infection (h.p.i) to ensure an amply wide time window to sample all transcripts. Our paired-end RNA se-  
109 quencing (RNA-seq) experiment yielded an average of **107.1** million total reads of **149bp** in length per  
110 time-point, which were simultaneously mapped to both the virus (THEV) and host (*Meleagris gallopavo*)  
111 genomes using the Hisat2 (25) alignment program. A total of **18.1** million reads from all time-points mapped  
112 to the virus genome; this provided good coverage/depth, leaving no regions unmapped. The mapped reads  
113 to the virus genome increased substantially from **432** reads at 4 h.p.i to **16.9** million reads at 72 h.p.i (**Table**  
114 **1**, **Figure 2a**). From the mapped reads, we identified a total of **2,457** unique THEV splice junctions from all  
115 time-points, with splice junctions from the later time-points being supported by significantly more sequence  
116 reads than earlier time-points. For example, all the **13** unique junctions at 4 h.p.i had less than 10 reads  
117 supporting each one, averaging a mere **2.8** reads/junction. Conversely, the **2374** unique junctions at 72 h.p.i  
118 averaged **898.4** reads/junction, some junctions having coverage as high as **322,677** reads. The substantial  
119 increases in splice junction and mapping reads to the THEV genome over time denotes an active infection  
120 and correlates with our quantitative PCR (qPCR) assay quantifying the total number of viral genome copies  
121 over time (**Figure 2b**).

122 Using StringTie (25), an assembler of RNA-seq alignments into potential transcripts, the mapped reads for  
123 each time point were assembled into transcripts using the genomic location of the predicted THEV ORFs as  
124 a guide. In the consolidated transcriptome, a composite of all unredudant transcripts from all time points,  
125 we counted a total of **29** novel transcripts. Although some exons in some transcripts match the predicted  
126 ORFs exactly, most of our identified exons are longer, spanning multiple predicted ORFs (**Figure 3**).

127 We validated the splice junctions in all transcripts by PCR amplification of viral cDNA, cloning, and Sanger  
128 sequencing (**Supplementary PCR methods**). During validation, we identified 5 additional transcripts some  
129 of which were further validated by 3' Rapid Amplification of cDNA Ends (3'RACE) data. The complete  
130 list of unique splice junctions mapped to THEV's genome has been submitted to the National Center for  
131 Biotechnology Information Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo>) under accession  
132 number GSE254416.

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

134 AdV gene expression occurs under exquisite temporal control with each promoter typically producing one or

135 few pre-mRNAs that undergo alternative splicing to yield the manifold repertoire of complex transcripts. To  
136 evaluate the activity of each promoter over time, *StringTie* and *Ballgown* (a program for statistical analysis  
137 of assembled transcriptomes) (26) were used to estimate the normalized expression levels of all transcripts  
138 for each time point in Fragments Per Kilobase of transcript per Million mapped reads (FPKM) units. Very few  
139 unique splice junctions, reads, and transcripts were counted at 4 h.p.i; hence, this time point was excluded  
140 in this analysis.

141 Considering individual mRNAs, TRXPT\_21 – from the E2 region – was the most significantly expressed  
142 at 12 h.p.i, constituting about **33.58%** of the total expression of all transcripts. Transcripts in the E3 and  
143 E4 regions also contributed significant proportions, and noticeably, some MLP region transcripts. The later  
144 time points were dominated by the MLP region transcripts – TRXPT\_10 and TRXPT\_14 were the most  
145 abundantly expressed at 24 and 72 h.p.i, respectively, as expected (**Figure 4a**). When we performed  
146 analysis of the FPKM values of transcripts per region we found a similar pattern: the E2 region was the  
147 most abundantly expressed at 12 h.p.i, after which the MLP region assumes predominance (**Figure 4b**).  
148 Secondly, we estimated relative abundances of all splice junctions at each time point using the raw reads.  
149 For individual junctions, we counted as significantly expressed only junctions with coverage of at least 1%  
150 of the total splice junction reads at the given time point. At 12 h.p.i, **18** junctions meet the 1% threshold, and  
151 were comprised of predominantly early region (E1, E2, E3, and E4) junctions, albeit the MLTU was the single  
152 most preponderant region overall, constituting **38.8%** of all the junction reads (**Table 2a** and **Supplementary**  
153 **Table S1a**). The topmost abundant junctions at 12 h.p.i remained the most significantly expressed at 24  
154 h.p.i also. However, here, the MLP-derived junctions were unsurprisingly even more preponderant overall,  
155 accounting for **45.7%** of all the junction reads counted (**Table 2b** and **Supplementary Table S1b**). At 72  
156 h.p.i, the trend of increased activity of the MLP continued as expected; at this time, the MLP region junctions  
157 were not only the most abundant overall – accounting for **67.4%** of all junction reads, – but also contained  
158 the most significantly expressed individual junctions (**Table 2c**, **Supplementary Table S1c** and **Figure 4c**).  
159 When we limited this analysis to only junctions in the final transcriptome, the relative abundances of the  
160 junctions for each region over time was generally similar to the pattern seen with all the junctions included  
161 (**Figure 4d**).

162 We also analyzed splice donor and acceptor site nucleotide usage over time to investigate any peculiarities  
163 that THEV may show, generally or over the course of the infection. We found that most splice donor-  
164 acceptor sequences were unsurprisingly the canonical GT-AG nucleotides. However, the splice acceptor-  
165 donor pairing became less specific over time, such that all combinations of nucleotide pairs were eventually  
166 detected (**Figure 5**)

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

168 This region in MAdVs is the first transcribed after successful entry of the viral DNA into the host cell nucleus,  
169 albeit at low levels (18). The host transcription machinery solely mediates the transcription of this region.  
170 After their translation, the E1 proteins in concert with a myriad of host transcription factors activate the other  
171 viral promoters (6). In MAdVs, this region is subdivided into E1a and E2b units, but the transcripts found in  
172 our data categorized under this region do not appear to so divided.

173 Only two ORFs (ORF1 [sialidase] and Hyd) are predicted in this region; however, we discovered **four** novel  
174 transcripts in this region, which collectively contain **3** unique splice junctions (**Figure 6**). Most of the ORFs  
175 of the novel transcripts are distinct from the predicted ORFs, but they all have the coding potential (CP)  
176 for the predicted Hyd protein as the 3'-most coding sequence (CDS) if secondary start codon usage is  
177 considered as reported for other AdVs (17, 18). The 5'-most CDS of TRXPT\_1 is multi-exonic, encoding  
178 a novel 17.9 kilodalton (kDa), 160 residue [amino acids (aa)] protein (ORF9). From its 5'-most start codon  
179 (SSC), TRXPT\_2 encodes the largest protein in this region – a 64.3 kDa, 580 aa protein (ORF10) with the  
180 same SSC as ORF9 (position 211bp). ORF10 spans almost the entire predicted ORF1 and Hyd, coming  
181 short in two regards: it is spliced from 1655bp to 1964bp (ORF1's C-terminus, including the stop codon), and  
182 it's stop codon (STC; position 2312) is 13 bp short of Hyd's STC. However, it has an SSC 102 bp upstream  
183 and in-frame with ORF1's predicted SSC. Thus, ORF10 shares substantial protein sequence similarity with  
184 ORF1 but not with Hyd, as the SSC of Hyd is not in-frame. Without its splice site removing the ORF1 STC,  
185 TRXPT\_2 would encode a longer variant of ORF1, starting from an upstream SSC. TRXPT\_3 is almost  
186 identical to TRXPT\_1, except for the lack of TRXPT\_1's second exon. Our RNA-seq data shows that all E1  
187 transcripts share the same transcription termination site (TTS; at position 2325bp). However, TRXPT\_3 and  
188 TRXPT\_4 seem to have transcription start sites (TSS) downstream of the TSS of TRXPT\_1 and TRXPT\_2  
189 (E1 TSS; position: 54bp). Given that studies in MAdVs show that E1 mRNAs share not only a common  
190 TTS but also the TSS, and only differ from each other regarding the internal splicing (18), it is likely that  
191 TRXPT\_3 and TRXPT\_4 are incomplete, and their actual TSS just like the TTS are identical for all E1  
192 transcripts. Regardless of the TSS considered for TRXPT\_3, the coding potential (CP) remains unaffected.  
193 Its 5'-most CDS, beginning at 1965bp and sharing the same STC as ORF9, produces a 13.1 kDa, 115  
194 residue protein (ORF4). ORF4 was predicted in an earlier study (27) but was excluded in later studies (1,  
195 12); however, our data suggests it is a bona fide ORF. Unlike TRXPT\_3, the CP of TRXPT\_4 is affected by  
196 the TSS considered; if we consider its unmodified TSS, then its CP is the same as TRXPT\_3 (ORF4 as the  
197 first CDS and Hyd as second CDS using the secondary SSC). However, if we assume that TRXPT\_4 uses  
198 the E1 TSS, then the 5'-most CDS is a distinct, novel, multi-exonic 15.9 kDa, 143 aa protein (ORF11) with

199 the same SSC as ORF9 and ORF10 but with a unique STC. The splice junctions of all transcripts in this  
200 region (except the junction for TRXPT\_4) were validated by cloning of viral cDNA and Sanger sequencing  
201 (**Supplementary PCR methods**).

202 During the validation of TRXPT\_2, ORF1 was present on the agarose gel (an unspliced band size) and  
203 Sanger sequencing results as a bona fide transcript (**Supplementary PCR methods**). This was corroborated  
204 by our 3'RACE experiment, which showed a transcript (TRXPT\_2B) spanning the entire ORF1 and  
205 Hyd ORFs without any splicing, with a poly-A tail immediately after the E1 TTS. The 5'-most CDS of this  
206 transcript (TRXPT\_2B) would encode ORF1. However, TRXPT\_2B has an upstream and in-frame SSC  
207 to the predicted SSC of ORF1, suggesting that the predicted ORF1 CDS is truncated – the actual ORF1  
208 (eORF1) that is expressed shares the same SSC as ORF10, but has a unique STC.

#### 209 **Early Region 2 (E2) and Intermediate Region (IM) transcripts**

210 The E2 TU expressed on the anti-sense strand is subdivided into E2A and E2B and encodes three classical  
211 AdV proteins – pTP and Ad-pol (E2B proteins), and DBP (E2A protein) – essential for genome replication  
212 (17, 18). Unlike MAdV where two promoters (E2-early and E2-late) are known (17), we discovered only a  
213 single TSS (E2 TSS; 18,751bp) from which both E2A and E2B transcription is initiated. However, similar  
214 to MAdVs, E2A and E2B transcripts have distinct TTSs, and the E2B transcripts share the TTS of the IVa2  
215 transcript of the IM region (17, 18) (**Figure 7**).

216 The E2A ORF, DBP is one of three THEV ORFs predicted to be spliced from two exons. The corre-  
217 sponding transcript (TRXPT\_21) found in our data matches this predicted splice junction precisely but with  
218 a non-coding additional exon at the 5'-end (E2-5'UTR) at position 18,684-18,751 bp. Thus, TRXPT\_21  
219 is a three-exon transcript encoding DBP (380 residues, 43.3 kDa) precisely as predicted. This transcript  
220 (TRXPT\_21) was also corroborated in a 3'RACE experiment. Additionally, from the 3'RACE, a splice variant  
221 of TRXPT\_21 which retains the second intron leading to a 2-exon transcript was found. This new transcript  
222 (TRXPT\_21B), albeit longer due to retaining the second intron and possessing a short 3' UTR, encodes  
223 a truncated isoform of DBP (tDBP) because the SSC utilized by TRXPT\_21, is followed shortly by STCs  
224 in the retained intron. The SSC 173 bp downstream of DBP's SSC yields tDBP (a 346 residue, 39.3 kDa  
225 product), which is in-frame of DBP but entirely contained in the second exon. TRXPT\_21 and TRXPT\_21B  
226 share a common TTS but TRXPT\_21B as seen in our 3'-RACE data, extends 39 bp into an adenine/thymine  
227 (A/T)-rich sequence before the poly-A tail sequence occur, suggesting this position (16,934bp) as the bona  
228 fide E2A TTS (**Figure 7**).

229 The E2B region transcripts also start with the E2-5'UTR but extend thousands of base pairs downstream to  
230 reach the TTS at 2334bp in the IM region, which is immediately followed by an A/T rich sequence (position

231 2323-2339bp) where polyadenylation probably occurs. Interestingly, the TTS of the E1 region (position  
232 2,325bp) on the sense strand is also in the immediate vicinity of this A/T rich sequence, which is almost  
233 palindromic; hence it likely serves as the polyadenylation signal for both E1 and E2B/IM transcripts. The  
234 E2B transcripts, TRXPT\_6 and TRXPT\_7 are almost identical except for an extra splice junction at the 3'-  
235 end of TRXPT\_6, making TRXPT\_6 a five-exon transcript and TRXPT\_7, four exons (**Figure 7**). TRXPT\_7  
236 has the CP for both classical proteins (pTP and Ad-pol) encoded in this region, of which the pTP ORF is  
237 predicted to be spliced from two exons just like in all other AdVs. The predicted splice junction of pTP  
238 is corroborated by our data; however, the full transcript is markedly longer than the predicted ORF: there  
239 are two novel non-coding 5' exons, the third exon (containing the SSC of pTP) is significantly longer than  
240 predicted, and the last exon containing the bulk of the CDS is more than triple the predicted size of pTP. The  
241 first two exons are 5'-UTRs because the SSC here is immediately followed by STCs; thus, the 5'-most SSC  
242 (position 10,995bp) of the third exon which matches the predicted SSC of pTP is utilized. The encoded  
243 product is identical to the predicted pTP protein (597 residues; 70.5 kDa). If secondary SSC (secSSC)  
244 usage is considered, with SSC at 6768bp and STC at 3430bp, the encoded product is identical to the  
245 predicted Ad-pol (polymerase) protein (1112 residues; 129.2 kDa). TRXPT\_6 differs from TRXPT\_7 by  
246 containing an extra splice site at 3447-3515bp. However, the CP remains similar to that of TRXPT\_7 except  
247 the Ad-pol encoded from the secSSC is a truncated isoform with a new STC resulting from the splice site.

248 While both TRXPT\_6 and TRXPT\_7 have the CP for Ad-pol with secSSC usage, in all AdVs studied, the two  
249 proteins (pTP and Ad-pol) are encoded by separate mRNAs with identical first three 5' exons and TTS, but  
250 the splice junction to the terminal exons are different. We checked for a longer splice junction between the  
251 third and fourth (terminal) exons of TRXPT\_7 with our junction validation method (targeted PCR, cloning,  
252 and Sanger sequencing) and discovered a unique splice junction (10,981-7062bp) not found in our RNA-  
253 seq data. If initiated from the E2 TSS and terminated at the E2 TTS, this transcript (TRXPT\_31) would  
254 encode Ad-pol exactly as predicted as its 5'-most CDS (**Figure 7**).

255 Our RNA-seq data also showed a novel short transcript (TRXPT\_15) entirely nested within the terminal  
256 exon of TRXPT\_7 but with a unique splice site. This transcript is an incomplete construction from the  
257 mapped reads as it contains a truncated CDS. However, we validated this splice junction to be genuine  
258 (**Supplementary PCR methods**).

259 The IM region is a single-transcript TU, encoding a single classical protein, IVa2. The promoter expressing  
260 this single transcript (TRXPT\_5) is embedded in E2B region and shares a TTS with E2B transcripts (17,  
261 18). TRXPT\_5 is a two-exon transcript spliced exactly as the last splice junction of TRXPT\_6. The first  
262 exon is a UTR, except the last 2 nucleotides, which connect with the first nucleotide of the second exon to

263 form the 5'-most SSC. This first SSC is 4 codons upstream and in-frame of the predicted IVa2 SSC. Except  
264 for the four extra N-terminus residues, the entire protein sequence is identical to the predicted IVa2.

265 **Early Region 3 (E3) transcripts.**

266 The E3 region is wholly contained in the MLTU and encodes proteins involved in modulating and evading  
267 the host immune defenses. In MAdVs, this region contains seven ORFs expressed from several transcripts  
268 which share the same TSS (from the E3 promoter) but have different TTSs (6, 17, 18). However, some  
269 E3 transcripts use the TSS of the MLP. Due to sharing the same TSS, in MAdVs, secSSC usage is heavily  
270 relied on for gene expression in this region except for 12.5K and transcripts using the MLP's TSS, as utilizing  
271 only the first SSC cannot produce all the other transcripts in this TU (17).

272 In THEV, only one ORF (E3) was predicted in this region. However, as the E3 TU is nested in the MLTU,  
273 transcripts from the L4P (100K, 22K, 33K, and pVIII) not only overlap the E3 region transcripts entirely as  
274 seen in our RNA-seq results, but also have their TSS and TTS in practically the same locations (**Figure 8**).  
275 Therefore, we have categorized these two groups together as E3 transcripts.

276 We identified seven novel transcripts here (**TRXPT\_22, TRXPT\_23, TRXPT\_24, TRXPT\_25, TRXPT\_26,**  
277 **TRXPT\_27, TRXPT\_29**) from our RNA-seq data, all originating from two distinct TSSs – we consider the  
278 first TSS (position 18,230bp) as corresponding to the L4P and the other at 18,727bp as corresponding to  
279 the E3 promoter (E3P). These E3 transcripts collectively have the CP for several predicted THEV ORFs:  
280 100K, 22K, 33K, pVIII, and E3, as well as Fiber (IV) and ORF7 belonging to the MLTU. But some of these  
281 CDSs are different than predicted due to either unknown exons or the presence of an in-frame upstream  
282 SSC. For instance, 33K is one of the few THEV ORFs predicted to be spliced from two exons; however,  
283 we discovered a significantly longer four-exon ORF (e33K) on TRXPT\_24 that contains it almost entirely.  
284 The first two exons of e33K were not predicted but the last two match the predicted exons and the CDS is  
285 in-frame, albeit the first 20bp of the predicted 33K (including the SSC at 20,142bp) is spliced out as part  
286 of the second intron of TRXPT\_24. Thus, the bona fide 33K (e33K) is a 19.8 kDa, 171 residue protein  
287 spanning four exons instead of the predicted 120 aa protein. TRXPT\_24 also has the CP for pVIII and  
288 E3 if we consider downstream SSC usage. However, the predicted E3 has an upstream in-frame SSC;  
289 thus, this longer version of E3 (eE3) is the genuinely expressed ORF. TRXPT\_29 is the shortest transcript  
290 in this TU. It is a two-exon transcript, both exons comprising the CDS. The product of TRXPT\_29 is a  
291 novel 73 residue protein (8.3KII) sharing the SSC of e33K but with a unique STC. TRXPT\_23 being spliced  
292 identically as TRXPT\_29 also encodes 8.3KII from its first SSC. Similarly, TRXPT\_22 also encodes a 73 aa  
293 novel protein (8.3KII) from its first SSC that shares over 80% similarity with 8.3KII, but it differs from 8.3KII at  
294 the C-terminus. Considering downstream SSC usage, both TRXPT\_22 and TRXPT\_23 can encode pVIII

295 and eE3 in that order, but TRXPT\_23 being longer, has the CP for the Fiber ORF also.

296 As the splice junctions of TRXPT\_22, TRXPT\_23, TRXPT\_24, and TRXPT\_29 essentially share the same  
297 genomic space, their validation was done with a single primer pair, and they were differentiated from each  
298 other by cloning and Sanger sequencing (**Supplementary PCR methods**). In addition to corroborating  
299 the splice junctions for the aforementioned transcripts, the Sanger sequencing results also showed another  
300 splice variant undetected in our RNA-seq transcriptome. This was a three-exon transcript (TRXPT\_30) with  
301 its first and last exons spliced identically as TRXPT\_23, but which also has the second exon of TRXPT\_24  
302 (**Figure 8**). The first CDS on TRXPT\_30 spans all three exons, producing a novel 140 residue, 15.7kDa  
303 protein. Interestingly, the last 81 C-terminus residues of this new protein (e22K) are identical to 22K (89  
304 residues), which is a single-exon ORF predicted to use the same SSC as 33K (20,142bp). Just as seen for  
305 33K, all the transcripts in this region exclude the first 20bp of 22K (including the SSC) as part of their introns;  
306 therefore, the first 7 residues of 22K are lacking in e22K due to splicing. Hence, we consider e22K as a  
307 long variant of the predicted 22K ORF. Albeit the TSS and TTS of TRXPT\_30 was not seen, we presume  
308 that they are similar to TRXPT\_23, in which case it would also have the downstream CP of TRXPT\_23.

309 TRXPT\_25 is the largest transcript in the TU. It also utilizes the L4P TSS but has a distinct TTS. It is  
310 a two-exon transcript, encoding a novel protein (t100K; 543 residues), which is a shorter isoform of the  
311 predicted 100K ORF. Considering secSSC usage on this transcript yields the predicted 22K ORF precisely.  
312 It also has the CP for pVIII and eE3 in that order. Furthermore, during the validation of TRXPT\_25's splice  
313 junction using primers that span its junction (18350-18717bp), we noticed a DNA band that corresponds to  
314 the full unspliced sequence (**Supplementary PCR methods**). As TRXPT\_25 only falls short of encoding  
315 the complete predicted 100K protein due to its splice junction, this band (which we cloned and validated by  
316 Sanger sequencing) suggests that the predicted 100K is indeed expressed. This transcript (TRXPT\_25B)  
317 although not seen in full, likely shares the same TSS and TTS as TRXPT\_25. Lastly, TRXPT\_26 and  
318 TRXPT\_27 both originate from the E3 TSS but have distinct TTSs. TRXPT\_26 is a three-exon transcript  
319 but the first two are UTRs. It encodes pVIII as the 5'-most ORF and has the CP for eE3 and Fiber in that  
320 order. TRXPT\_27 on the other hand, is only a two-exon transcript but similar to TRXPT\_26, only the terminal  
321 exon contains the CDSs. It encodes Fiber as the 5'-most ORF, and ORF7 downstream with secSSC usage.  
322 TRXPT\_13, which is an L4 transcript that uses the MLP TSS is discussed under the MLTU transcripts.

### 323 **Early Region 4 (E4) transcripts**

324 This TU is found at the tail-end (3'-end) of the genome and expressed from the anti-sense strand. Based  
325 on nucleotide position, ORF7 and ORF8 were predicted in this region (1); however, as ORF7 is neither on  
326 the same strand as ORF8 nor transcribed from a promoter in the E4 region, only ORF8 can legitimately

327 be classified as a transcript in this TU. This is corroborated by our RNA-seq data, as only one transcript  
328 was identified in this region on the anti-sense strand (**Figure 9**). The transcript (TRXPT\_28) spans 25192-  
329 26247bp and is spliced at 25701-26055bp, making a two-exon transcript. The second exon fully matches  
330 the predicted ORF8 with 12 extra base pairs at the 3'-end. However, there is an SSC in the first exon at  
331 position 26246bp (192bp upstream of the predicted SSC). The encoded protein from this SSC is in-frame  
332 with the predicted SSC found in the second exon; hence, we consider this protein (eORF8 – 26.4 kDa, 229  
333 aa), a longer isoform of the predicted ORF8, the genuinely expressed ORF with an identical C-terminus to  
334 the predicted ORF8 protein.

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

336 The MLTU transcripts dominate the late phase (i.e, after DNA replication) of the AdV infectious cycle.  
337 The MLP produces all late mRNAs by alternative splicing and alternative polyadenylation of a primary  
338 transcript, grouped into five transcript classes (L1-L5). Most of THEV's coding capacity falls within this  
339 TU. Specifically, about 13 out of the 23 predicted ORFs were assigned to this TU, some of which we have  
340 categorized under the E3 TU instead. Our RNA-seq data revealed 12 transcripts (**TRXPT\_8, TRXPT\_9,**  
341 **TRXPT\_10, TRXPT\_11, TRXPT\_12, TRXPT\_13, TRXPT\_14, TRXPT\_16, TRXPT\_17, TRXPT\_18,**  
342 **TRXPT\_19, TRXPT\_20**) in this TU, the majority of which have the 5' untranslated TPL sequence as seen  
343 in all AdVs. For three transcripts (**TRXPT\_16, TRXPT\_17, TRXPT\_18**), a different leader sequence (sTPL)  
344 is used, which differs from the TPL in only one regard: the first TPL exon is substituted for a different first  
345 exon, found between the first and second TPL exons. Also, TRXPT\_20 seems to include only the third TPL  
346 exon (**Figure 10**).

347 We identified five TTSs (10,549bp, 12,709bp, 16,870bp, 17,891bp, 20,865bp) in this TU, which we consider  
348 as corresponding to the five late mRNA classes (L1-L5), respectively, as found in all AdVs. L1 mRNAs  
349 include TRXPT\_8, which comprises the TPL (non-coding) and the CDS-containing terminal exon. This  
350 transcript encodes the 52K ORF exactly as predicted with the SSC beginning from the first nucleotide of  
351 the terminal exon. L2 mRNAs include TRXPT\_16, TRXPT\_17, and TRXPT\_18, all of which consist of the  
352 sTPL (also non-coding) followed by their respective terminal exons. TRXPT\_16 encodes pIIIa exactly as  
353 predicted as the 5'-most ORF, and also has the CP for the ORFs, III and pVII in that order. TRXPT\_17  
354 encodes the ORF, III (penton), and TRXPT\_18 encodes the ORF pVII exactly as predicted. The L3 mRNAs  
355 include TRXPT\_14 and TRXPT\_20, of which TRXPT\_14 utilizes the full TPL whereas TRXPT\_20 uses  
356 only the third TPL exon (TPL3). Both transcripts have the CP for the ORF, hexon (II) but hexon is the  
357 only ORF encoded on TRXPT\_14, whereas the 5'-most ORF on TRXPT\_20 is pX (pre-Mu) followed by  
358 pVI and hexon in that order. L4 mRNAs include TRXPT\_9, TRXPT\_10, TRXPT\_11, and TRXPT\_13 all of

359 which begin with the TPL followed by three (TRXPT\_9, TRXPT\_10, and TRXPT\_13) or four (TRXPT\_11)  
360 coding exons. These are the largest transcripts found in the transcriptome, each one possessing the CP  
361 for several similar late proteins. Normally, MLTU transcripts encoding particular ORFs splice the TPL onto  
362 a splice site just upstream of the ORF to be expressed (17). While this holds true for most MLTU ORFs,  
363 several late ORFs (pVI, protease, and ORF7) do not have such close proximity splicing but are contained in  
364 larger transcripts such as these L4 mRNAs, strongly suggesting the use of non-standard ribosomal initiation  
365 mechanisms such as secSSC utility and ribosome shunting found in other AdVs for their translation (17,  
366 28). TRXPT\_9 and TRXPT\_10 are very similar but not identical. The last exon of TRXPT\_9 seems to be  
367 truncated and probably shares the same TTS as the other L4 mRNAs. They are both 6-exon transcripts  
368 encoding pVII as the 5'-most ORF (fourth exon) and also have the CP for pX, pVI, hexon, a longer variant of  
369 protease (eProt) – uses an upstream in-frame SSC than predicted, and ORF12 (a novel unpredicted 120 aa  
370 protein). TRXPT\_10 (and TRXPT\_9 with the L4 TTS) additionally has the CP for pVIII and eE3. Conversely,  
371 TRXPT\_11 is a seven-exon mRNA with hexon as it's 5'-most ORF but it also has the CP for eProt, ORF12,  
372 e33K, and also pVIII and eE3 in that order. TRXPT\_13 seems to be an E3 ORF utilizing the MLP TSS as  
373 it encodes classical L4P genes such as pVIII and eE3 in that order similar to TRXPT\_22 (E3 TU) but lacks  
374 TRXPT\_22's novel first ORF (8.3KII).

375 Lastly, the L5 class includes only TRXPT\_12 which contains the TPL and a coding terminal exon. Its 5'-  
376 most ORF is fiber (IV) but it also has the CP for the THEV specific gene, ORF7. TRXPT\_12's CP is identical  
377 to TRXPT\_27 of the the E3 TU but they differ in their 5'-UTRs.

378 **DISCUSSION/CONCLUSIONS**

379 While the advent of next-generation sequencing has rendered easier the study of large and complex eu-  
380 karyotic transcriptomes, the study of the smaller and compact viral transcriptomes remains unintuitively  
381 challenging, as several transcripts may have significant overlaps due to genome economization. Char-  
382 acterizing AdV transcriptomes is even more difficult due to the wide array of mRNAs produced via very  
383 complex alternative splicing combined with alternative polyadenylation, all initiated from relatively few pro-  
384 moters. This makes AdV transcriptomes some of the most intricate for a virus. The challenge is further  
385 compounded by the fact that the standard software programs used in the RNA-seq analysis pipelines are  
386 not designed primarily for such compact, gene-dense, and complex transcriptomes as AdVs. Furthermore,  
387 there is no prior transcriptomic studies for THEV. Our approach to properly handle this complex data was  
388 to use standard RNA-seq analysis programs coupled with some custom analysis and validating all splice  
389 junctions with independent methods. Our work provides the first insights into the splicing patterns of THEV,  
390 which is expectedly similar to other MAdVs but with key differences. Our work shows 34 transcripts in  
391 the THEV transcriptome grouped into five TUs, of which the E3 TU shows great complexity of alternative  
392 splicing.

393 An unexpected observation is that the pileup of mapped reads to THEV seems consistently skewed over  
394 similar regions of the genome at all time points. As AdVs gene expression is temporally regulated, we  
395 expected to see unambiguous differences in the pileup of reads over different regions of the genome at  
396 different time points, indicating the different stages of infection. While this could simply mean that the  
397 infection was not well synchronized, we speculate that the temporal gene expression regulation of THEV is  
398 probably different from MAdVs. This is supported by a previous study stating the same conclusion with its  
399 finding that almost all THEV transcripts were detectable by at 4h.p.i, and by 8h.p.i, mRNA for all predicted  
400 ORFs (including the late genes) were present (24). Conversely, despite the overall pileup similarity, a close  
401 inspection shows that the relative proportions of reads over some regions show some variation over time.  
402 The breakdown of transcripts detected at different time points in **Figure 3b** seems to support this different  
403 temporal regulation of THEV. Specifically, the MLP of THEV is active significantly earlier in infection – as  
404 early as 4h.p.i and more pronounced at 12h.p.i (**Figure 3b** and **Table 2a**), – whereas the late phase shift in  
405 MAdVs occurs after 24h.p.i. This also lends credence to our speculation. However, generally speaking, the  
406 overall temporal gene expression regulation known in MAdVs – early regions showing their peak expression  
407 at earlier time points followed by predominance of the MLTU at later time points – also holds true for THEV.  
408 Further studies would be necessary to establish the precise temporal regulation of THEV transcription.

409 The use of short read deep sequencing to reconstruct full AdV mRNA structures provides excellent results,  
410 especially for mapping the splice sites. However, due to the substantial overlapping nature of AdV mRNAs  
411 coupled with the fragmentation step in the library preparation protocol, mapping the precise TSS and TTS  
412 of the assembled transcripts is difficult. Also, similar transcripts with substantial overlaps may be assembled  
413 as one longer mRNA since the short reads alone do not provide enough context for the transcript assembler  
414 (StringTie) to distinguish them. In our results, we see transcripts in the same TU initiated or terminated in  
415 the same approximate area (10-70bp and 1-300bp apart for TSS and TTS, respectively) but not precisely  
416 at the same position. We consider the most upstream TSS or most downstream TTS for the transcripts  
417 involved but we present them unchanged in all the figures shown. Also, by comparison to the more well-  
418 studies MAdV transcriptomes, we think that a few long transcripts in the MLTU (TRXPT\_9, TRXPT\_10,  
419 and TRXPT\_11) are probably a result of fusing some L4P-derived transcripts to the terminal exons of the  
420 bona fide MLTU transcripts by StringTie, making them significantly longer. These mRNAs do not only have  
421 unusually many exons for an AdV, but their last three or four exons are also identical to the L4P-derived  
422 mRNAs. Future studies using long read sequencing technologies are necessary to provide conclusive data  
423 for precisely mapping the TSS and TTS, as well as teasing apart the bona fide structures of the long MLTU  
424 transcripts. Furthermore, it is not unreasonable to presume that several splice variants were undiscovered  
425 in our work as evidenced firstly by finding unique transcripts using 3'RACE and during our splice junction  
426 validation steps. And secondly, recent studies (17, 18, 22) are still discovering novel mRNA variants for  
427 even the best studied MAdVs decades later. Another observation made is that all the TTSs in THEV's  
428 transcriptome are in close proximity to A/T-rich sequences which we presume to be polyadenylation signal  
429 sequences (PASS). Interestingly, some of these PASSs are located in the immediate vicinity of two closely  
430 located TTSs expressed on opposite strands. Namely, the E1 and E2B/IM TTSs have an almost palindromic  
431 PASS between them, as do the E4 (anti-sense strand) and the sense strand TRXPT\_12 and TRXPT\_27.

432 An interesting finding of our analysis is that while most of the predicted ORFs are precisely encoded by the  
433 spliced transcripts, we found a few that seem to be truncated predictions, as either an upstream in-frame  
434 SSC (eORF1, eE3, and eProt) or unknown upstream exons spliced onto them (eIVa2, e33K, and eORF8)  
435 were found. Other ORFs were identified that were either shorter (tDBP, t100K) or longer (e22K) isoforms of  
436 some predicted ORF but we found evidence to support the predicted ORF itself, making them all possible  
437 genuinely translated variants. We also found several novel unpredicted ORFs. Taken together, we surmise  
438 that further studies will likely yield even more unpredicted novel ORFs or variants of predicted ORFs.

439 Eukaryotic mRNAs are typically functionally monocistronic, the 5'-most AUG normally being used as the  
440 translation reading frame. However, depending on the sequence context, in some organisms, the initiating

441 codon may even be a non-AUG start codon. AdV mRNAs, which mostly span more than one ORF, are  
442 known to be functionally polycistronic, employing non-standard mechanisms of translation initiation, namely,  
443 secSSC usage and ribosome shunting (6, 22). Albeit there is no reliable method of predicting how efficiently  
444 any given AUG will be used, AdVs use secondary AUGs as initiation codons for most E1b proteins and for  
445 some E3 proteins. In fact, recent studies show that secSSC usage is found transcriptome-wide. This is  
446 thought to occur because translation initiation at the first SSC is inefficient, allowing downstream SSCs to be  
447 employed for initiation (17). The ribosomal shunting or jumping mechanism is utilized for MLTU transcripts  
448 that have the TPL. This mechanism allows the ribosome to translocate to a downstream initiating codon  
449 under the direction of the shunting elements in the TPL, even if a start codon in a good Kozak sequence  
450 context is bypassed. Thus, predicting the protein(s) that are expressed from an AdV mRNA becomes highly  
451 uncertain as any one of the SSC may be selected (6, 22). Almost all the THEV transcripts in our data have  
452 the CP for several ORFs, some spanning as many as six ORFs but the majority spanning at least two ORFs.  
453 Therefore, we believe our data supports the usage of these special ribosome initiation mechanisms as a  
454 several predicted and novel ORFs found on mRNA in our data have no conceivable mechanism of being  
455 translated if only the typical ribosome scanning mechanism is employed. Interestingly, several distinct  
456 transcripts have identical CPs. This is not unique to THEV but is observed in human AdVs in a recent  
457 study (17). They proposed that this may permit protein production to be fine-tuned through alteration in the  
458 balance between different mRNA groups expressing that ORF.

459 It is well established that AdV alternative splicing undergoes a regulated temporal shift in splice site usage.  
460 This was thought to be limited to certain TUs; however, recent studies suggest that AdVs routinely produce  
461 different combinations of splice acceptor–donor pairs and that this is observed in all TUs (6, 17, 22, 29).  
462 The mechanistic details of this phenomenon has been best studied for the E1A and L1 units. The studies  
463 show that AdVs (specifically, late phase AdV-infected nuclear extract) modulate the activities of the splicing  
464 factor U2AF and the cellular SR family of splicing factors (reviewed in reference (29)) and encode several  
465 mostly late phase proteins (E4-ORF3, E4-ORF6, E4-ORF4, L4-33K, and L4-22K) that influence the RNA  
466 splice site used. This phenomenon seems to occur in the THEV transcriptome also, as the stringency of  
467 splice acceptor-donor pairs selected decreased measurably from the onset of the late phase (see **Figure**  
468 **5**). In fact, recent studies of some human AdVs show that virtually unlimited number of combinatorial  
469 alternative splicing events resulting in menagerie of novel transcripts are produced in an AdV lytic infection  
470 (17, 22). It is unlikely that all repertoire of mRNA produced via this mechanism will actually be translated.  
471 However, it has been speculated that the plasticity in alternative RNA splicing enables the AdVs to fine-  
472 tune protein synthesis by providing different alternatively spliced variants encoding the same protein under

473 changing conditions. And also that the capacity to produce novel exon combinations will offer the virus  
474 an evolutionary advantage to change the gene expression repertoire and protein production in a changing  
475 environment (17, 22).

476 Summarizing all the main points above, we see that the THEV transcriptome bares remarkable overall  
477 similarity to the better studied MAdVs. The transcriptome organization into five TUs, the overall regulation of  
478 early and late genes, and the production of a broad repertoire of transcripts via virtually unlimited alternative  
479 splicing. However, the THEV transcriptome appears to be less sophisticated (i.e, encode less genes) than  
480 MAdVs primarily because the MAdV genomes are close to twice a long as that of THEV's, which rationally  
481 should encode less genes. The lack of subdivision of the E1 region into E1a and E1b is one of the most  
482 obvious examples. Also, the MAdV E4 region encodes several proteins unlike in THEV where only one  
483 transcript coding for only one protein was found. The most conspicuous example is found in examining the  
484 complexity of the MLTU leader sequences. While the majority of THEV's MLTU transcripts begin with the  
485 TPL (267bp long) just like MAdVs and also utilizes a variant leader sequence (sTPL), it is well established  
486 that a significantly more diverse 5'UTRs are employed for MAdV MLTU transcripts, including the TPL (used  
487 for majority of transcripts), the so-called x, y, and z leaders, and the i-leader. Granted, the MAdV MLTU  
488 transcripts infrequently incorporate the the non-TPL leaders, their absence in our data could mean that the  
489 5'UTR diversity of THEV's MLTU mRNA are indeed more limited due to its smaller genome size. It is also  
490 possible that later studies could uncover more variety not seen our results.

491 **MATERIALS AND METHODS**

492 **Cell culture and THEV Infection**

493 The Turkey B-cell line (MDTC-RP19, ATCC CRL-8135) was grown as suspension cultures in 1:1 complete  
494 Leibovitz's L-15/McCoy's 5A medium with 10% fetal bovine serum (FBS), 20% chicken serum (ChS), 5%  
495 tryptose phosphate broth (TPB), and 1% antibiotics solution (100 U/mL Penicillin and 100ug/mL Strepto-  
496 mycin), at 41°C in a humidified atmosphere with 5% CO<sub>2</sub>. Infected cells were maintained in 1:1 serum-  
497 reduced Leibovitz's L15/McCoy's 5A media (SRLM) with 2.5% FBS, 5% ChS, 1.2% TPB, and 1% antibiotics  
498 solution (100 U/mL Penicillin and 100ug/mL Streptomycin). A commercially available HE vaccine was pur-  
499 chased from Hygieia Biological Labs as a source of THEV-A (VAS strain). The stock virus was titrated using  
500 an in-house qPCR assay with titer expressed as genome copy number (GCN)/mL, similar to Mahshoub *et al*  
501 (30) with modifications. Cells were infected in triplicates at a multiplicity of infection (MOI) of 100 GCN/cell,  
502 incubate at 41°C for 1 hour, and washed three times to get rid of free virion particles. Samples in tripli-  
503 cates were harvested at 4-, 12-, 24-, and 72-h.p.i for total RNA extraction. The infection was repeated but  
504 samples in triplicates were harvested at 12-, 24-, 36-, 48-, and 72-h.p.i for PCR validation of novel splice  
505 sites. Still one more independent infection was done at time points ranging from 12 to 168-h.p.i for qPCR  
506 quantification of virus titers.

507 **RNA extraction and Sequencing**

508 Total RNA was extracted from infected cells using Thermofishers' RNAqueous™-4PCR Total RNA Isolation  
509 Kit (#AM1914) per manufacturer's instructions. An agarose gel electrophoresis was performed to check  
510 RNA integrity. The RNA quantity and purity was initially assessed using nanodrop, and RNA was used only  
511 if the A260/A280 ratio was 2.0 ± 0.05 and the A260/A230 ratio was >2 and <2.2. Extracted total RNA sam-  
512 ples were sent to LC Sciences, Houston TX for poly-A-tailed mRNA sequencing where RNA integrity was  
513 checked with Agilent Technologies 2100 Bioanalyzer High Sensitivity DNA Chip and poly(A) RNA-  
514 seq library was prepared following Illumina's TruSeq-stranded-mRNA sample preparation protocol.  
515 Paired-end sequencing was performed on Illumina's NovaSeq 6000 sequencing system.

516 **Validation of Novel Splice Junctions**

517 All splice junctions identified in this work are novel except one predicted splice site each for pTP, DBP, and  
518 33K, which were corroborated in our work. However, these predicted splice junctions had not been exper-

519 imentally validated hitherto, and we identified additional novel exons, giving the complete picture of these  
520 transcripts. The novel splice junctions discovered in this work using the StringTie transcript assembler were  
521 validated by PCR, cloning, and Sanger Sequencing (**Supplementary PCR methods**). Briefly, we designed  
522 primers that span a range of novel exon-exon boundaries for each specific transcript in a transcription unit  
523 (TU). We designed a universal forward or reverse primers for each respective TU and paired them with  
524 primers binding specific positions in each transcript. Each forward primer contained a KpnI restriction site  
525 and reverse primers, an XbaI site in the primer tails. After first-strand cDNA synthesis of total RNA ex-  
526 tracted from THEV infected MDTC-RP19 cells with SuperScript™ IV First-Strand Synthesis System, these  
527 primers were used in a targeted PCR amplification, the products analyzed with agarose gel electrophoresis  
528 to confirm expected band sizes, cloned by traditional restriction enzyme method, and Sanger sequenced to  
529 validate these splice junctions at the sequence level.

### 530 **3' Rapid Amplification of cDNA Ends (3'-RACE)**

531 We performed a rapid amplification of sequences from the 3' ends of mRNAs (3'-RACE) experiment us-  
532 ing a portion of the extracted total RNA of infected MDTC-RP19 cells used for the RNA-seq experiment  
533 as explained above. We followed the protocol described by Green *et al* (31) with modifications. Briefly,  
534 1ug of total RNA was reverse transcribed to cDNA using SuperScript™ IV First-Strand Synthesis System  
535 following the manufacturing instructions using an adapter-primer with a 3'-end poly(T) and a 5'-end BamHI  
536 restriction site. A gene-specific sense primer with a 5'-end KpnI restriction site paired with an anti-sense  
537 adapter-primer with a 5'-end BamHI site were used to amplify target sections of the cDNA using Invitrogen's  
538 Platinum™ Taq DNA polymerase High Fidelity, following manufacturer's instructions. The PCR amplicons  
539 were restriction digested, cloned, and Sanger sequenced.

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

541 Our sequence reads were analyzed following a well-established protocol described by Pertea *et*  
542 *al* (25), using Snakemake - version 7.24.0 (32), a popular workflow management system to  
543 drive the pipeline. Briefly, sequencing reads were trimmed with the Trim-galore - version  
544 0.6.6 (33) program to achieve an overall Mean Sequence Quality (Phred Score) of 36. Trimmed  
545 reads were mapped simultaneously to the complete genomic sequence of avirulent turkey hemor-  
546 rhagic enteritis virus (<https://www.ncbi.nlm.nih.gov/nuccore/AY849321.1/>) and *Meleagris gallopavo*  
547 (<https://www.ncbi.nlm.nih.gov/genome/?term=Meleagris+gallopavo>) using Hisat2 - version 2.2.1 (25)

548 with default settings. The generated alignment (BAM) files from each infection time point were filtered  
549 for reads mapping to the THEV genome using Samtools – version 1.16.1 and fed into StringTie –  
550 version 2.2.1 (25) to assemble the transcripts, using a GTF annotation file derived from a GFF3 annotation  
551 file obtained from NCBI, which contains the predicted ORFs of THEV as a guide. GFFCOMPARE – version  
552 0.12.6 was used to merge all transcripts from all time points without redundancy and using a custom R  
553 script, adenovirus transcripts units (regions) were assigned to each transcript, generating the transcriptome  
554 of THEV. StringTie set to expression estimation mode was used to calculate FPKM scores for all  
555 transcripts after which Ballgown – version 2.33.0 in R was used to perform the statistical analysis on the  
556 transcript expression levels. Samtools was also used to count the total sequencing reads for all replicates  
557 at each time point and Regtools – version 1.0.0 was used to count all junctions, the reads supporting  
558 them, and extract all other information related to the junction. See **Supplementary Computational**  
559 **Analysis** for the details of transcript expression level estimations and splice junction read counts.

## 560 DATA AVAILABILITY

561 The raw sequence data (FastQ), transcript expression counts, and total unique junctions have been de-  
562 posited at the National Center for Biotechnology Information Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo>) under accession number GSE254416.  
563  
564 Data is also available on request by contacting the designated corresponding author

## 565 CODE AVAILABILITY

566 All the code/scripts in the entire analysis pipeline are available on github ([https://github.com/Abraham-Quaye/thev\\_transcriptome](https://github.com/Abraham-Quaye/thev_transcriptome))  
567

568 **ACKNOWLEDGMENTS**

569 BYU high performance computing systems - Memory-intensive analysis were run here.

570 REFERENCES

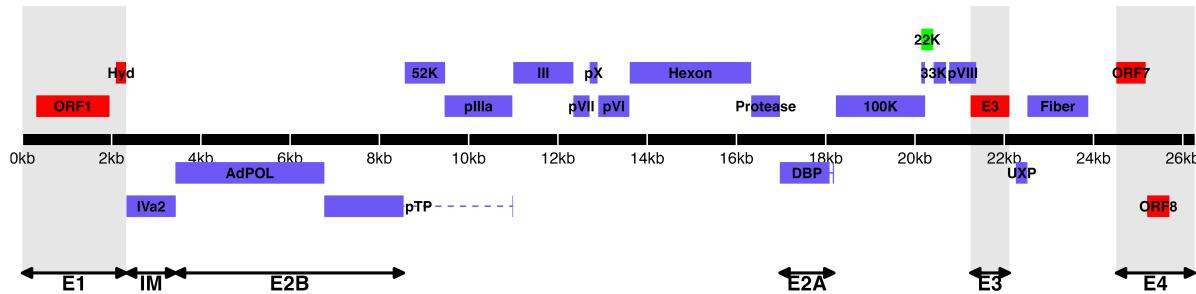
- 571 1. Davison A, Benko M, Harrach B. 2003. Genetic content and evolution of adenoviruses. *The Journal*  
572 of general virology
- 573 2. Harrach B. 2008. Adenoviruses: General features, p. 1–9. *In* Mahy, BWJ, Van Regenmortel, MHV  
574 (eds.), *Encyclopedia of virology* (third edition). Book Section. Academic Press, Oxford.
- 575 3. Upton C, Slack S, Hunter AL, Ehlers A, Roper RL. 2003. Poxvirus orthologous clusters: Toward  
576 defining the minimum essential poxvirus genome. *Journal of virology* 77:7590–7600.
- 577 4. McGeoch D, Davison AJ. 1999. Chapter 17 - the molecular evolutionary history of the herpesviruses,  
578 p. 441–465. *In* Domingo, E, Webster, R, Holland, J (eds.), *Origin and evolution of viruses*. Book  
Section. Academic Press, London.
- 579 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*  
580 Report of the International Committee on Taxonomy of Viruses 125–141.
- 581 6. Guimet D, Hearing P. 2016. 3 - adenovirus replication, p. 59–84. *In* Curiel, DT (ed.), *Adenoviral*  
582 vectors for gene therapy (second edition). Book Section. Academic Press, San Diego.
- 583 7. Kovács ER, Benkő M. 2011. Complete sequence of raptor adenovirus 1 confirms the characteristic  
584 genome organization of siadenoviruses. *Infection, Genetics and Evolution* 11:1058–1065.
- 585 8. Davison AJ, Wright KM, Harrach B. 2000. DNA sequence of frog adenovirus. *J Gen Virol* 81:2431–  
586 2439.
- 587 9. Kovács ER, Jánoska M, Dán Á, Harrach B, Benkő M. 2010. Recognition and partial genome char-  
588 acterization 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.
- 589 10. Katoh H, Ohya K, Kubo M, Murata K, Yanai T, Fukushi H. 2009. A novel budgerigar-adenovirus  
590 belonging to group II avian adenovirus of siadenovirus. *Virus Research* 144:294–297.
- 591 11. Beach NM. 2006. Characterization of avirulent turkey hemorrhagic enteritis virus: A study of the  
592 molecular basis for variation in virulence and the occurrence of persistent infection. Thesis.

- 593 12. Beach NM, Duncan RB, Larsen CT, Meng XJ, Sriranganathan N, Pierson FW. 2009. Comparison of  
594 12 turkey hemorrhagic enteritis virus isolates allows prediction of genetic factors affecting virulence.  
595 J Gen Virol 90:1978–85.
- 596
- 597 13. Gross WB, Moore WE. 1967. Hemorrhagic enteritis of turkeys. Avian Dis 11:296–307.
- 598
- 599 14. Rautenschlein S, Sharma JM. 2000. Immunopathogenesis of haemorrhagic enteritis virus (HEV) in  
600 turkeys. Dev Comp Immunol 24:237–46.
- 601 15. Larsen CT, Domermuth CH, Sponenberg DP, Gross WB. 1985. Colibacillosis of turkeys exacerbated  
602 by hemorrhagic enteritis virus. Laboratory studies. Avian Dis 29:729–32.
- 603 16. Dhama K, Gowthaman V, Karthik K, Tiwari R, Sachan S, Kumar MA, Palanivelu M, Malik YS, Singh  
604 RK, Munir M. 2017. Haemorrhagic enteritis of turkeys – current knowledge. Veterinary Quarterly  
605 37:31–42.
- 606
- 607 17. Donovan-Banfield I, Turnell AS, Hiscox JA, Leppard KN, Matthews DA. 2020. Deep splicing plasticity  
608 of the human adenovirus type 5 transcriptome drives virus evolution. Communications Biology 3:124.
- 609 18. Zhao H, Chen M, Pettersson U. 2014. A new look at adenovirus splicing. Virology 456-457:329–341.
- 610
- 611 19. Wolfrum N, Greber UF. 2013. Adenovirus signalling in entry. Cell Microbiol 15:53–62.
- 612
- 613 20. Falvey E, Ziff E. 1983. Sequence arrangement and protein coding capacity of the adenovirus type 2  
614 "i" leader. Journal of Virology 45:185–191.
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.
22. Westergren Jakobsson A, Segerman B, Wallerman O, Bergström Lind S, Zhao H, Rubin C-J, Pet-  
tersson U, Akusjärvi G. 2021. The human adenovirus 2 transcriptome: An amazing complexity of  
alternatively spliced mRNAs. Journal of Virology 95.

- 615 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  
616 cells. *Nature* 489:101–108.
- 617 24. Aboeza Z, Mabsoub H, El-Bagoury G, Pierson F. 2019. In vitro growth kinetics and gene expression  
618 analysis of the turkey adenovirus 3, a siadenovirus. *Virus Research* 263:47–54.
- 619 25. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. 2016. Transcript-level expression analysis of  
620 RNA-seq experiments with HISAT, StringTie and ballgown. *Nature Protocols* 11:1650–1667.
- 621 26. Jack Fu [Aut], Alyssa C. Frazee [Aut, Cre], LeonardoCollado-Torres [Aut], Andrew E. Jaffe [Aut],  
622 Jeffrey T. Leek[Aut, Ths]. 2017. Ballgown. Bioconductor.
- 623 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,  
624 hemorrhagic enteritis virus. *Virology* 249:307–315.
- 625 28. Yueh A, Schneider RJ. 1996. Selective translation initiation by ribosome jumping in adenovirus-  
626 infected and heat-shocked cells. *Genes & Development* 10:1557–1567.
- 627 29. Akusjarvi G. 2008. Temporal regulation of adenovirus major late alternative RNA splicing. *Frontiers  
628 in Bioscience Volume:5006*.
- 629 30. 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  
630 vaccines. *Journal of Virological Methods* 239:42–49.
- 631 31. Green MR, Sambrook J. 2019. Rapid amplification of sequences from the 3' ends of mRNAs: 3'-  
632 RACE. *Cold Spring Harbor Protocols* 2019:pdb.prot095216.

- 633 32. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok  
634 SO, Kanitz A, Wilm A, Holtgrewe M, Rahmann S, Nahnsen S, Köster J. 2021. Sustainable data  
analysis with snakemake. *F1000Research* 10:33.
- 635 33. Krueger F, James F, Ewels P, Afyounian E, Weinstein M, Schuster-Boeckler B, Hulselmans G, Scla-  
636 mons. 2023. FelixKrueger/TrimGalore: v0.6.10 - add default decompression path. Zenodo.

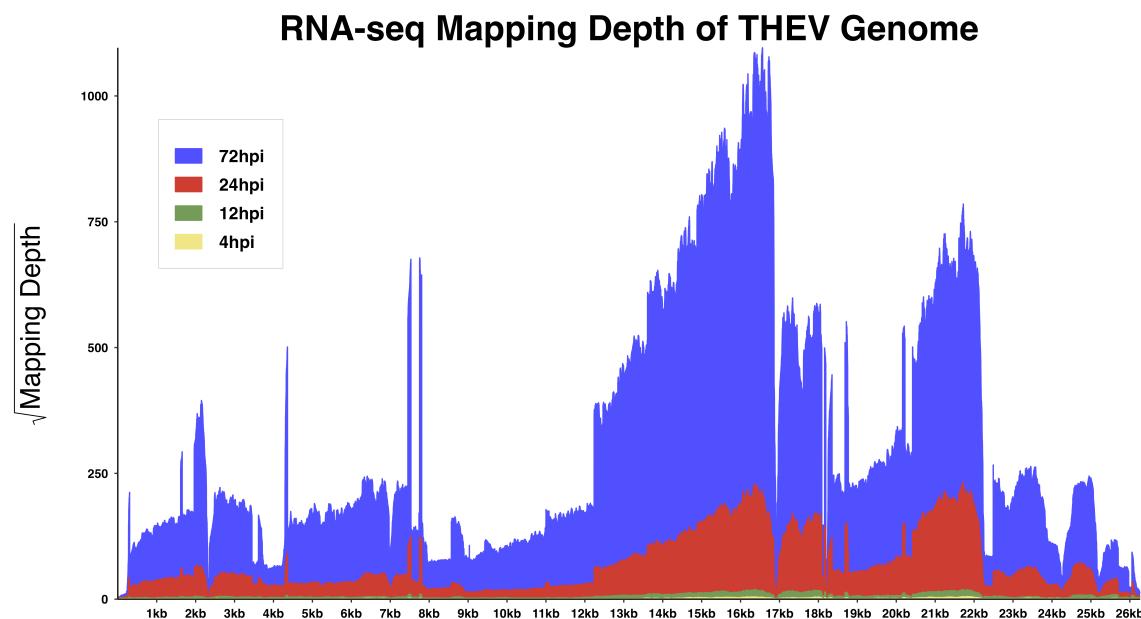
637 **TABLES AND FIGURES**



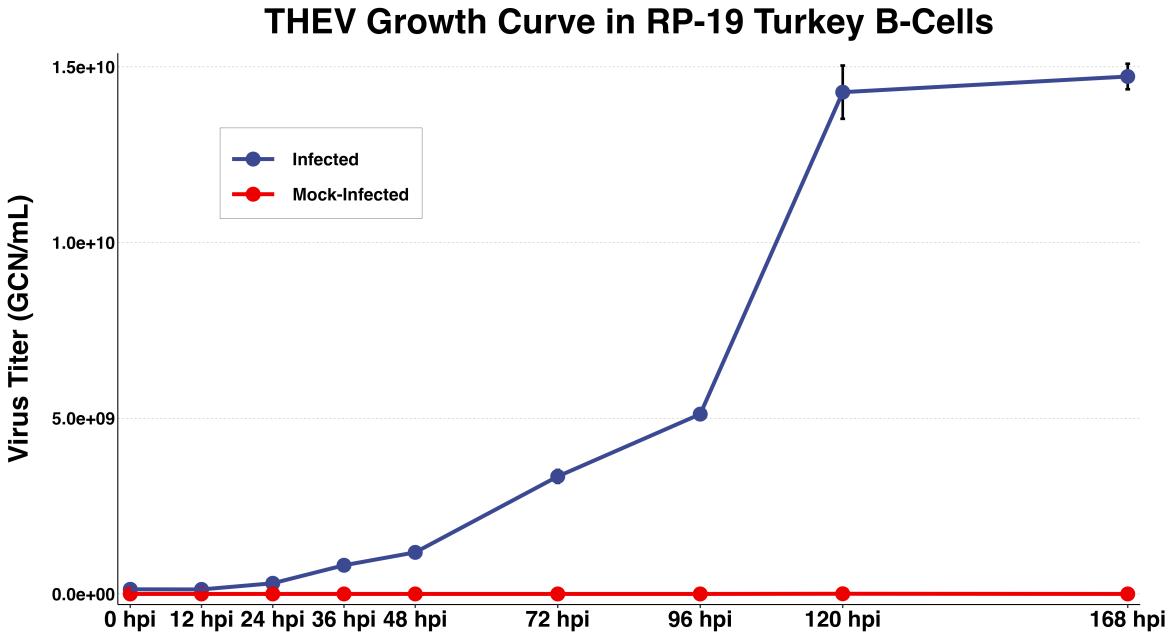
638

639 **Figure 1. Predicted ORF map of THEV virulent strain.** The central horizontal line represents the  
 640 double-stranded DNA marked at 5kb intervals as white line breaks. Blocks represent viral genes. Blocks  
 641 above the DNA line are transcribed rightward, those below are transcribed leftward. pTP, DBP and  
 642 33K predicted to be spliced are shown as having tails. Shaded regions indicate regions containing  
 643 "genus-specific" genes (colored red). Genes colored in blue are "genus-common". Gene colored in light  
 644 green is conserved in all but Atadenoviruses. The UXP (light blue) is an incomplete gene present in almost  
 645 all AdVs. Regions comprising the different transcription units are labelled at the bottom (E1, E2A, E2B,  
 646 E3, and E4); the unlabeled regions comprise the MLTU.

A



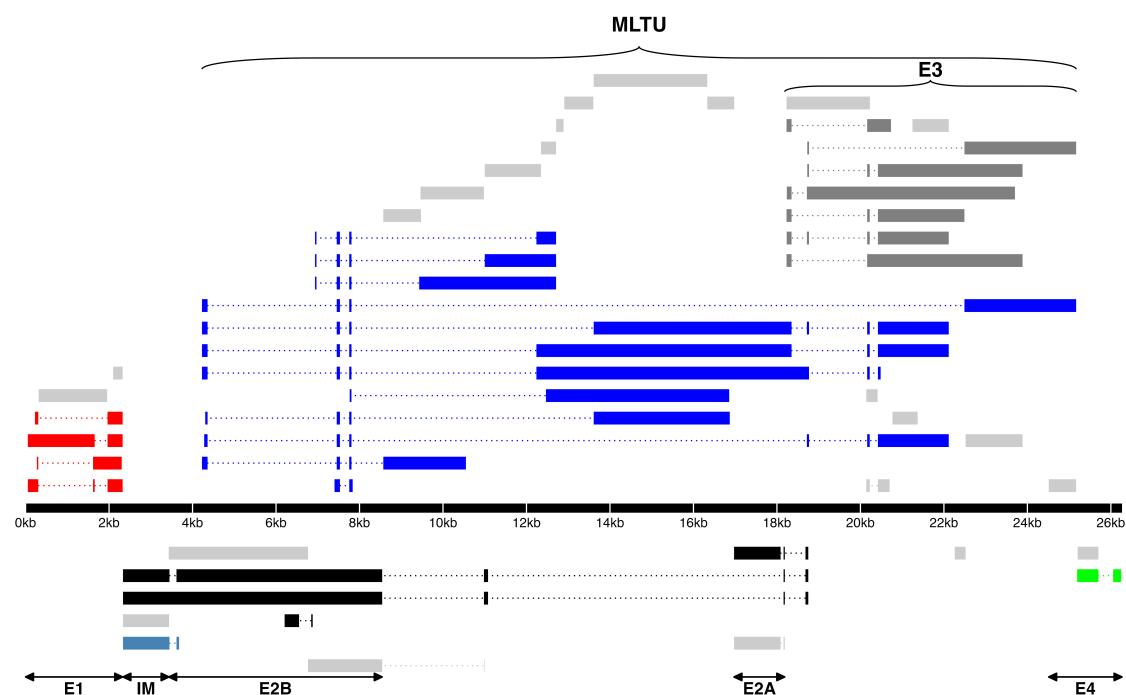
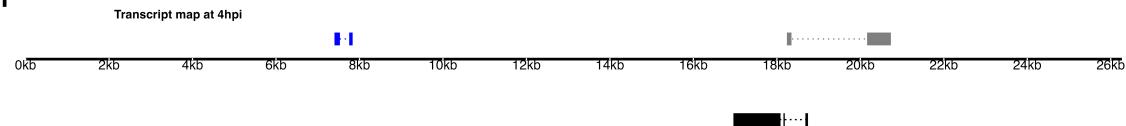
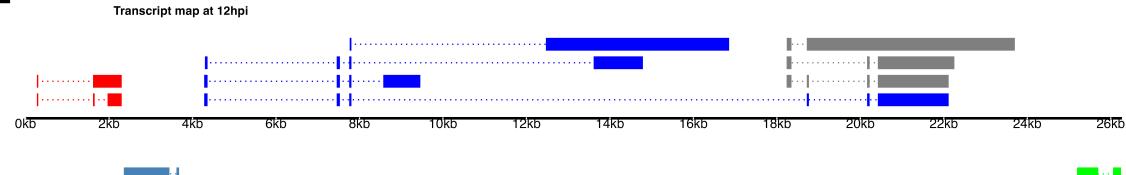
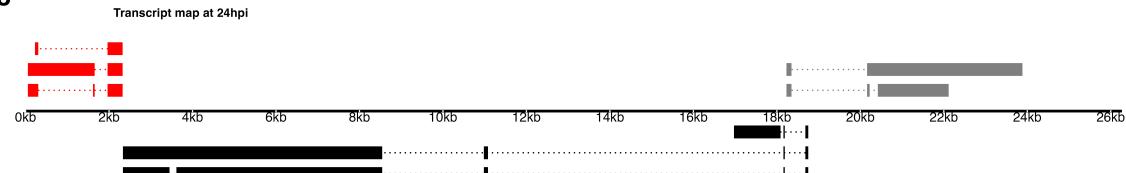
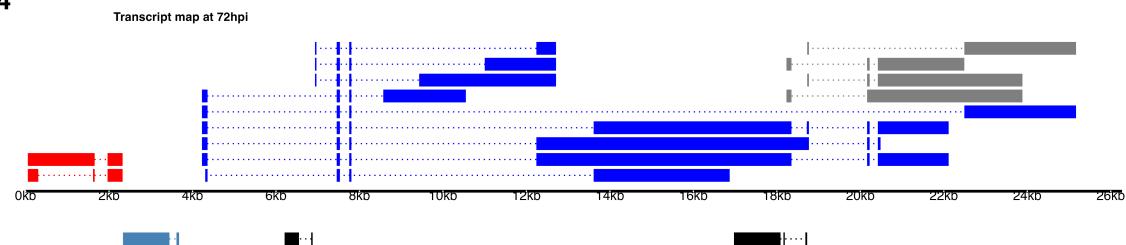
B



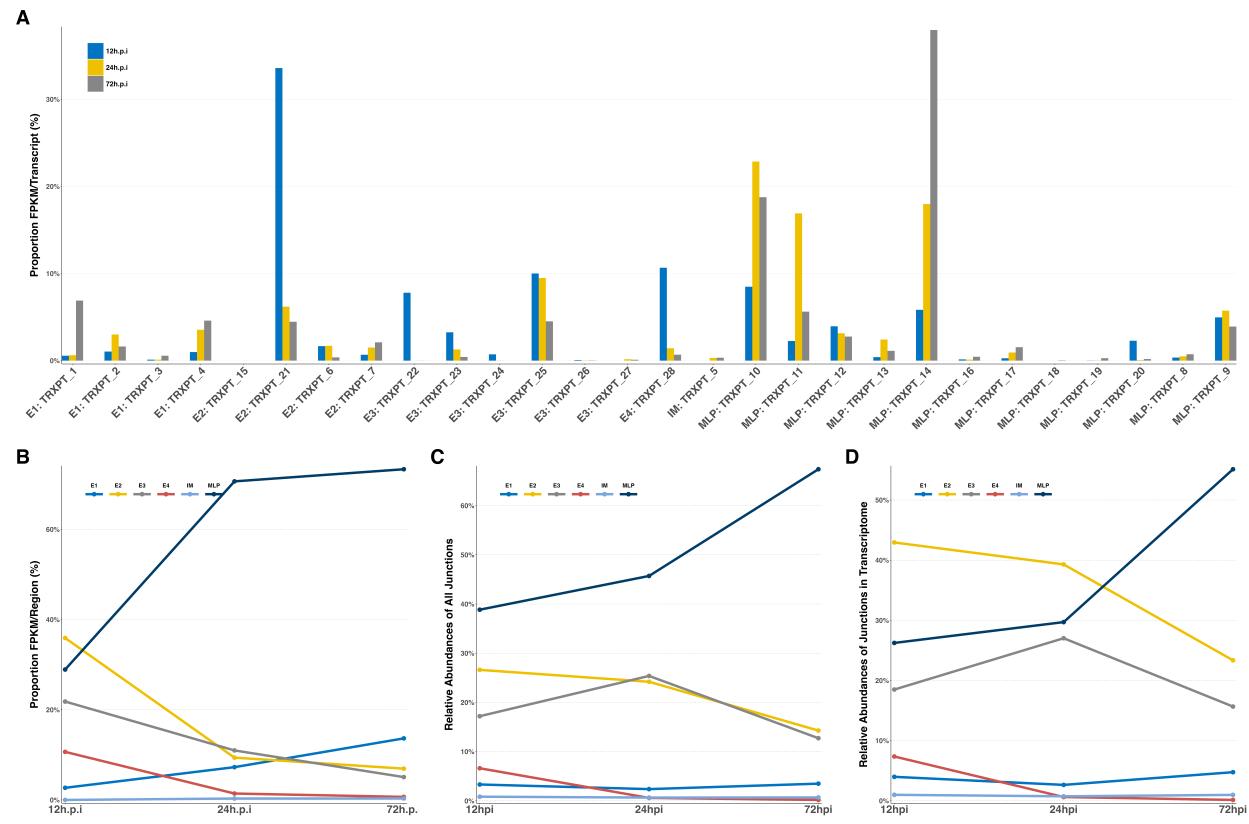
647

648 **Figure 2: Increasing levels of THEV over time. a) Per base coverage of sequence reads mapping to**  
 649 **THEV genome by time point.** The pileup of mRNA reads mapping to THEV genome at the base-pair level  
 650 for each indicated time point. b) **Growth curve of THEV (VAS vaccine strain) in MDTC-RP19 cell line.**  
 651 Virus titers were quantified with a qPCR assay. There is no discernible increase in virus titer up 12 h.p.i,  
 652 after which a steady increase in virus titer is measured. The virus titer expands exponentially beginning

653 from 48 h.p.i, increasing by orders of magnitude before reaching a plateau at 120 h.p.i. GCN: genome copy  
654 number.

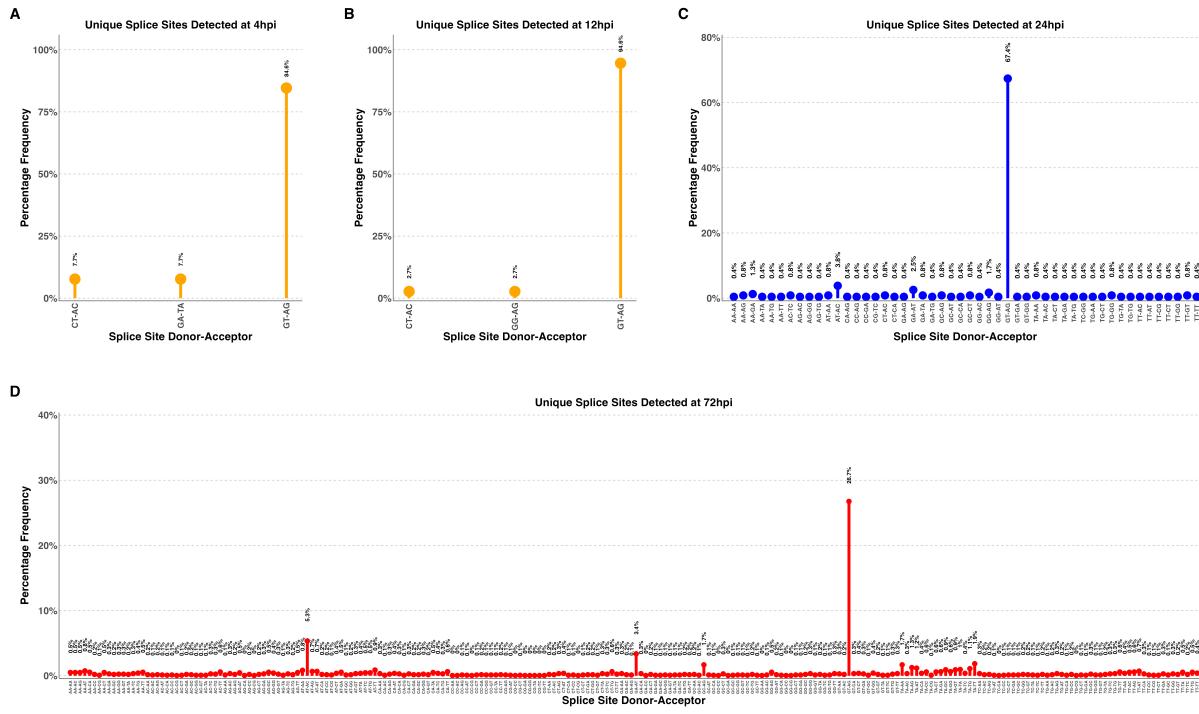
**A****B1****B2****B3****B4**

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

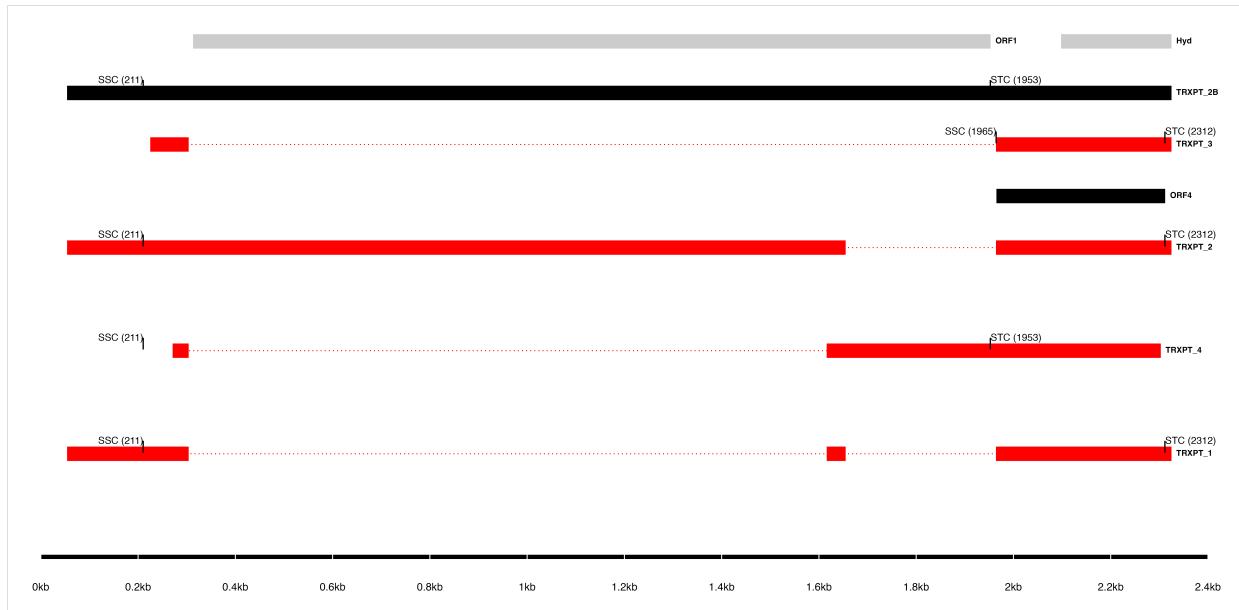


663 **Figure 4: Changes in splicing and expression profile of THEV over time.** **a)** Normalized (FPKM)  
 664 expression levels of transcripts over time. The expression levels (FPKM) of individual transcripts as a  
 665 percentage of the total expression of all transcripts at each time point are indicated. Only transcripts from  
 666 our RNA-seq data are included here. **b)** Normalized (FPKM) expression levels of transcripts by region over  
 667 time. The expression levels of each region/TU as a percentage of the total expression of all transcripts at  
 668 each time point are indicated. Region expression levels were calculated by summing up the FPKMs of all  
 669 transcripts categorized in that region. **c)** Relative abundances of all splice junctions grouped by region/TU  
 670 over time. After assigning all 2,457 unique junctions to a TU and the total junction reads counted at each  
 671 time point for each region, the total junction reads for each TU plotted as percentage of all junction reads at  
 672 each time point.

673 each time point is indicated. Note that the junction read counts are not normalized. **d) Relative abundances**  
 674 *of junctions in transcriptome grouped by region/TU over time*. This is identical to **(c)**, except that only the  
 675 junctions found in the full transcriptome obtained from the RNA-seq data were included.



676  
 677 **Figure 5: Changes in splice donor-acceptor nucleotides over time.** The splice donor-acceptor  
 678 nucleotides of THEV just like other AdVs is mostly the canonical GT-AG. At early time points (4h.p.i and  
 679 12h.p.i [(a) and (b)]) the junction nucleotides used appear to be well scrutinized or restricted, utilizing  
 680 mostly the canonical splice nucleotides. However, as the infection progresses to the late stages (24h.p.i  
 681 and 72h.p.i [(c) and (d)]), the selectivity of specific splice acceptor-donor pairs seems to degenerate  
 682 significantly, such that all combinations of nucleotides are utilized.

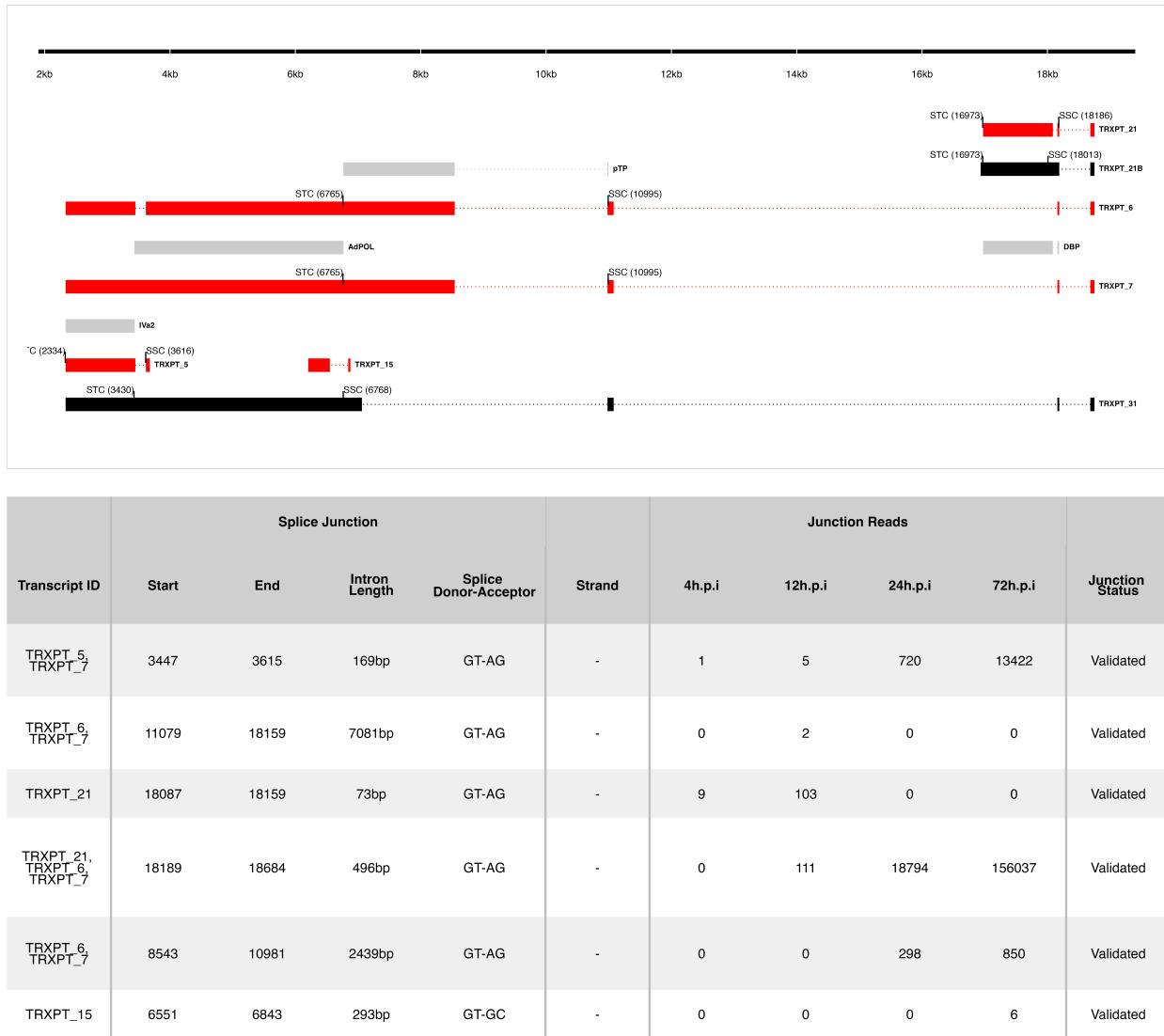


Transcript ID	Splice Junction					Strand	Junction Reads				Junction Status
	Start	End	Intron Length	Splice Donor-Acceptor			4h.p.i	12h.p.i	24h.p.i	72h.p.i	
TRXPT_4 TRXPT_4	304	1616	1313bp	GT-AG		+	0	9	1019	25041	Validated*
TRXPT_3	304	1964	1661bp	GT-AG		+	0	2	168	1588	Validated
TRXPT_2 TRXPT_1	1655	1964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

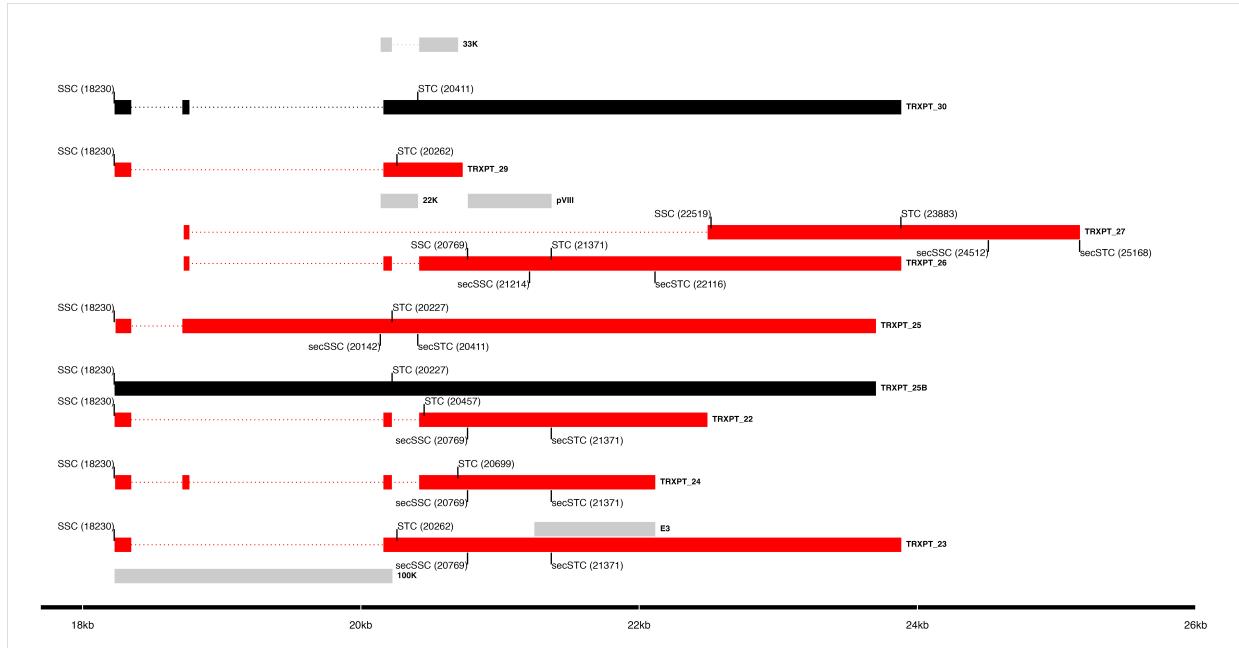
\*Not validated for TRXPT\_4

683

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

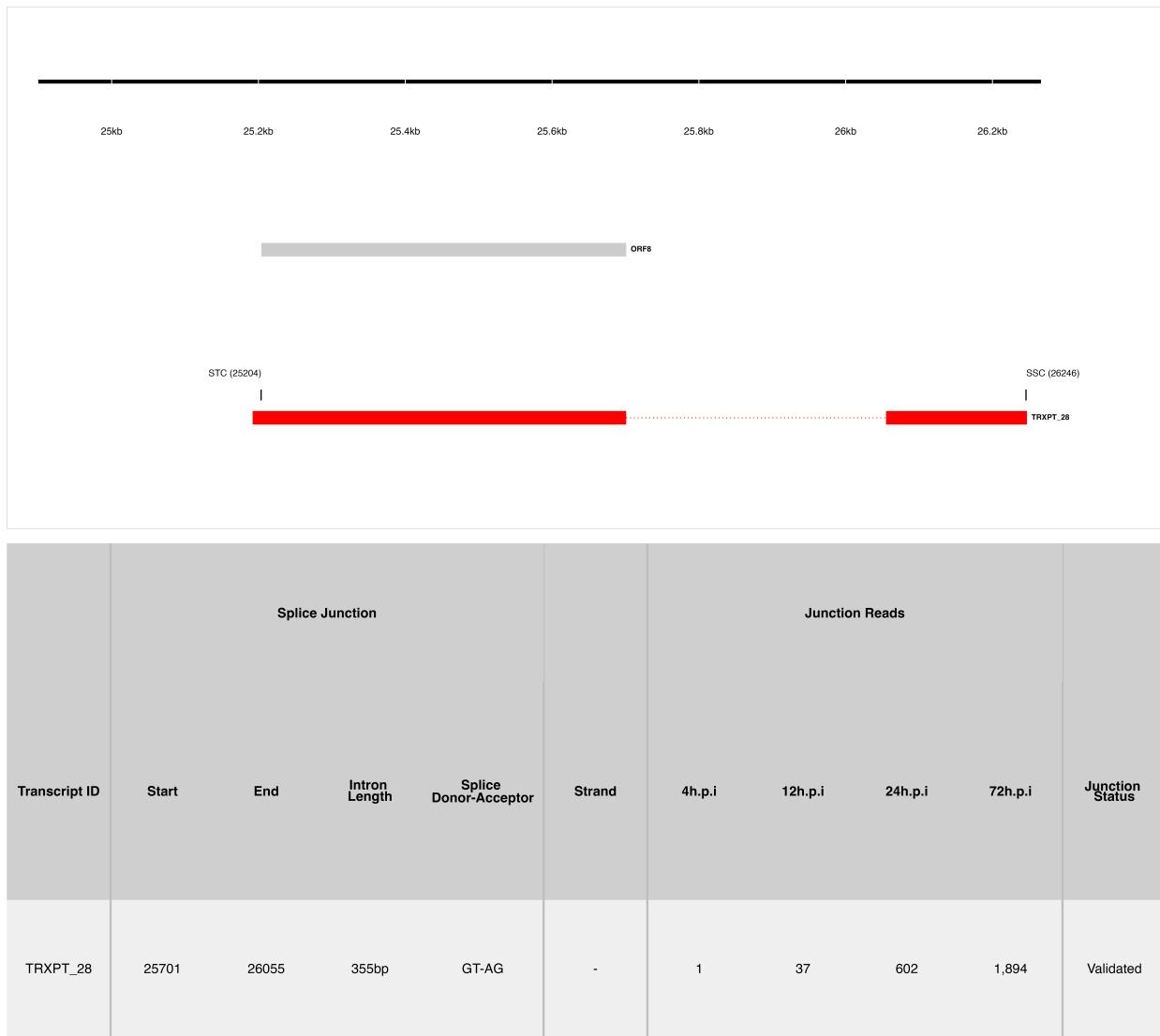


692 **Figure 7: The splice map of the E2 and IM TUs.** Exons are depicted as boxes connected by introns  
 693 (dotted lines). Red transcripts are generated from RNA-seq data and predicted ORFs are colored grey.  
 694 TRXPT\_21B discovered by 3'RACE is colored black. Each transcript or ORF is labelled with its name to  
 695 the right. The SSC and STC of the 5'-most CDS of each transcript is indicated with the nucleotide position  
 696 in brackets. The region of the virus is depicted at the bottom as a black line with labels of the nucleotide  
 697 positions for reference. The table shows sequence reads covering the splice junctions with information  
 698 about their validation status using cloning and Sanger sequencing.

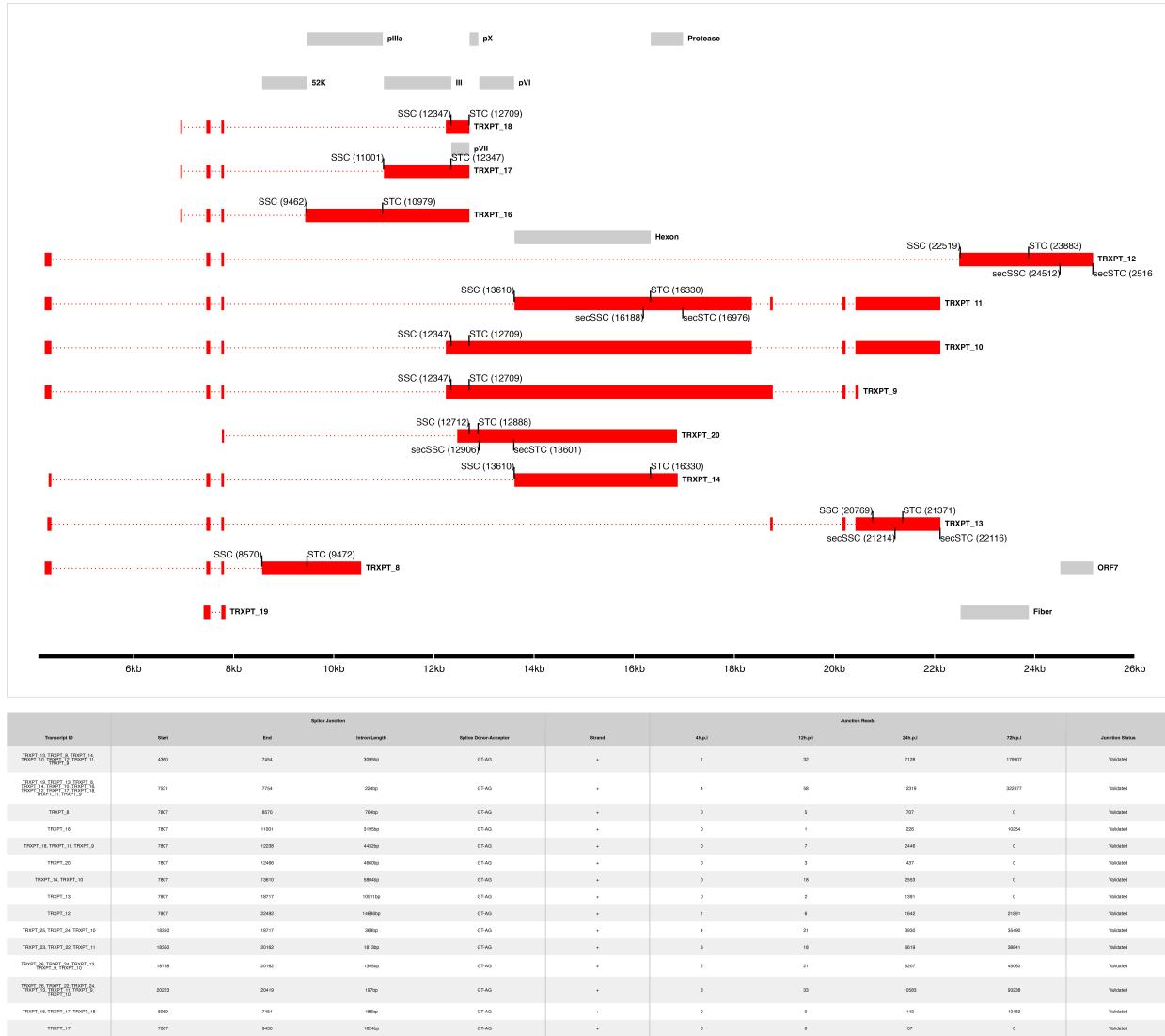


Transcript ID	Splice Junction					Junction Reads					Junction Status
	Start	End	Intron Length	Splice Donor-Acceptor	Strand	4h.p.i	12h.p.i	24h.p.i	72h.p.i		
TRXPT_25, TRXPT_24, TRXPT_10	18350	18717	368bp	GT-AG	+	4	21	3930	35490	Validated	
TRXPT_23, TRXPT_22, TRXPT_11	18350	20162	1813bp	GT-AG	+	3	18	6619	38841	Validated	
TRXPT_26, TRXPT_24, TRXPT_13, TRXPT_9, TRXPT_10	18768	20162	1395bp	GT-AG	+	2	21	5207	45062	Validated	
TRXPT_26, TRXPT_22, TRXPT_14, TRXPT_13, TRXPT_11, TRXPT_9, TRXPT_10	20223	20419	197bp	GT-AG	+	3	33	10583	93238	Validated	
TRXPT_27	18768	22492	3725bp	GT-AG	+	0	0	101	1950	Validated	

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



709 **Figure 9: The splice map of the E4 TU.** Exons are depicted as boxes connected by introns (dotted lines).  
 710 The transcript from RNA-seq data is colored red and the predicted ORF, grey. The transcript and ORF are  
 711 labelled with their names to the right. The start codon (SSC) and stop codon (STC) of the 5'-most CDS  
 712 is indicated with the nucleotide position in brackets. The region of the virus is depicted at the bottom as a  
 713 black line with labels of the nucleotide positions for reference. The table shows sequence reads covering  
 714 the splice junction with its validation status using cloning and Sanger sequencing.  
 715



716

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

Table 1: Table 1: Overview of sequencing results

Metric	4h.p.i	12h.p.i	24h.p.i	72h.p.i	Total
<b>Total reads</b>	1.17e+08	7.63e+07	1.20e+08	1.15e+08	4.28e+08
<b>Mapped (Host)</b>	1.04e+08	6.79e+07	1.06e+08	8.38e+07	3.62e+08
<b>Mapped (THEV)</b>	4.32e+02	6.70e+03	1.18e+06	1.69e+07	1.81e+07
<b>Mean Per Base Coverage/Depth</b>	2.42	37.71	6,666.96	95,041.7	101,749
<b>Total unique splice junctions</b>	13	37	236	2374	2,457
<b>Junction coverage Total (at least 1 read)</b>	37	605	115075	2132806	2.25e+06
<b>Junction coverage Mean reads</b>	2.8	16.4	487.6	898.4	351.3
<b>Junction coverage (at least 10 reads)</b>	0	13	132	1791	1,936
<b>Junction coverage (at least 100 reads)</b>	0	1	53	805	859
<b>Junction coverage (at least 1000 reads)</b>	0	0	18	168	186

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

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

Table 3: Table 2b: Most abundant splice junctions at 24h.p.i

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

Table 4: Table 2c: Most abundant splice junctions at 72h.p.i

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

<sup>724</sup> **SUPPLEMENTARY MATERIALS**

<sup>725</sup> **Supplementary Table S1A**

Table 5: 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%

<sup>726</sup> **Supplementary Table S1B**

Table 6: 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	+	29,209	25.4%
24hpi	E2	-	27,833	24.2%
24hpi	E1	+	2,724	2.4%
24hpi	Unassigned	-,+/-	1,312	1.1%

Time	Region	Strand	Total Reads	Percentage
24hpi	IM	-	744	0.6%
24hpi	E4	-	664	0.6%

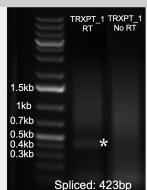
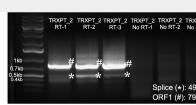
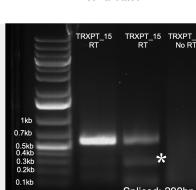
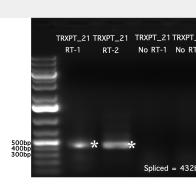
727 **Supplementary Table S1C**

Table 7: Table S1c: Most Transcriptionally Active Regions of THEV at 72h.p.i

Time	Region	Strand	Total Reads	Percentage
72hpi	MLP	+	1,437,273	67.4%
72hpi	E2	-	304,274	14.3%
72hpi	E3	+	271,392	12.7%
72hpi	E1	+	74,135	3.5%
72hpi	Unassigned	-,+/-	27,680	1.3%
72hpi	IM	-	14,484	0.7%
72hpi	E4	-	3,568	0.2%

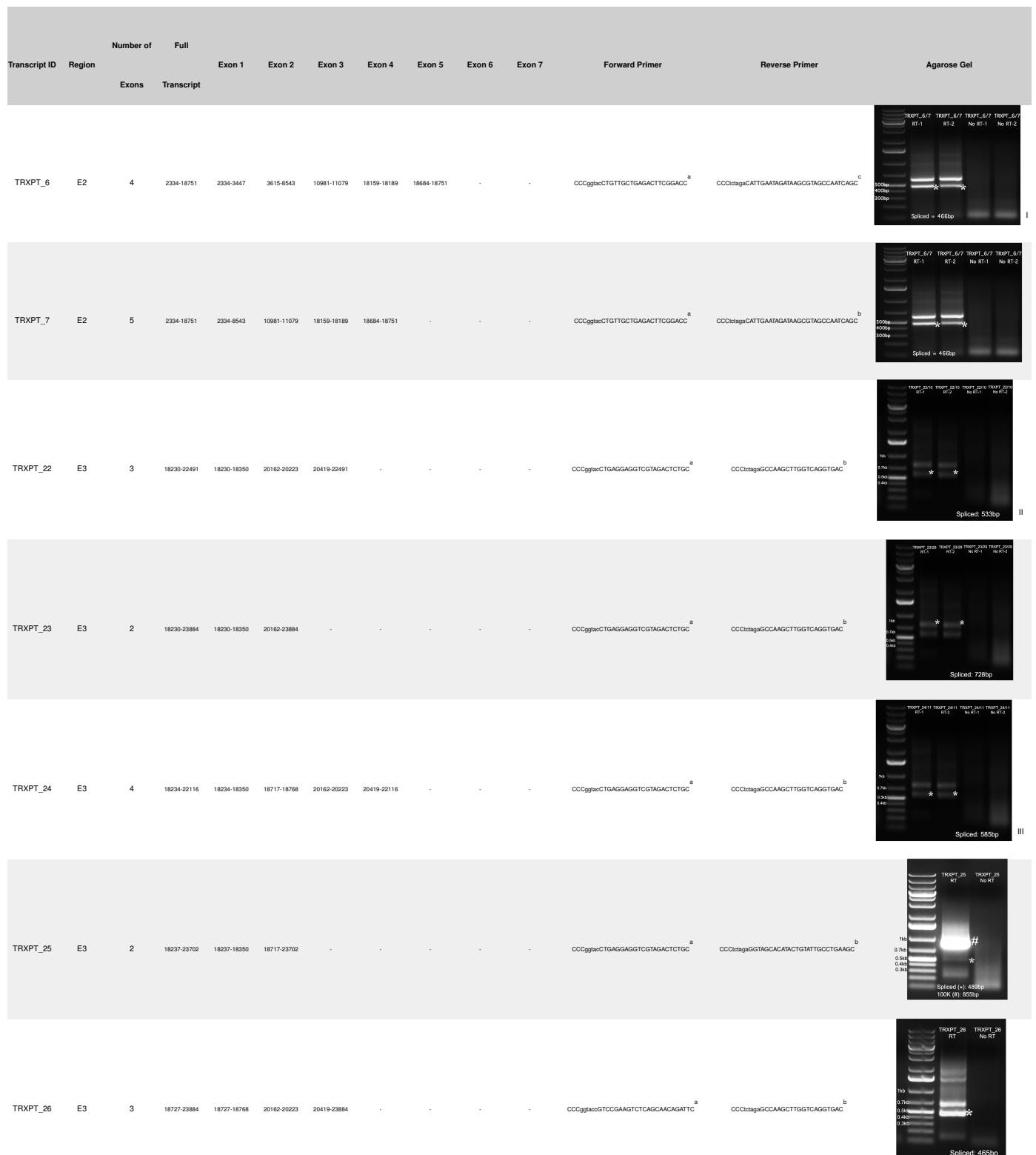
728 **Supplementary PCR Methods**

**Table 8: 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_1	E1	3	54-2325	54-304	1616-1655	1964-2325	-	-	-	CCGgtacCTCTGTTGAATGTGGGGGG <sup>a</sup>	CCCltagaCGTCAGTAGTCAGGAATTCTAGTG <sup>b</sup>	
TRXPT_2	E1	2	54-2325	54-1655	1964-2225	-	-	-	-	CCGgtacGAGGCCTGTTGGATTGTTGC <sup>a</sup>	CCCltagaCGTCAGTAGTCAGGAATTCTAGTG <sup>b</sup>	
TRXPT_3	E1	2	225-2325	225-304	1964-2325	-	-	-	-	CCGgtacCATTCCCGTACACGGGTGTG <sup>a</sup>	CCCltagaCGTCAGTAGTCAGGAATTCTAGTG <sup>b</sup>	
TRXPT_4	E1	2	271-2303	271-304	1616-2303	-	-	-	-	CCGgtacGTCATCACAACTGACCTTGTGTC <sup>a</sup>	CCCltagaCGTCAGTAGTCAGGAATTCTAGTG <sup>b</sup>	Not Validated
TRXPT_15	E2	2	6206-6878	6206-6551	6843-6878	-	-	-	-	CCGgtacCCTTAAAATCAAGCCTATGGCTGTAAAC <sup>a</sup>	CCCltagaGTGTCATTGTCACGCTGTAGTAG <sup>b</sup>	
TRXPT_21	E2	3	16973-18751	16973-18087	18159-18189	18684-18751	-	-	-	CCGgtacCTGTTCTGAGACTTCGGACC <sup>a</sup>	CCCltagaGAACCCAGATATTGGCTCCAAGG <sup>b</sup>	
TRXPT_31	E2	4	2334-18751	2334-7062	10981-11079	18159-18189	18684-18751	-	-	CCGgtacCTAGTGGCAGTGTCAAGAGTCC <sup>a</sup>	CCCltagaCATGCCAGGTATAATTGCGGGAGTAG <sup>b</sup>	

<sup>a</sup> Primer binds inside first exon; <sup>b</sup> Primer binds inside terminal exon; <sup>c</sup> Primer binds inside fourth exon; <sup>d</sup> Agarose gel identical to TRXPT\_7 due to identical splicing; <sup>e</sup> Agarose gel identical to last 3 exons of TRXPT\_10 due to identical splicing; <sup>f</sup> Agarose gel identical to last 4 exons of TRXPT\_11 due to identical splicing; <sup>g</sup> Agarose gel identical to TRXPT\_23 due to identical splicing; <sup>h</sup> Agarose gel identical to TRXPT\_9 due to identical splicing; <sup>i</sup> Agarose gel identical to TRXPT\_14 due to identical splicing;

<sup>IV</sup> due to identical splicing; <sup>V</sup> Agarose gel identical to TRXPT\_23 due to identical splicing; <sup>VI</sup> Agarose gel identical to TRXPT\_9 due to identical splicing; <sup>VII</sup> Agarose gel identical to TRXPT\_14 due to identical splicing;

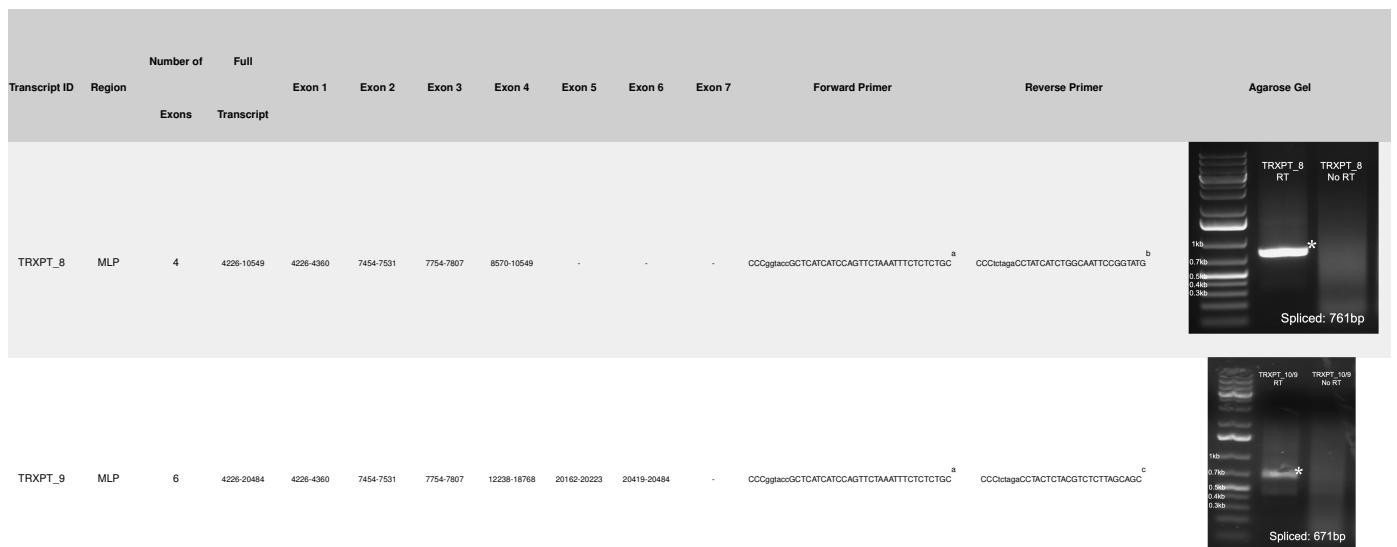


Transcript ID	Region	Number of Exons		Full Transcript							Forward Primer	Reverse Primer	Agarose Gel
		Exon	Transcript	Exon 1	Exon 2	Exon 3	Exon 4	Exon 5	Exon 6	Exon 7			
TRXPT_27	E3	2	18727-25168	18727-18768	22492-25168	-	-	-	-	-	CCGgtaccGTCGAAGTCTCAGAACAGATTC <sup>a</sup>	CCGttagaTGCAATGCTATCCCTCGCTG <sup>b</sup>	
TRXPT_29	E3	2	18230-20732	18230-18350	20162-20732	-	-	-	-	-	CCGgtaccCTGAGGAGGTGTAGACTCTGC <sup>a</sup>	CCGttagaGCCAAGCTGGTCAGGTGAC <sup>b</sup>	
TRXPT_28	E4	2	25192-26247	25192-25701	26055-26247	-	-	-	-	-	CCGgtaccGGACACGTGTTCGTAGAGAAAC <sup>a</sup>	CCGttagaCAGTGCAATCCGACGGCTG <sup>b</sup>	
TRXPT_5	IM	2	2334-3678	2334-3447	3615-3678	-	-	-	-	-	CCGgtaccCTGGTGAGATCTCCAAACAGAAAG <sup>a</sup>	CCGttagaCGCAACCTGTAGGTCGATTAC <sup>b</sup>	
TRXPT_10	MLP	7	4226-22116	4226-4360	7454-7531	7754-7807	12238-18350	20162-20223	20419-22116	-	CCGgtaccGCTCATATCCAGTTCTAAATTCTCTCTGC <sup>a</sup>	CCGttagaGCTACTCTAAGTCTCTTACAGCAGC <sup>c</sup>	
TRXPT_11	MLP	6	4226-22116	4226-4360	7454-7531	7754-7807	13610-18350	18717-18768	20162-20223	20419-22116	CCGgtaccGCTCATATCCAGTTCTAAATTCTCTCTGC <sup>a</sup>	CCGttagaGCTTCAGTATTAGCAGCTGCACAACC <sup>c</sup>	
TRXPT_12	MLP	4	4226-25168	4226-4360	7454-7531	7754-7807	22492-25168	-	-	-	CCGgtaccGCTCATATCCAGTTCTAAATTCTCTCTGC <sup>a</sup>	CCGttagaTTTCCAGCTGAACCTGGAG <sup>b</sup>	

<sup>a</sup> Primer binds inside first exon; <sup>b</sup> Primer binds inside terminal exon; <sup>c</sup> Primer binds inside fourth exon; <sup>I</sup> Agarose gel identical to TRXPT\_7 due to identical splicing; <sup>II</sup> Agarose gel identical to last 3 exons of TRXPT\_10 due to identical splicing; <sup>III</sup> Agarose gel identical to last 4 exons of TRXPT\_11 due to identical splicing; <sup>IV</sup> Agarose gel identical to TRXPT\_23 due to identical splicing; <sup>V</sup> Agarose gel identical to TRXPT\_9 due to identical splicing; <sup>VI</sup> Agarose gel identical to TRXPT\_14 due to identical splicing;

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	-	CCGgtaccGCTCATCACAGTTCTAAATTCTCTCTGC <sup>a</sup>	CCCltagaGCCAAGCTGGTCAGGTGAC <sup>b</sup>	
TRXPT_14	MLP	4	4304-16870	4304-4360	7454-7531	7754-7807	13610-16870	-	-	-	CCGgtaccGCTCATCACAGTTCTAAATTCTCTCTGC <sup>a</sup>	CCCltagaGCCAGCTTCACTGACAGCTGACACC <sup>b</sup>	
TRXPT_16	MLP	4	6934-12709	6934-6969	7454-7531	7754-7807	9430-12709	-	-	-	CCGgtaccGGATCTCCAGATTCTGGTCTGTG <sup>a</sup>	CCCltagaGCCTGTCCAACACCTGC <sup>b</sup>	
TRXPT_17	MLP	4	6934-12709	6934-6969	7454-7531	7754-7807	11001-12709	-	-	-	CCGgtaccGGATCTCCAGATTCTGGTCTGTG <sup>a</sup>	CTCCCCATCTAGACCTTTCATCTAAC <sup>b</sup>	
TRXPT_18	MLP	4	6934-12709	6934-6969	7454-7531	7754-7807	12238-12709	-	-	-	CCGgtaccGGATCTCCAGATTCTGGTCTGTG <sup>a</sup>	CCCltagaGTTCTCGTCTTACGTCGTC <sup>b</sup>	
TRXPT_19	MLP	2	7401-7836	7401-7531	7754-7836	-	-	-	-	-	-	-	N/A
TRXPT_20	MLP	2	7765-16856	7765-7807	12466-16856	-	-	-	-	-	CCGgtaccGAGGATTGAAGCCAATTCCCTCAACG <sup>a</sup>	CCCltagaCTGCAGGCACACAGGTG <sup>b</sup>	

<sup>a</sup> Primer binds inside first exon; <sup>b</sup> Primer binds inside terminal exon; <sup>c</sup> Primer binds inside fourth exon; <sup>d</sup> Agarose gel identical to TRXPT\_7 due to identical splicing; <sup>e</sup> Agarose gel identical to last 3 exons of TRXPT\_10 due to identical splicing; <sup>f</sup> Agarose gel identical to last 4 exons of TRXPT\_11 due to identical splicing; <sup>g</sup> Agarose gel identical to TRXPT\_23 due to identical splicing; <sup>h</sup> Agarose gel identical to TRXPT\_9 due to identical splicing; <sup>i</sup> Agarose gel identical to TRXPT\_14 due to identical splicing;



<sup>a</sup>Primer binds inside first exon; <sup>b</sup>Primer binds inside terminal exon; <sup>c</sup>Primer binds inside fourth exon; <sup>I</sup>Agarose gel identical to TRXPT\_7 due to identical splicing; <sup>II</sup>Agarose gel identical to last 3 exons of TRXPT\_10 due to identical splicing; <sup>III</sup>Agarose gel identical to last 4 exons of TRXPT\_11 due to identical splicing; <sup>IV</sup>Agarose gel identical to TRXPT\_23 due to identical splicing; <sup>V</sup>Agarose gel identical to TRXPT\_9 due to identical splicing; <sup>VI</sup>Agarose gel identical to TRXPT\_14 due to identical splicing;

729 In the table above, the restriction sites in the primer tails are shown in lowercase letters. All the primer  
 730 melting temperatures (TMs) are 58-60°C using a hot start Taq DNA polymerase. The PCR reaction mix  
 731 was done per manufacturer's instructions. The PCR cycling conditions were as follows: Initial denaturation  
 732 – 95°C for 1 minute; cyclical denaturation – 95°C for 30 seconds, annealing – variable temperature (53°C-  
 733 56°C) for 30 seconds, primer extension – 68°C for variable time, and final elongation – 68°C for 5 minutes.  
 734 We used 35 cycles of amplification.

### 735 Supplementary Computational Analysis

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

744 StringTie was also used to estimate the normalized expression levels (FPKM) of all the transcripts and  
 745 Ballgown in R was used to perform statistical analysis and comparisons of the transcript expression levels,

746 which instructive in understanding the temporal regulation THEV gene expression.

747 In these steps above, each sample (replicate of each time point) was processed independently and merged  
748 only in the final transcriptome assembly or during analysis with `Ballgown`. In the subsequent steps de-  
749 scribed below, all samples for each time point were processed together.

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

