Conservation of the E8 CDS of the E8E2 protein among mammalian papillomaviruses

Mikk Puustusmaa¹ and Aare Abroi²

¹Department of Bioinformatics, University of Tartu, Riiia 23, Tartu 51010, Estonia
²Estonian Biocentre, Riiia 23, Tartu 51010, Estonia

Papillomaviridae are small dsDNA viruses with a limited coding capacity. To fulfill all of the functional requirements for propagation and spreading, papillomaviruses use double coding and alternative protein isoforms. E8 E2 is an alternative E2 protein isoform that is generated by fusing the short E8 CDS that completely overlaps E1 to the ‘hinge’ and the DNA-binding region of the E2 protein via alternative transcription/splicing. The papillomaviruses in which E8 E2 mRNA sequences have been described exhibit a sparse phylogenomic distribution. Thus, it is not clear whether E8 E2 is an ancestral protein that has not been described for other papillomavirus types or whether it randomly appears because of the conservation of the E1 protein and occurs only coincidentally. We searched for potential E8 coding sequences in a non-redundant set of papillomaviruses and applied SynPlot2 and an in-house-developed algorithm (cRegions) to determine the most plausible of the above-mentioned scenarios. Beginning with nine experimentally described E8 E2 mRNAs, we predicted the potential E8 CDSs for more than 300 mammalian papillomavirus genomes. According to our analysis, E8 E2 is not a result of E1 coding and represents a protein in its own right, and it most likely has an ancestral origin that precedes the divergence of major mammalian papillomavirus genera.

INTRODUCTION

The cottontail rabbit papillomavirus (CRPV, also referred to as Shope PV) is a member of the family Papillomaviridae and was the first identified DNA tumour virus. It was discovered as early as 1932. PVs are species specific, and different members of the family Papillomaviridae most likely productively infect almost all mammalian species (including marine mammals and bats). Furthermore, papillomaviruses have been found in several sauropsids (birds and reptiles), although none of these viruses have been identified in invertebrates (Rector & Van Ranst, 2013). PVs have a circular dsDNA genome with a size between ~6900 bp and ~8600 bp (according to the PaVE database on 2 February 2016; http://pave.niaid.nih.gov/). The PV genome encodes four core genes (E1, E2, L1 and L2) and various oncogenes and regulatory proteins. Extensive studies on the molecular biology of papillomaviruses began after 1980, when Dvoretzky and colleagues propagated the bovine papillomavirus 1 (BPV1) genome in a cell culture system (Dvoretzky et al., 1980). Since then, PVs have been used as a model system to study eukaryotic DNA replication and oncogenesis as well as different virological aspects. Many molecular details of PVs (i.e. DNA replication, oncogenesis, transcription and stable maintenance) have been studied and are well understood. Despite the existence of several systems for generating viral particles (Biryukov & Meyers, 2015), the full replication cycle of the virus, including first rounds of transcription, and the establishment of viral infection, has not yet been extensively studied using native virions.

The E2 (full-length) protein is a replication protein and a major regulator of PV gene transcription. E2 functions as a repressor or an activator of transcription, depending on the location of the E2 binding sites within the enhancer/promoter region. In addition to the full-length E2 protein, different E2 isoforms have been characterized. Several PVs express an alternatively spliced protein known as E8’E2. mRNAs capable of encoding E8’E2 are generated by splicing, fusing the E8 exon to the splice-acceptor site (3′ss) located at the beginning of the ‘hinge’ region of the E2 protein (Fig. 1). The E8 coding sequence (CDS) overlaps with the E1 CDS and is in the +1 frame relative to the E1 reading frame (defined as frame 0). As a result, the E8’E2 protein consists of an E8 peptide, the E2 ‘hinge’ region and the E2 DNA binding domain (DBD). Thus, E8’E2 is a DNA-binding protein that is able to compete with full-length E2 for binding to E2-binding sites and to form heterodimers with full-length E2 (Kurg et al., 2006, 2009; McBride et al., 1989). Disrupting the expression of the E8’E2 protein leads to an increase in viral early transcripts

Ten supplementary figures and one supplementary table are available with the online Supplementary Material.
and viral DNA replication in several PV types, indicating that E8-E2 is involved in the regulation of these processes (Isok-Paas et al., 2015; Kurg et al., 2010; Lace et al., 2008; Sankovski et al., 2014; Stubenrauch et al., 2000). Therefore, in these PV types, E8-E2 appears to be an important factor involved in the regulation of early transcription and the establishment of infection. In human papillomavirus 31 (HPV31), the E8-E2 protein represses transcription and extra-chromosomal replication, and many interactions responsible for these repressive functions have been described (Ammermann et al., 2008; Fertey et al., 2010; Powell et al., 2010). In addition to HPV31, the transcriptional repression activity of E8-E2 protein has been described for BPV1, *Sylvilagus floridanus* papillomavirus 1 (SfPV1), HPV16 and HPV5 (Choe et al., 1989; Jeckel et al., 2003; Lace et al., 2008, 2012; Sankovski et al., 2014, 2015). In BPV1, the E8 peptide of the E8-E2 protein functions as a nuclear matrix-targeting signal (NMTS) (Sankovski et al., 2015). Repression of early gene expression appears to be a common property of E8-E2 in all PVs in which this protein has been described.

Transcripts corresponding to the E8-E2 protein have been described for BPV1 (Choe et al., 1989), HPV1 (Palermo-Dilts et al., 1990), HPV5 (Sankovski et al., 2014), HPV11 (Rotenberg et al., 1989), HPV16 (Doorbar et al., 1990), HPV18 (Kurg et al., 2010), HPV31 (Stubenrauch et al., 2000), HPV33 (Snijders et al., 1992) and SfPV (Jeckel et al., 2003) [historically referred to as CRPV1 (Bernard et al., 2010)]. To the best of our knowledge, evidence of the presence of E8-E2 at the protein level in virus-transformed cells has been provided only for BPV1 (Hubbert et al., 1988; Kurg et al., 2006; Lambert et al., 1989; Yang et al., 1991). The distribution of these nine PV types with experimentally verified E8-E2 mRNAs is sparse, when they are highlighted on a phylogenetic tree (Fig. 2). This distribution may be caused by at least two different processes. First, the conserved amino acids in the E1 protein within the E8 region drive conservation of the DNA sequence, and the E8 ORF has been preserved by pure coincidence during evolution. Second, E8-E2 has an ancestral origin. However, if E8-E2 has an ancestral origin, what about more than 300 PV types without experimentally verified E8-E2 transcripts [e.g. the large, growing group of *Gamma-papillomaviruses* (Bravo & Félez-Sánchez, 2015)]? Whether all PVs encode E8-E2 as an overlapping gene; is E8-E2 caused by a selection on the E1 reading frame (and thus encodes E8), and whether E8 is a protein in its own right and essential to all other PVs, remain unknown.

In this work, an algorithm was developed to detect E8 CDSs in PV genomes. Two complementary methods [cRegions and SynPlot2 (Firth, 2014)] were applied to examine whether the coding potential of E8 was determined by the overlapping E1 sequences. Our analysis showed that the presence of the E8 CDS is not caused by the conservation of amino acids in the E1 gene. Additionally, most mammalian PVs encode E8 with clearly recognizable E8 peptide motifs.

### RESULTS

#### Predicting the E8 CDSs in the available papillomaviruses

To investigate the existence of the E8 CDS in PV genomes, bioinformatics analysis was performed on a set of over 300 PV genomes. The final dataset used for the E8 prediction consisted of 318 different PV genomes, including the types from Table 1, in which the E8-E2 mRNA was previously described and confirmed experimentally. Using our search algorithm (see the Methods section), potential E8 CDSs were found in 308 PV genomes (Fig. 2, Table S1, available in the online Supplementary Material). All of the PVs from Table 1, which

---

![Fig. 1. Human papillomavirus 16 (HPV16) E1 protein-reading frame and adjacent reading frames. Two splice donor sites (5'ss) within the E1 reading frame are depicted. The first is E1-E4 5'ss at 880 nt, and the second is E8-E2 5'ss at 1302 nt. The E8 dual-coding region is located from 1265 nt to 1302 nt in the +1 reading frame with respect to E1. The E2 reading frame is in the +1 reading frame with respect to E1. (Based on NC_001526.2 and PaVE HPV16REF).](image-url)
functioned as a positive control group, had correct E8 CDS predictions (Table S1). Based on the experimentally confirmed data in Table 1, AG|GTA was preferred as the 5’ss sequence, and the first hit was always preferred. Subsequent analyses showed that the selected restrictions were wide enough and the lengths of all of the detected potential E8 CDSs were between 21 and 59 nt (Fig. S2, search criteria 17 and 62, respectively). The E8 initiation codon was between positions 305 and 420 with respect to the E1 initiation codon (Fig. S3, search criteria 300 and 800, respectively).

Potential E8 CDSs were not detected in 10 PV types. These 10 PV types included six sauropsids’ PVs: CcPV1 (loggerhead

Fig. 2. Described and predicted E8 CDSs in an E1-based phylogenetic tree. The blue squares represent PVs in which the E8E2 mRNA sequence has been experimentally confirmed (for details, see Table 1). Green arcs of the circle represent PVs in which a potential E8 CDS was detected using the method described in this work. Red sections represent PVs in which a potential E8 CDS was not detected, and yellow squares represent PVs in which a predicted potential E8 CDS was questionable in the current study. Orange arcs show PVs that encode E1’E4 (based on the PaVE database). Each colored full clade (sector) represents a group of PVs with similar putative E8 peptide sequences. The E1 protein alignment was performed with MUSCLE v3.8.31 (default settings, --refine option). The analysis involved 318 E1 amino acid sequences. Evolutionary analyses were conducted in MEGA6 (Tamura et al., 2013) (maximum likelihood method, bootstrap 100). The tree was visualized in iTOL (Letunic & Bork, 2007) and is available at http://itol.embl.de/tree/193401210654614521505680.
**Table 1.** Summary of the published experimental data on E8-E2 mRNAs

<table>
<thead>
<tr>
<th>PV</th>
<th>E1 atg*</th>
<th>E1 atg†</th>
<th>E8 atg-E1</th>
<th>E8 5’ss‡</th>
<th>E8 length (nt)§</th>
<th>E8 CDS frame</th>
<th>5’ss seq#</th>
<th>E8-E2 3′ss#</th>
<th>E2 atg**</th>
<th>E2 CDS frame††</th>
<th>E1-E4 3′ss‡‡</th>
<th>Taxonomy ID§§</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>SfPV1</td>
<td>1362</td>
<td>1720</td>
<td>358</td>
<td>1751</td>
<td>32</td>
<td>+1</td>
<td>CAG</td>
<td>GTA</td>
<td>3714</td>
<td>3112</td>
<td>+1</td>
<td>3714</td>
<td>10623</td>
</tr>
<tr>
<td>BPV1</td>
<td>849</td>
<td>1204</td>
<td>355</td>
<td>1235</td>
<td>32</td>
<td>+1</td>
<td>CAG</td>
<td>GTA</td>
<td>3225</td>
<td>2608</td>
<td>+1</td>
<td>3225</td>
<td>10559</td>
</tr>
<tr>
<td>HPV1a</td>
<td>812</td>
<td>1200</td>
<td>388</td>
<td>1231</td>
<td>32</td>
<td>+1</td>
<td>CAG</td>
<td>GTA</td>
<td>3200</td>
<td>2637</td>
<td>+1</td>
<td>3200</td>
<td>10583</td>
</tr>
<tr>
<td>HPV5</td>
<td>961</td>
<td>1328</td>
<td>367</td>
<td>1359</td>
<td>32</td>
<td>+1</td>
<td>GAG</td>
<td>GTA</td>
<td>3322</td>
<td>2723</td>
<td>+1</td>
<td>3322</td>
<td>333923</td>
</tr>
<tr>
<td>HPV11</td>
<td>832</td>
<td>1241</td>
<td>409</td>
<td>1272</td>
<td>32</td>
<td>+1</td>
<td>CAG</td>
<td>GTA</td>
<td>3325</td>
<td>2723</td>
<td>+1</td>
<td>3325</td>
<td>10580</td>
</tr>
<tr>
<td>HPV16</td>
<td>858</td>
<td>1265</td>
<td>407</td>
<td>1302</td>
<td>38</td>
<td>+1</td>
<td>CAG</td>
<td>GTA</td>
<td>3358</td>
<td>2756</td>
<td>+1</td>
<td>3358</td>
<td>333760</td>
</tr>
<tr>
<td>HPV18</td>
<td>914</td>
<td>1323</td>
<td>409</td>
<td>1357</td>
<td>35</td>
<td>+1</td>
<td>CAG</td>
<td>GTA</td>
<td>3434</td>
<td>2817</td>
<td>+1</td>
<td>3434</td>
<td>333761</td>
</tr>
<tr>
<td>HPV31</td>
<td>862</td>
<td>1259</td>
<td>397</td>
<td>1296</td>
<td>38</td>
<td>+1</td>
<td>CAG</td>
<td>GTA</td>
<td>3295</td>
<td>2693</td>
<td>+1</td>
<td>3295</td>
<td>10585</td>
</tr>
<tr>
<td>HPV33</td>
<td>879</td>
<td>1276</td>
<td>397</td>
<td>1316</td>
<td>41</td>
<td>+1</td>
<td>CAG</td>
<td>GTA</td>
<td>3351</td>
<td>2749</td>
<td>+1</td>
<td>3351</td>
<td>10586</td>
</tr>
</tbody>
</table>

*Location of the E1 initiation codon (atg) in the PV genome (the number shows the position of the A nucleotide in the PV genome).
†Location of the E8 initiation codon in the PV genome (the number shows the position of the A nucleotide in the PV genome).
‡Location of the E8-E2 splicing-donor site (5′ss) in the PV genome (the number shows the position of the G nucleotide in the PV genome, which is the last nucleotide in exon CAGGTA).
§Length of the E8 CDS (shown in nucleotides).
||Reading frame of the E8 CDS with respect to the E1 reading frame (the E1 reading frame is considered the 0 frame).
¶Splicing donor of E8-E2.
#Location of the E8-E2 splicing-acceptor site (3′ss) in the PV genome.
**Location of the E2 initiation codon in the PV genome (the number shows the position of the A nucleotide in the PV genome).
††E4 reading frame (E1-E4) with respect to E2 (the E2 reading frame is considered the 0 frame).
‡‡E1-E4 splicing-acceptor site in the PV genome.
§§NCBI taxonomy identification number.
turtle PV1), CmPV1 (green sea turtle PV1), FcPV1 (common chaffinch PV1), FlPV1 (yellow-necked spurfowl PV1), PaPV1 (Adélie penguin PV1) and PePV1 (African grey parrot PV1).

In our phylogenetic tree based on the E1 protein sequences (Fig. 2) and in previously published trees based on L1 nucleotide sequences (http://PaVE.niaid.nih.gov/#explore/phylogenetic_tree) or concatenated E1/E2/L1/L2 sequences (Bravo & Félez-Sánchez, 2015), all of these sauropsid PV types clustered together to form a monophyletic clade with high confidence. However, a potential E8 was predicted in the E1 gene of MsPV1 (diamond python PV1). The putative E8 peptide was more similar to a mammalian PV than it was to other sauropsid PVs. The remaining four PV genomes without a detected E8 were RferPV1 (greater horseshoe bat PV1), HPV41, EdPV1 (North American porcupine PV1) and EhelPV1 (straw-coloured fruit bat PV1). Three PVs (HPV41, EdPV1 and EhelPV1) clustered together in the E1-based phylogenetic tree (Fig. 2). Interestingly, the E8 CDS was located in the +1 reading frame with respect to E1 in all 308 PV genomes in which it was found (both experimental and predicted; Table S1).

**Splice-site preferences in different genera**

A short summary of the predicted E8 5’ss are shown in Table 2. For all of the tested PV genomes, the 5’ss deviated from AG|GTA in eight PVs. The 5’ss was AG|GTT in TtPV2 (Tursiops truncatus PV2), BPV7 and HPV123 and AG|GTG in MfPV7 (Macaca fascicularis PV7), TmPV2 (Trichechus manatus latirostris PV2), OaPV3 (Ovis aries PV3) and RrupPV1 (Rupicapra rupicapra PV1) and northern fulmar PV1 (FgPV1).

**E8 CDS conservation**

The bioinformatics tool cRegions was developed to distinguish whether the conservation of amino acids in certain parts of a protein was facilitated by selection at the amino acid level or other mechanisms (i.e. dual-coding areas) at the nucleic acid level (the working principle of cRegions is illustrated in Fig. 3). Both cRegions and SynPlot2, developed by Dr Firth (2014), were used to assess whether the E8 coding sequence is a simple incidental consequence of selection at the protein level on E1, or is a protein in its own right. SynPlot2 has been previously used in a similar analysis of RNA viruses (Firth, 2014). Both tools were designed to find regions of overlapping or embedded functional elements within protein coding genes. The programs had the following similar limitations: the position of functional elements must be conserved with respect to the amino acid sequence; a sufficient number of sequences is required (for higher confidence); amino acids at these positions (or at least several of them) must have redundant codons and the alignment must be confidential, and the embedded element must be conserved in these sequences. To better meet these criteria, each papillomavirus genus in this work was analysed separately.

The data presented in Fig. 4 show that there are four distinctive signals in the E1 protein coding sequence of **Alphapapillomaviruses** where the nucleotide sequence is more conserved

| PV* | Total† | NAG|GTA‡ | NAG|GTG§ | NAG|GTC|| NAG|GTT¶ | Not detected# | 5’ss** |
|-----|--------|-------|-------|-------|-------|-------|-------|--------|-------|
| AlphaPV | 77 | 76 | 1 | — | — | — | — | 66 CAG|GTA |
| BetaPV | 49 | 49 | — | — | — | — | — | 49 GAG|GTA |
| GammaPV | 65 | 64 | — | — | 1 | — | — | 56 CAG|GTA |
| DeltaPV | 14 | 14 | — | — | — | — | — | 14 CAG|GTA |
| LambdaPV | 11 | 11 | — | — | — | — | — | 11 CAG|GTA |
| ChiPV | 11 | 11 | — | — | — | — | — | 11 CAG|GTA |
| Others* | 91 | 75 | 4 | — | 2 | 10 | — | 91 CAG|GTA |

*Papillomaviridae* genus.
†Total count of PV types for which the E8 CDS was predicted.
‡, §, ||, ¶Number of 5’ss in the predicted E8 CDSs.
#Number of PV types for which the E8 CDS was not detected.
**5’ss sequences in the predicted E8 CDSs.
than expected from the amino acid frequencies (Fig. 4). The first signal is located at 15 nt to 24 nt from the E1 start in the multiple-sequence alignment (MSA) (e.g. 877 to 883 in HPV16) and corresponds to the splice-donor site used to produce the E1–E4 transcript. The second signal extends from 308 nt to 319 nt (e.g. 1122 to 1133 in HPV16) and is a conserved region that was described by Campione-Piccardo et al. (1991), corresponding to the highly conserved amino acid motif LKRK (Leu-Lys-Arg-Lys), which functions in the protein as one part of the bipartite nuclear localization signal (Bergvall et al., 2013). Recently, this DNA region (cis-element) was shown to be an important part of the E8 promoter (at least in HPV16) (Straub et al., 2015). The third signal extends from 473 nt to 522 nt (e.g. 1265 to 1302 in HPV16) and corresponds to the dual-coding region of E8. The last signal is the overlap of the E1 and E2 genes, which is approximately 60 nt in length and extends from 2138 nt to 2195 nt (e.g. 2756 to 2814 in HPV16), together with the region preceding the E1–E2 overlap. The E1–E2 overlap region serves as a positive control for the applied method.

All four elements (the 5’ss at the beginning of E1, the potential E8–E2 promoter, the E8 CDS and the E1–E2 overlap) were observed in other PV genera (Figs S4–S8). Both cRegions and SynPlot2 recognized the E8 region as a potential double coding element in all genera with more than 10 PV types (i.e. for alpha, beta, gamma, delta, lambda and chiPVs). Thus, the E8 CDS is in its own right and not a result of conservation of the E1 sequences.

**Detection of other overlapping functional elements**

Closer inspection of all of the analysed PV genera revealed many clear signals with currently unknown functions. BetaPVs had many high-signal values prior to the conserved region described by Campione-Piccardo (Fig. S4). An interesting signal was found by both programs in the Delta genus before the E1–E2 overlap (Fig. S5). To the best of our knowledge, the P2443 promoter is located in the region just upstream of the E2 ORF in BPV1 at least, which is a DeltaPV (Hermonat et al., 1988). Additionally, a signal was detected in the Gamma genus after the E1–E4 splicing-donor site and before the known conserved region described by Campione-Piccardo. The consensus sequence in this region is CAGGGAATTCCCTGGC (the logo is shown in Fig. S9).

**Three distinct E8 groups**

An MSA was generated for all predicted E8 peptides (Table S1; including sequences from Table 1). The results showed that the E8 peptides could be classified into three distinct groups. First, the ‘Alpha’ group consisted of all
known Alphapapillomaviruses and three Saimiri sciureus PVs from the Dyoomikron genus (SscPV1, SscPV2, and SscPV3). Two main features that distinguished this group from the other groups were the highly conserved alanine (A) and isoleucine (I) residues in the second and third positions, respectively, and the tryptophan (W) in the sixth position (Fig. 5a). Second, the ‘Cetacea’ group was the smallest and primarily consisted of PV types that infect Cetacea (Rector & van Ranst, 2013). In addition to Omikronpapillomavirus and Upsilonpapillomaviruses, one Dyodelta papillomavirus (PphPV4, harbour porpoise PV4) was present that also infects Cetacea (Rector & van Ranst, 2013). This group has two peculiar members: one from the Dyodelta genus SsPV1 (domestic pig PV1) and the second from the Omega genus UmPV1 (polar bear PV1). This group has a specific motif, GIWHK, and a second highly conserved tryptophan at position 13 (Fig. 5b). Group ‘Cetacea’ had the longest E8 peptide of the groups (median group length of 18 versus 12 for group ‘Alpha’ and 10 for group ‘Diverse majority’). Third, the ‘Diverse majority’ group was the largest and consisted of the remaining PVs (Fig. 5c). In the consensus sequence, there was a highly conserved lysine (K) at the second position and a leucine (L) at the third position, in addition to the characteristic ‘KLL’ motif in the centre. Three predicted E8 peptides in this group were truncated before generating the logo (Fig. 4).

http://jgv.microbiologyresearch.org 2339

Fig. 4. Overlapping functional elements in the E1 CDS of Alphapapillomaviruses. (a) cRegions analysis of the arithmetic mean of the RMSD statistical values over codon position three, with a sliding-window size of three. (b) SynPlot2 results using a three-codon sliding window. Alphapapillomaviruses E1 protein sequences were aligned with MUSCLE, and the –refine option was used. The corresponding DNA alignment was performed with Pal2Nal. The data in the figures show four distinctive signals where the nucleotide sequence is more conserved than expected from the amino acid frequencies.
Three E8 groups are homologues

The E8 peptide sequence of group ‘Alpha’ and group ‘Cetacea’ had common motifs, whereas the ‘Diverse majority’ group did not show significant similarity to the other two groups. To examine the homologous nature of these sequences, a less biased set of 57 PV sequences was used (see the Methods section, full alignment in http://bioinfo.ut.ee/cRegions/manuscript/supplementary_alignments/MAFFT_PAL2NAL_E1_57.fas). The predicted E8 CDSs were highlighted on the nucleotide MSA, based on the E1 amino acid alignment. As shown in Fig. S10, all of the E8 sequences were coded in the same regions with respect to the E1 alignment. Group ‘Alpha’ and group ‘Cetacea’ had start codons at the same position, whereas the ‘Diverse majority’ group’s start codons were shifted by a few codons. Similar results were obtained when the initial protein alignment was performed using MUSCLE. In most of the viruses shown in Fig. S10, the splice-consensus site was also well aligned. Thus, these three E8 groups most likely have a common ancestor.

DISCUSSION

In our study, starting from nine experimentally characterized E8’E2 mRNA sequences, the E8 CDSs and E8 5’ss sites were predicted for more than 300 PV types. According to the PaVE database, all of these PVs encoded E1’E4 [except for MpvPV1 (Mustela putorius PV1)] (Fig. 1). This observation is important because all of the experimentally verified E8’E2 transcripts used the same 3’ss as E1’E4. Therefore, almost all of the PV types have the potential to encode E8’E2. However, no potential E8 CDS was predicted for 10 PVs, and the
prediction for one PV type was most likely a false positive. As shown in Fig. 2, nine of the E8-deficient PV types formed two specific clades on the phylogenetic tree. One clade contained seven PV types isolated from sauropsids (reptiles and birds). This clade most likely diverged very early in papillomavirus evolution. None of these seven PV types had annotated E1‘E4 splice sites in the PaVE database or the conserved E8 promoter element TAAAGCGAAGT (starting position 566 in http://bioinfo.ut.ee/cRegions/manuscript/ supplementary_alignments/E1_318PV_mafft_Pal2Nal.fas), although CcPV1 and CmPV1 contained the conserved amino acid motif KRRL (Lys-Arg-Arg-Leu, starting position 190 in the http://bioinfo.ut.ee/cRegions/manuscript/supplementary alignments/E1_protein_318PV_mafft.fas). None of the E8 CDSs predicted with the default parameters reached the borders of the used parameters (Figs S2 and S3). To obtain reasonable assurance that the apparently E8-deficient PV types did not encode E8 elsewhere in the E1 gene, another search was performed with less strict parameters (the ‘aGgt-Atg/3 residual’ restriction was also removed). None of the identified short ORFs had any resemblance to any other experimentally confirmed or predicted E8 CDSs. Thus, it is not correct to call them E8 (or E8 homologues) without experimental verification. Using the default parameters, a potential E8 was found in one PV type within an E8-deficient clade (FgPV1) (Fig. 2) with the peptide sequence MGRETQGPESASAMAAPK. This E8 peptide did not fit into any of the three E8 groups, and FgPV1 did not have an annotated E1‘E4 splice site in the PaVE database. Thus, the predicted E8 CDS of FgPV1 was most likely a false positive, and experimental analyses must be performed to confirm its presence. Due to the long evolutionary distance (assuming a host–virus coevolution probable divergence time of ~315 million years ago) and the evolutionary rate of the PV E1 protein (1.76 × 10⁻⁸ substitutions/nt/year for Lambda-papillomaviruses, 7.1 × 10⁻⁹ substitutions/nt/year for mammalian PVs and 1.1 × 10⁻⁶ substitutions/nt/year for non-mammalian anmioite PV) (Herbst et al., 2009; Rector & Van Ranst, 2013; Rector et al., 2007; Shah et al., 2010), the fact that an E8 has not been found does not necessarily mean that these PV types do not have a functional E8 analogue.

All of the predicted putative E8 peptides can be divided into three groups based on their E8 peptide MSA. Group ‘Alpha’ (sequence logo in Fig. 5a) consists of Alphapapillomaviruses plus three PVs from the genus Dyoornikronpapillomavirus. The genus Dyoornikronpapillomavirus is the closest relative of AlphaPVs in our E1 protein-based phylogenetic tree (Fig. 2, light blue) and the L1 nucleotide-based tree (http://PaVE.niaid.nih.gov/#explore/phylogenetic_tree); these viruses were isolated from the cervix of the common squirrel monkey. Additionally, a conserved KWK motif (Lys-Trp-Lys) occurs in this group (Fig. 5a residue positions 5–7), which is important for the function of E8‘E2. The KWK motif is necessary for both the repression of transcription and the repression of the viral replication, at least in HPV16 and HPV31 (Straub et al., 2014; Stubenrauch et al., 2001; Zobel et al., 2003). Group ‘Cetacea’ (sequence logo in Fig. 5b) was the smallest group and primarily included papillomaviruses that infect cetacea. All of the PVs in this group were located in the same branch of our E1-based phylogenetic tree (Fig. 2, light purple); however, in the L1-based tree, these PVs did not form a monophyletic clade (http://PaVE.niaid.nih.gov/#explore/phylogenetic_tree). The group ‘Cetacea’ E8 motif resembled the Alphapapillomaviruses E8 motif, exhibiting a small amino acid at the second position (Ala or Gly), an isoleucine at the third position and the WK motif (Trp-Lys) in the middle. In the E1–E2–L1–L2-based tree (Bravo & Félez-Sánchez, 2015) and in our E1 protein-based tree, the PV types with this motif clustered close to the alphasPVs. The majority of the PV types in this clade were of genital and/or mucosal origin. Group ‘Diverse majority’ (sequence logo in Fig. 5c) contained all of the remaining PVs (Fig. 2, light green), which presented a wide phylogenomic distribution. The highly conserved lysine at the second position (Fig. 5c) of the E8 motif is functionally important, and changing the lysine to an alanine eliminates the repressive function of E8‘E2 in BPV1 (Sankovski et al., 2015). This E8 motif is likely the most ancestral and is approximately 150 million years old according to the divergence time of the PV genera encoding this type of E8 (Shah et al., 2010). Three predicted E8 peptides did not resemble any other known or predicted E8 peptides: MrPV1 (Ricketts’s big-footed bat PV1, ‘MRPKENWMSRWKK’), ZcPV1 (California sea lion PV1, ‘MLKTVMGAHRDHRHR’), and FgPV1 (‘MGRETQGPESASAMAAPK’). These E8 peptides did not fit well into any group.

All of the experimentally verified E8 CDSs (Table 1) were located in the +1 reading frame with respect to E1 (the E1 reading frame is considered the 0 frame). This common property was not considered in our algorithm; however, all of the predicted E8 CDSs (Table S1) were also in the +1 frame. Being located in the same frame and exhibiting a relatively similar position within E1 (Table S1 and Fig. S10 MAFFT), are good arguments for the homologous nature of all of the predicted E8 CDSs. The transcriptional repression activity is described for E8‘E2 representatives from the ‘Alpha’ and ‘Diverse majority’ groups. The E8‘E2 proteins in all other PV types are most likely also involved in the regulation of early transcription and the establishment of infection.

The generation of E8‘E2 mRNA appears to be robust, as E8‘E2 mRNA has been isolated from different HPV types, from different sources for each of them (Chiang et al., 1991; Doorbar et al., 1990; Ferty et al., 2011; Isok-Paas et al., 2015; Kurg et al., 2010; Lace et al., 2008; Renaud & Cowsert, 1996; Rotenberg et al., 1989; Straub et al., 2014; Toots et al., 2014; Wang et al., 2011). Thus, our predictions for E8 5’ss are valid for a wide spectrum of conditions.

Most proteins (domains) must preserve their three-dimensional structure to maintain their functions; therefore, the same or a similar amino acid sequence must be conserved. Nevertheless, the protein-coding sequence can change during evolution (accumulating neutral substitutions) without altering the amino acid sequence, primarily because most amino acids can be encoded by different
synonymous codons. If conserved amino acids occur in certain regions in homologous (orthologous or paralogous) proteins, then these amino acids are considered important for the protein’s function. However, conservation of an amino acid sequence may also be caused by conservation at the nucleic acid level. Therefore, certain regions in the protein might be conserved, not because conservation of the amino acid sequence (function) is necessary, but because of conservation at the DNA level (or RNA level, for RNA viruses). There can be different reasons for this conservation, such as the existence of splice sites, internal promoters, dual-coding regions, mRNA secondary structures, ribosomal frameshift signals, etc. It is important to distinguish between these levels of conservation (i.e. protein and nucleic acid levels). cRegion was developed to identify regions inside protein-coding sequences, where selective pressure is stronger at the nucleotide level relative to the amino acid level in the major reading frame. The cRegions algorithm can identify functional elements if divergence occurs at the DNA level and the protein sequence is sufficiently conserved to generate a reasonable MSA. However, cRegion has certain deficiencies. First, this method cannot detect regions composed of amino acids with non-redundant codons (Met and Trp). Second, a sufficient mutational/evolutional time is required (i.e. the protein coding sequences should have had sufficient time to accumulate substitutions at the DNA level). In other words, if only the amino acid sequence of a protein is important, then the gene should have had enough time to use synonymous codons at that position, whereas if the conservation is at the DNA level, then the codons will be preserved. Third, a large number of homologous sequences are required to obtain an accurate assessment of observed nucleotide frequencies. Finally, cRegions cannot detect long double-coding regions because the codon usage of the analysed genes will be too biased.

cRegions compares observed and predicted nucleotide frequencies. As noted by an anonymous reviewer, a quite similar approach at the codon level has been successfully applied to detect packaging signals in the influenza A virus (Gog et al., 2007). The differences between observed and predicted values depend on how these predictions are obtained. One important parameter that impacts predicted frequencies is codon usage bias (noted also in Gog et al., 2007). Thus, it is important to realize that assessments of codon usage for E1 genes must be used to obtain relevant predictions. In the case of papillomaviruses, the virus may have adapted to its host and/or to tissue-specific codon usage. Therefore, papillomaviruses should not be analysed together because of their different host and tissue tropisms (Egawa et al., 2015). Instead, each genus should be analysed separately (assuming that tissue tropism and spatio-temporal gene expression are similar within the genus).

It is important to assess whether the E8 CDS is not a mere artefact of coding restrictions from the E1 gene. Both methods detected four constant distinct signals, including the E8 region in all PV genera with more than 11 PV types (i.e. alpha, beta, gamma, delta, lambda and chiPVs), which accounted for a total of 227 PV types (Fig. 4, Figs S4–S8). Other conserved signals included the 5’ss of the E1 E4 transcript, the conserved region described by Cam-pione-Piccardo and colleagues, and the E1–E2 gene overlap. The analysis with both programs revealed many other conserved elements within the E1 gene. Similar results were obtained when the initial E1 alignment was performed with MAFFT instead of MUSCLE and when other characteristics were used to compare the observed and predicted nucleotide distributions (i.e. the maximal difference or Chi-square test). In the PVs, the detection of these elements appears to be robust. Identifying overlapping functional elements using bioinformatics within viral sequences may help discover many hidden secrets of viruses.

CONCLUSIONS

According to our analysis, the E8’E2 protein is predicted (or shown) in 307 PV types and the presence of the E8 CDS is not a consequence of E1 conservation. According to our data and the PaVE data for E6 and E7, E8’E2 is as common protein as well-known PV oncogenes (E6 and E7) and deserves to be annotated in major databases.

METHODS

Predicting E8 CDSs in PV genomes. The PaVE database (http://pave.niaid.nih.gov/) was used to collect information on PV genomes. PaVE provides organized and curated information on papillomavirus genomics and tools (Van Doorslaer et al., 2013), including uniform protein names and corrected CDS annotations. The final dataset consisted of 318 PV genomes, excluding non-reference genomes (2 February 2016). All of the available published data concerning experimentally verified E8’E2 mRNAs were collected (Table 1) to assess the parameters for our search algorithm. The collected information presented in Table 1 demonstrates that the minimal distance between the start codon of the E1 gene and the start codon of the E8 protein CDS was 355 nt (BPV1), and the maximum was 409 nt (HPV18). Therefore, the first parameter for the algorithm was the minimal distance to begin the search for the E8 CDS (ATG codon) after the E1 start codon (from the PaVE annotation). In this work, the default value was 300 nt. The second parameter was window size, which indicated how far the algorithm should search. The default value for the second parameter was 500 nt. Table 1 shows that the shortest E8 CDS was 32 nt, and the longest was 41 nt (counted from the A nucleotide of the E8 initiator codon to the end of the E8 exon aG|gta). To avoid excluding any shorter or longer E8 sequences, the algorithm searched for E8 sequences that were between 17 and 62 nt in length. Therefore, the third and the fourth parameters were the minimal and maximal allowed sizes of the E8 CDS. The short strict splice donor site (5’ss) consensus NAG|GTA was used to mark the end of the E8 CDS by default because only the [C/G]|AG|GTA 5’ss was observed in the collected data (Table 1). When an E8 CDS that satisfied the default requirements was not found, the script used different 5’ss consensus sequences in the following order: NAG|GTG -> NAG|GTT -> NAG|GTC. This order is based on studies that have analysed mammalian splice-site consensus sequences (Abril et al., 2005; Burset et al., 2000; Roca et al., 2013). All of the described E8’E2 mRNA sequences exhibit a splice-
acceptor site (3’s) within the E2 gene, and in all cases, the E1’ E4 transcript uses the same 3’ss as E8’E2 (Table 1). According to previously described data, the E1’ E4 3’s is located in a fixed position with respect to the E2 codons. Therefore, the last restriction for the search algorithm was that the length of the E8 CDS divided by three must produce a residual of two. A summary of the script workflow is shown in Fig. S1. The algorithm was written in Python.

**Multiple-sequence alignment (MSA).** By default, protein MSAs were generated in MUSCLE v3.8.31 (Edgar, 2004) using the default settings. Occasionally, the –refine option was employed to improve the alignment. To illustrate the location of the predicted E8 CDSs with respect to the E1 protein alignment, a less biased set of PVs was generated (Fig. S10). First, the pairwise sequence similarities of the E1 proteins were reduced within each genus (alpha, beta, gamma, delta and lambda), until there were nine sequences per genus. All of the PVs from group ‘Cetacea’ (see results) were included. EcPV5 (Equus caballus PV5), CPV15 (Canis familiaris PV15), ZcPV1, HPV2, 27, 57 and 35 were excluded from the primary data because of an insertion in the E8 coding region, to allow better visualization. The E1 protein-based alignments were generated with MAFIT (Katoh & Standley, 2013) and converted to the respective nucleotide alignment with Pal2Nal. After the alignment, HPV76 and MrPV1 were removed for the reasons described above. The final set contained 57 PV types (Fig. S10).

**Consensus-sequence logos.** The predicted putative E8 peptide sequences were aligned using MUSCLE with the default settings. Logos were created with the WebLogo 3.4 online tool (http://weblogo.three- plusone.com/) using the default settings. For clarity, potential misidentified start codons of E8 for three HPVs (HPV48, HPV163 and HPV168) were corrected prior to MSA (see results).

**cRegions and SynPlot2.** The cRegions tool compares observed and predicted nucleotide frequencies. The underlying methodology of the tool is as follows. First, an MSA is generated from homologous protein sequences. Next, a corresponding nucleotide MSA is generated from the protein MSA with the Pal2Nal (Suyama et al., 2006) tool using the protein’s coding sequences. Third, the frequencies of each nucleotide are calculated in each column (each position) in the corresponding nucleotide MSA. Then, predictions related to the nucleotide frequencies are generated in each column of the MSA, based on the amino acid frequencies at the respective position. For these predictions, the average codon usage of E1 genes (from the same set of sequences for which the prediction was calculated) and the frequencies of the amino acids in each column are used (Fig. 3). In the final step, the observed and predicted nucleotide frequencies are compared. The root-mean-square deviation (RMSD) is then employed to compare the predicted and observed nucleotide frequencies.

\[
\text{RMSD} = \sqrt{\frac{1}{n} \left( (A_{\text{obs}} - A_{\text{pre}})^2 + (C_{\text{obs}} - C_{\text{pre}})^2 + (G_{\text{obs}} - G_{\text{pre}})^2 + (T_{\text{obs}} - T_{\text{pre}})^2 \right)}
\]

The subscript ‘obs’ indicates observed frequencies, and the subscript ‘pre’ indicates predicted frequencies.

In this work, a sliding-window mode was used instead of the raw RMSD values. The sliding-window mode indicates that the arithmetic mean is calculated over the RMSD values but includes only the third codon position. Each data point in the resulting graph (Fig. 3a) shows the position of the nucleotide that was first included in the calculation of the arithmetic mean. A window size of 3 was used to detect small conserved regions. Additionally, if one column in the MSA contained gaps of more than 20%, the position was excluded from the calculation of the arithmetic mean. The cRegions web-tool is available at (http://bioinfo.ut.ee/cRegions/).

SynPlot2 (Firth, 2014) was used via the web interface (http://guinenever. otago.ac.nz/cgi-bin/ael/synplot.pl). The E1 protein sequences were aligned with MUSCLE using the –refine option. The corresponding DNA alignment was generated with Pal2Nal, and the output was used as the input for SynPlot2. The sliding window parameter was set to 3, indicating (2n+1) three codons. Identical alignments were used for the cRegions and SynPlot2 algorithms. All of the MSAs used in this analysis are available in the http://bioinfo.ut.ee/cRegions/manuscript/supplementary_alignments.

**ACKNOWLEDGEMENTS**

This work was supported by the IUT20-27 from the Estonian Research Council and by the European Regional Development Fund through the Center of Excellence in Chemical Biology 3.2.0101.08-0017. The development of the cRegions webpage was supported by the European Regional Development Fund through the Research Internationalization Programme (ELIXIR). The authors also thank Sten Ilmarj and Hedi Peterson for their involvement in the very early stages of this work. M. P. and A. A. designed the experiments; M. P. performed the experiments; M. P. and A. A. analysed the data; and M. P. and A. A. wrote the paper.

**REFERENCES**


