RESEARCH ARTICLE

REVISED A unique insert in the genomes of high-risk human papillomaviruses with a predicted dual role in conferring oncogenic risk [version 2; peer review: 2 approved]

Noam Auslander, Yuri I. Wolf, Svetlana A. Shabalina, Eugene V. Koonin

National Center for Biotechnology Information, National Institutes of Health, USA, Bethesda, Maryland, 20814, USA

v2 **First published:** 02 Jul 2019, **8**:1000 (https://doi.org/10.12688/f1000research.19590.1)

Latest published: 01 Oct 2019, **8**:1000 (https://doi.org/10.12688/f1000research.19590.2)

Abstract

The differences between high risk and low risk human papillomaviruses (HR-HPV and LR-HPV, respectively) that contribute to the tumorigenic potential of HR-HPV are not well understood but can be expected to involve the HPV oncoproteins, E6 and E7. We combine genome comparison and machine learning techniques to identify a previously unnoticed insert near the 3’-end of the E6 oncoprotein gene that is unique to HR-HPV. Analysis of the insert sequence suggests that it exerts a dual effect, by creating a PDZ domain-binding motif at the C-terminus of E6, as well as eliminating the overlap between the E6 and E7 coding regions in HR-HPV. We show that, as a result, the insert might enable coupled termination-reinitiation of the E6 and E7 genes, supported by motifs complementary to the human 18S rRNA. We hypothesize that the added functionality of E6 and positive regulation of E7 expression jointly account for the tumorigenic potential of HR-HPV.

Keywords

Papillomaviruses, cervical cancer, oncogenic risk, translation terminatio-reinitiation, machine learning

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
REVISED **version 2** published 01 Oct 2019	✓	✓ report
	↑	↑
version 1 published 02 Jul 2019	✓ report	✓ report

1 **Alison McBride**, National Institute of Allergy and Infectious Diseases (NIAID), Bethesda, USA

2 **Elizabeth A. White**, University of Pennsylvania Perelman School of Medicine, Philadelphia, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding authors: Noam Auslander (noamaus@gmail.com), Eugene V. Koonin (koonin@ncbi.nlm.nih.gov)

Author roles: Auslander N: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Writing – Original Draft Preparation; **Wolf YI**: Investigation, Methodology; **Shabalina SA**: Formal Analysis, Methodology; **Koonin EV**: Conceptualization, Investigation, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by U.S. Department of Health and Human Services (Intramural funds).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Auslander N, Wolf YI, Shabalina SA and Koonin EV. **A unique insert in the genomes of high-risk human papillomaviruses with a predicted dual role in conferring oncogenic risk [version 2; peer review: 2 approved]** F1000Research 2019, **8** :1000 (https://doi.org/10.12688/f1000research.19590.2)

First published: 02 Jul 2019, **8**:1000 (https://doi.org/10.12688/f1000research.19590.1)

REVISED **Amendments from Version 1**

The following changes have been made.

Reviewer 1:

1. We now mention the most recent number of HPV types that have been formally recognized 198), and the reference supporting this has been updated.
2. We added numerous references to recent publications that describe analyses of HPV evolution, oncogenicity, oncoprotein interactions and E6 interactions with PDZ domains.
3. We now include all 64 alpha HPV types from PaVE, and all analyses and Figure 1– Figure 3 include those HPV strains.
4. Resolution of the new figures has been improved.
5. The legend to Figure 2B has been modified as suggested.
6. The legend to Figure 2C describes all symbols and colored blocks.
7. Figure 3A has been updated to show the motifs clearly and arrange the strains by the order in the phylogenetic tree.
8. Figure 3B has been updated, confusing labelling of 18S rRNA removed.

Reviewer 2:

1. Many references to the current literature were added, and several other features of HR-HPV E6 are now discussed.
2. The suggested references to HR-HPV E6 specific binding partners have been included.
3. The discussion on the importance of the PDZ binding motif in distinguishing HR from LR HPV has been tempered, and the context for its potential relevance has been clarified.
4. We now discuss the recent analysis of 5570 cervical HPV16 genomes (Mirabello and colleagues) and have substantially expanded the discussion of the effect of the insert on E7 production.
5. Figure 1B has been updated, phylogenetic trees are now presented in a linear format.

In addition, although the referees have suggested a statistical review of the permutation test that was used in this work to assess the significance of the results, we are confident that this is unnecessary because the statistical technique we used is simple and standard.

Any further responses from the reviewers can be found at the end of the article

Introduction

Persistent infections with carcinogenic human papillomaviruses (HPVs) are the main cause of cervical neoplasia and cancer, with over 99% of the cervical lesions containing HPV sequences[1–3]. There are currently 198 HPV types that have been officially numbered[4], of which approximately a third are predominantly detected in the cervical epithelium and belong to the *Alphapapillomavirus* genus[5,6]. The viruses of this genus are further grouped into high-risk (HR) and low-risk (LR) HPV types based on their association with cervical cancer and pre-cancerous lesions[7,8]. Most of the HR-HPV variants belong to the *Human papillomavirus 16* (alpha-9) or *Human papillomavirus 18* (alpha-7) species groups[9].

Phylogenetic trees constructed from alignments of complete HPV genomes cluster all oncogenic types together, suggesting a common ancestor for the HR-HPVs. However, in separate trees built from different regions of the genome, the carcinogenic potential co-segregates with the early (those produced prior to replication) but not with the late genes[10,11]. The early HPV proteins E6 and E7 have transforming properties[12–14] and are required for the malignant conversion. The involvement of these proteins in tumorigenesis is thought to stem from their interactions with the tumor suppressors p53 and pRB, respectively, as well as other proteins involved in tumorigenesis[14–20]. Variations in E6 and E7 proteins have been suggested to determine the oncogenic potential of HPV[21,22] but the potential discriminating features of the oncogenic variants are frequently observed in LR-HPVs as well[23–27]. A notable molecular feature that distinguishes HR from LR-HPV (along with pronounced sequence differences) is the presence of a PDZ-domain recognition motif at the extreme C terminus of the HR-E6 oncoprotein, as opposed to LR-E6[28–30], which enables interactions of HR-E6 with numerous PDZ domain proteins[31–33].

Both oncogenes are transcribed from an early promoter as a single E6-E7 polycystronic pre-mRNA. The transcriptional level and translational efficacy of this RNA are regulated by the alternative RNA splicing machinery of host cells[34,35]. Alternative splicing of introns in the E6 gene produces multiple splice isoforms of E6-E7 pre-mRNA[34,36]. In HR-HPV16 and HPV18, E6*I is one of the major splice isoforms that functions as the mRNA for the production of E7 via translation reinitiation[37], In contrast, unspliced E6 mRNA is responsible for the production of E6[34,37]. Another recent study has shown that HPVs generate single-stranded circular RNAs (circRNAs), some of which encompass the E7 gene (circE7)[38].

Identification of signatures of the HR-HPV genotypes that differentiate them from the majority of alpha papillomaviruses that lack oncogenic potential could help elucidate the genetic basis of the carcinogenic properties of HPVs, thus contributing to a better understanding of the biological mechanisms exploited by the virus to trigger neoplasia. However, at present, genomic determinants of the HPV oncogenic risk are not well understood, and the exact nature of the genetic changes that led to the emergence of the HR-HPV oncogenicity remains unknown.

To better understand HPV carcinogenesis, we revisited the search for specific genomic determinants of HR-HPV types and identified a previously unreported insert of 30 to 60 base pairs (bp) at the 3'-end of the E6 oncoprotein coding region that is present in all HR-HPV types but not in LR-HPV. This insert introduces a new stop codon, separating the nucleotide sequence coding for E6 from that coding for E7, thus, eliminating the overlap between E6 and E7 that is characteristic of the LR-HPV types. The insert confers a PDZ binding motif at the end of E6 oncoprotein which is likely important for the oncogenic potentiation. Additionally, the insert places the termination codon of E6 upstream but in close proximity of the initiation codon of E7. Furthermore, the insert contains sequences complementary to human 18S rRNA in the regions of hairpins 26 and 27 that are known to interact with viral RNAs and are specifically involved in IRES (Internal Ribosomal Entry Sites) binding and cap-independent translation[39–42]. We hypothesize that the insert separating the coding sequence of E6 and E7 was the primary cause of the emergence of high oncogenic potential alpha-HPV.

Results

The complete genome sequences and amino acid sequences of HPV E1, E2, E6, E7, L1 and L2 proteins were collected for all sequenced alpha-HPV strains (**pave.niaid.nih.gov**[43]). We then constructed a global multiple sequence alignment of the whole genome nucleotide sequences and the amino acid sequences alignments for each protein. Maximum likelihood phylogenetic analysis of these alpha-HPVs, based on the complete nucleotide sequence, as well as the amino acid sequences of most of the early proteins, identified HR-HPV as a clade, whereas phylogenies of L1 and L2 did not support the monophyly of the HR-HPV (Figure 1), in agreement with previous findings[9,10]. These observations are compatible with a major role of recombination in HPV evolution.

We next sought to identify genomic features that might partition alpha-HPV species in accord with their oncogenic risk, focusing on E6 and E7 oncogenes.

First, we searched for regions of insertions and deletions within the genome nucleotide sequences of E6 and E7 that might differentiate between the risk groups. Specifically, we identified sequences with high frequency of deletions or insertions that are located within high confidence alignment regions (See *Methods* for details). We then applied Support Vector Machine (SVM), a linear classification technique, with a leave-one-out cross-validation, aiming to identify regions that differentiate between high-risk and low-risk strains. This approach resulted in the identification of genomic regions that separated HR-HPV from LR-HPV with high accuracy (>0.75, with statistical significance; see *Methods*). Among these, we found one prominent insert (between 30 and 60 nucleotides) located in the 3'-terminal region of the E6 gene (Figure 2A). We also performed a similar search for regions separating HR-HPV from LR-HPV, using the amino acid sequence of E6 and E7 oncoproteins. For the purpose of classification, we coded the amino acids with numbers based on their frequencies and the BLOSUM62[44] matrix (see *Methods*). As expected, the divergent region in E6 was identified from the amino acid sequences as well (Figure 2B). In contrast, in the E7 protein sequences, we did not find any significant differences between the high risk and low risk HPV strains (*Extended data:* Figure S2[45]).

The discriminating region identified in the E6 gene contains an insert with a conserved sequence in all HR-HPV strains. The insert contains an in-frame stop codon for the E6 coding sequence, which eliminates the overlap between the coding sequences of E6 and E7 that is characteristic of the LR-HPV genomes, but results in nearly identical lengths of the E6 proteins in HR and LR-HPV strains albeit with unrelated C-terminal sequences of 8-19 amino acids (Figure 2B). The unique C-terminal sequence of HR-HPV E6 that originates from the insert contains a PDZ domain-binding motif X-T-X-V/L at the very C-terminus of E6 in almost all HR-HPVs. Indeed, several PDZ domain-containing proteins have been identified as binding partners of HR-E6, including hDlg, hScrib, MAGI-1, MAGI-2, MAGI-3, and MUPP1[29,32,33,46–48]. These interactions that are

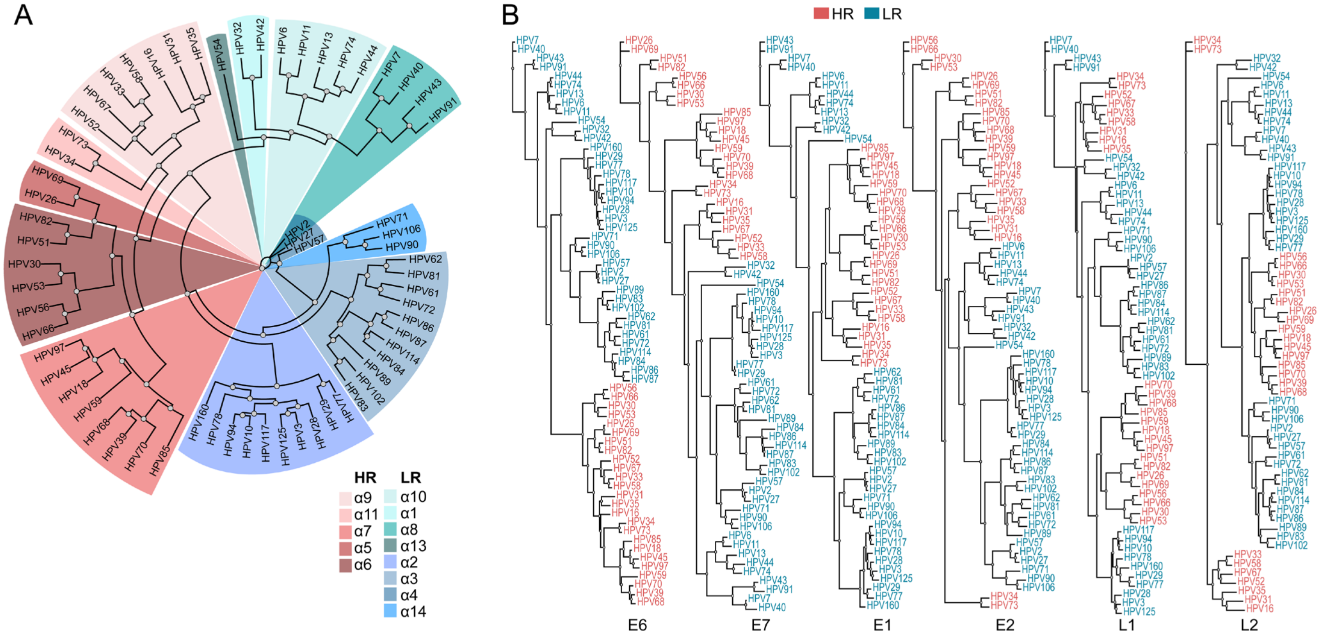

Figure 1. Phylogenetic trees of alpha-HPV. (**A**) Maximum likelihood tree obtained with the whole genome nucleotide sequences alignment of alpha-HPV, colored by the different alpha-class categories. (**B**) Maximum likelihood trees obtained with alignments of E6, E7, E1, E2, L1 and L2 amino acid sequences of alpha-HPV, colored by the oncogenic risk groups.

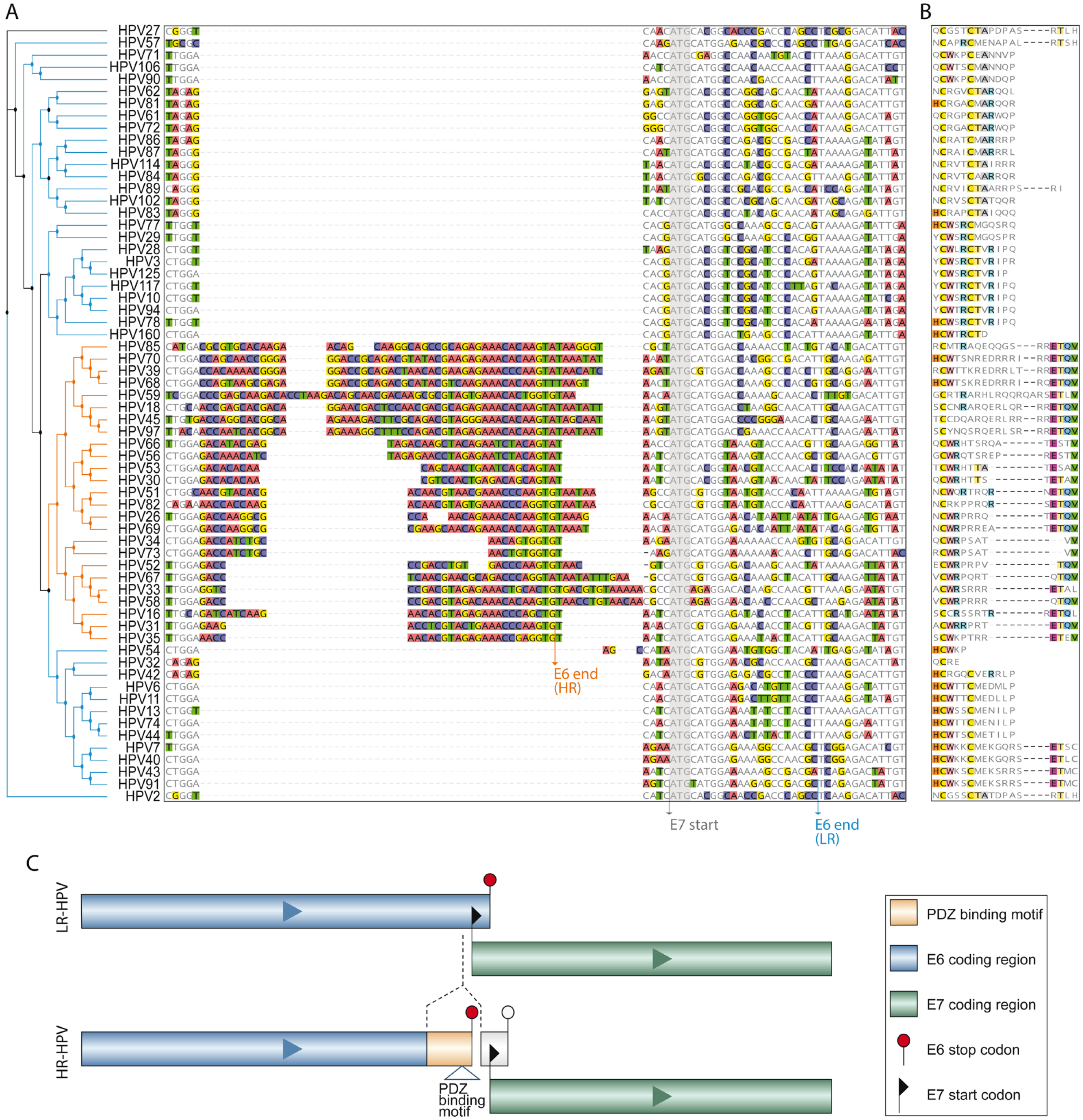

Figure 2. HR-HPV-specific sequence insert. (**A**) Nucleotide sequence alignment of alpha-HPVs (blue, LR-HPV; orange, HR-HPV sequences). Arrows: Grey, E7 start codon for most HPV types; Blue and Orange, E6 stop codons that are distinct between LR-HPV and HR-HPV. (**B**) Alignment of the C-terminal amino acid sequences of E6 proteins of alpha-HPVs. (**C**) Schematic representation of the E6/E7 separation caused by the insertion in the 3'-terminal region of E6 and the proximity of E6 stop and E7 starts in HR-HPV.

unique for HR-HPV are likely to contribute to HR-HPV induced oncogenesis[49].

We observed that the sequence similarity between the insert sequences among HR-HPV strains is more pronounced at the nucleotide level than at the amino acid level (See METHODS for details and *Extended data:* Figure S3[45]). Combined with the separation between the coding regions of E6 and E7 resulting from the insertion, and the proximity of E6 stop codon to the E7 start codon, this finding led us to hypothesize that the insert has an additional role as a regulatory element. Furthermore, as E7 has been previously identified as the dominant oncogene[50], the lack of genomic determinants of HR-HPV within the E7 gene is compatible with the possibility that the insert contains regulatory elements enhancing E7 expression in HR-HPV strains.

Several cases of coupled termination-reinitiation for polycistronic mRNA with proximate stop and start codons are evident for translation of eukaryotic virus genes[51–57]. The efficiency of this process depends on the close proximity of the termination and reinitiation sites[57,58], and the presence of motifs complementary to the 18S rRNA in the mRNA sequence[51,52,57,59]. Given the proximity of the E6 termination site to the E7 initiation site codon that results from the insertion in the HR-HPV strains conferred by the insert, we investigated the possibility of coupled termination-reinitiaion of E6 and E7 in these strains. Notably, within the inserted sequence in the vicinity of the E7 start codon, we identified two conserved regions that are complementary to the sequences in 18S rRNA hairpins 26 and 27 which are commonly involved in the interactions between ribosomes and virus IRES and responsible for cap-independent translation[60] (Figure 3A). The recognition by host 18S rRNA was shown to be important factor for active termination-reinitiation

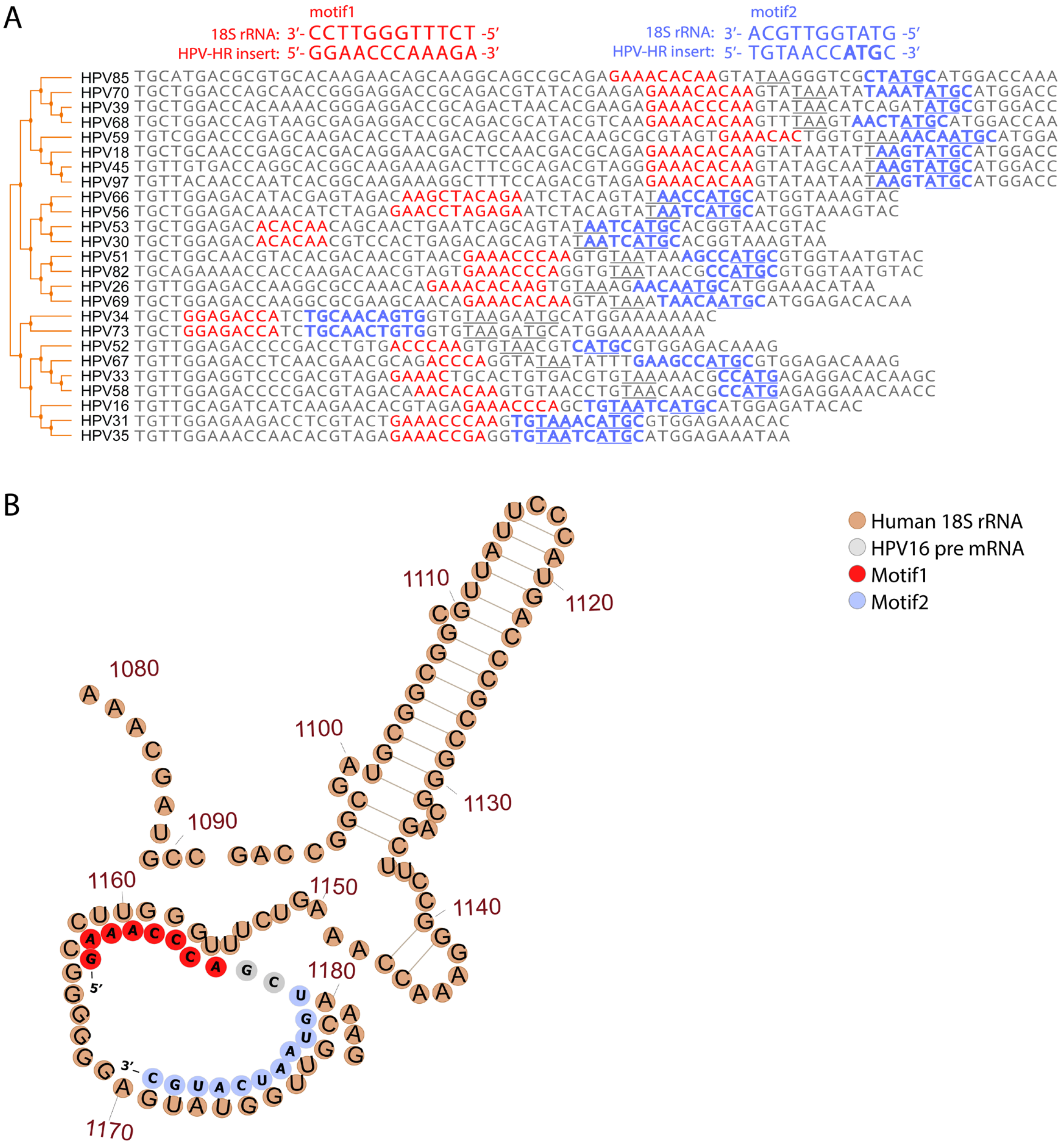
A
motif1
18S rRNA: 3′- CCTTGGGTTTCT -5′
HPV-HR insert: 5′- GGAACCCAAAGA -3′
motif2
18S rRNA: 3′- ACGTTGGTATG -5′
HPV-HR insert: 5′- TGTAACCATGC -3′
HPV85 TGCATGACGCGTGCACAAGAACAGCAAGGCAGCCGCAGAGAAACACAAGTATAAGGGTCGCTATGCATGGACCAAA
HPV70 TGCTGGACCAGCAACCGGGAGGACCGCAGACGTATACGAAGAGAAACACAAGTATAAATATAAATATGCATGGACC
HPV39 TGCTGGACCACAAAACGGGAGGACCGCAGACTAACACGAAGAGAAACCCAAGTATAACATCAGATATGCGTGGACC
HPV68 TGCTGGACCAGTAAGCGAGAGGACCGCAGACGCATACGTCAAGAAACACAAGTTTAAGTAACTATGCATGGACCAA
HPV59 TGTCGGACCCGAGCAAGACACCTAAGACAGCAACGACAAGCGCGTAGTGAAACACTGGTGTAAAACAATGCATGGA
HPV18 TGCTGCAACCGAGCACGACAGGAACGACTCCAACGACGCAGAGAAACACAAGTATAATATTAAGTATGCATGGACC
HPV45 TGTTGTGACCAGGCACGGCAAGAAAGACTTCGCAGACGTAGGGAAACACAAGTATAGCAATAAGTATGCATGGACC
HPV97 TGTTACAACCAATCACGGCAAGAAAGGCTTTCCAGACGTAGAGAAACACAAGTATAATAATAAGTATGCATGGACC
HPV66 TGTTGGAGACATACGAGTAGACAAGCTACAGAATCTACAGTATAACCATGCATGGTAAAGTAC
HPV56 TGCTGGAGACAAACATCTAGAGAACCTAGAGAATCTACAGTATAATCATGCATGGTAAAGTAC
HPV53 TGCTGGAGACACACAACAGCAACTGAATCAGCAGTATAATCATGCACGGTAACGTAC
HPV30 TGCTGGAGACACACAACGTCCACTGAGACAGCAGTATAATCATGCACGGTAAAGTAA
HPV51 TGCTGGCAACGTACACGACAACGTAACGAAACCCAAGTGTAATAAAGCCATGCGTGGTAATGTAC
HPV82 TGCAGAAAACCACCAAGACAACGTAGTGAAACCCAGGTGTAATAACGCCATGCGTGGTAATGTAC
HPV26 TGTTGGAGACCAAGGCGCCAAACAGAAACACAAGTGTAAAGAACAATGCATGGAAACATAA
HPV69 TGCTGGAGACCAAGGCGCGAAGCAACAGAAACACAAGTATAAATAACAATGCATGGAGACACAA
HPV34 TGCTGGAGACCATCTGCAACAGTGGTGTAAGAATGCATGGAAAAAAAC
HPV73 TGCTGGAGACCATCTGCAACTGTGGTGTAAGATGCATGGAAAAAAAA
HPV52 TGTTGGAGACCCCGACCTGTGACCCAAGTGTAACGTCATGCGTGGAGACAAAG
HPV67 TGTTGGAGACCTCAACGAACGCAGACCCAGGTATAATATTTGAAGCCATGCGTGGAGACAAAG
HPV33 TGTTGGAGGTCCCGACGTAGAGAAACTGCACTGTGACGTGTAAAAACGCCATGAGAGGACACAAGC
HPV58 TGTTGGAGACCCCGACGTAGACAAACACAAGTGTAACCTGTAACAACGCCATGAGAGGAAACAACC
HPV16 TGTTGCAGATCATCAAGAACACGTAGAGAAACCCAGCTGTAATCATGCATGGAGATACAC
HPV31 TGTTGGAGAAGACCTCGTACTGAAACCCAAGTGTAAACATGCGTGGAGAAACAC
HPV35 TGTTGGAAACCAACACGTAGAGAAACCGAGGTGTAATCATGCATGGAGAAATAA
B
1080
1090
1100
1110
1120
1130
1140
1150
1160
1170
1180
5′
3′
Human 18S rRNA
HPV16 pre mRNA
Motif1
Motif2

Figure 3. Sequences complementary to 18S rRNA sequence in the HR-HPV-specific insert. (**A**) Comparison of 18S rRNA complementary sequences for different HR-HPV strains. The distal and proximal to the E7 start motifs (motifs 1 and 2) to the E7 start codon are shown in red and blue, respectively. The E6 stop codons (TAA) and E7 start codons are marked in bold. (**B**) An illustration of the potential interactions between folded HPV16 pre-mRNA (grey, red and blue) and the human 18S rRNA (brown). The two motifs are marked on the HPV-structure in red (motif 1) and blue (motif 2).

for polycistronic mRNA with proximate stop and start codons[51,52,57,59]. The first region of complementarity consistently forms an internal loop and a relaxed, unpaired structure in the predicted optimal and sub-optimal E6-E7 mRNA folding of HR-HPV strains[58] (*Extended data:* Figure S4[45], see Methods). The second region of 18S RNA complementarity overlaps with the E7 start site, and hence might function independently or cooperate in the regulation of E7 translation (Figure 3B). Also, these regions of complementarity might be important for generation of circE7 and for translation regulation of E7 encoded by circE7 or by alternative transcripts through cap-independent mechanisms. These findings suggest that the insertion could enable coupled termination-reinitiation of E6 and E7 proteins and could regulate E7 translation encoded by E6E7 alternative transcripts and circE7, enhancing their combined expression in HR-HPV, and thus promoting the oncogenic transformation induced by these viruses.

Furthermore, it has been recently reported that oncogenic human HPVs generate single-stranded circular RNAs (circRNAs), some of which encompass the E7 gene (circE7)[38]. CircE7 is represented in the TCGA RNA-Seq data from HPV-positive cancers, and specific disruption of circE7 in CaSki cervical carcinoma cells reduces E7 protein levels and inhibits cancer cell growth both *in vitro* and in tumor xenografts[38]. Given that the insert identified here is located in the 5'UTR of cicrE7 and in the proximity of the circE7 backsplice junction, it might increase the E7 translation efficacy from circE7 as well as facilitate the generation of the circE7 RNA.

Discussion

The genus *Alphapapillomavirus* includes HPV types that are uniquely tumorigenic. However, the events in the HPV genome evolution that led to the carcinogenic potential of some alpha-HPV types remain poorly understood. Here, we revisited this problem by performing a search for genomic determinants of the oncogenic risk of alpha-HPV types and identified a unique insert in the 3'-terminal regions of the E6 oncoprotein genes in all HR-HPV strains. To the best of our knowledge, this insert in HR-HPV genomes has not been reported previously. The insertion maintains closely similar lengths of E6 proteins in HR-HPV and LR-HPV types, which could explain why it has been overlooked in previous HPV genome analyses.

We hypothesize that the insertion makes a dual contribution to the oncogenicity of the HR-HPV types. First, the inserted sequence changes the C-terminal amino acid sequence of E6 and creates a PDZ domain-binding motif that is unique to HR-HPV types. The experimentally demonstrated interaction between the E6 proteins of HR-HPV and several PDZ domain-containing proteins is thought to contribute to HPV-induced tumorigenesis[30,32,61]. Interestingly, PDZ-binding motifs have been identified also in several other oncogenic viruses, such as HTLV-1, adenovirus RhPV1 and beta-HPV8[62,63]. Second, the insert eliminates the overlap between the E6 and E7 coding regions, implying a possible role as a regulatory element. We found that almost all HR-HPV genomes contain the sequence T-A/G-T-A-A-T/A in the insert near the end of the E6 coding sequence, which is closely similar to the sequence of the early promoter at the 5' end of E6 that is employed for the synthesis of the E6-E7 mRNA[64,65]. However, for the HR-HPV strains, unlike the case of the LR-HPV[66], there are no reports of an independent E7 promoter, so that E6 and E7 are both translated from a polycistronic mRNA. Thus, the promoter-like sequence within the insert is likely to be spurious.

In HR-HPV strains, E6 and E7 proteins are translated from a polycistronic pre-mRNA[67], and translation reinitiation has been suggested as the mechanism[37,66,68]. However, the close proximity of the E6 stop codon to the E7 start codon in HR-HPV (only a few nucleotides separating these codons) could inhibit re-initiation[66,68]. Therefore, it has been suggested that ribosomal reinitiation is enabled through the formation of the E6*I splice variant (removal of intron I) in which the intercistronic distance between the translation termination codon of E6*I and the E7 initiation codon is increased[37]. Thus, E7 protein translation might be enhanced by the removal of the intron I region from pre-mRNAs by splicing, whereas retention of that sequence is required for E6 translation[69]. A comprehensive characterization of HPV16 and HPV18 expression by RNA-Seq analysis in invasive cervical cancer has shown that E6*I is the most abundant transcript isoform for both viral types[70]. The insert identified here is located in the 5'UTR immediately upstream of the E7 start codon in the E6*I transcript, and there are two alternative branch sites for E6*I splicing to produce a E7 mRNA with variable 5'UTRs[71]. However, several studies have reported that E7 translation is independent of splicing within the E6 open reading frame[68,72,73], undermining this interpretation.

Further investigating the potential regulatory role of the inserted sequence, we identified two conserved motifs that are complementary to the human 18S rRNA (Figure 3); interaction of such motifs with the rRNA has been shown to play a role in coupled termination-reinitiation for several viral genes[51,52,57,59]. The first motif forms an internal loop in the predicted mRNA secondary structure of E6-E7, whereas the second one overlaps with the E7 initiation codon. Given the evolutionary conservation of these motifs and the close proximity of E6 termination site to the E7 initiation site, it appears plausible that coupled termination-reinitiation promoted by the insert sequence is an important mechanism for E7 translation in HR-HPV strains.

The folding of these regions of rRNA complementarity in E6-E7 mRNAs is typically relaxed in the predicted optimal and sub-optimal secondary structures of the HR-HPV strains. The insert is present not only in polycistronic *E6E7* pre-mRNAs, but also in some alternatively spliced *E6E7* transcripts, including the most common spliced isoform *E6*I* in HR- HPV16 and HPV18[70]. The biologically functional circE7 that has been identified in TCGA RNA-Seq data from HPV-positive cancers[38], contains the unique insert in its 5'UTR. Thus, the regions of complementarity to human 18S rRNA could modulate the translation rate of the E7 encoded by alternative transcripts and circRNAs in HR-HPV.

Recent analysis of 5570 cervical HPV16 genomes[74], in which the E6 gene was found to be highly variable whereas the E7 gene was strikingly devoid of genetic variants in precancer and cancer cases, support the notion that idea that E7 is a the contributor in HR-HPV-induced carcinogenesis[50,75]. Thus, enhancement of E7 production by regulatory elements contained in the insert is likely to substantially affect the oncogenicity of the HR-HPV strains[76]. The insert is likely to enable coupled termination-reinitiation of the E6 and E7 genes, which is supported by the presence, within the insert sequence, of motifs complementary to the human 18S rRNA. In addition to the enhancement of reinitiation during the E6-E7 mRNA translation, this interaction might affect the dynamics and efficiency of E7 translation from alternative transcripts in HR-HPV. Furthermore, these motifs might also regulate E7 translation encoded by circE7 through cap-independent mechanisms. The dynamics and the rate of translation of E7 alternative isoforms appear to vary and seem to influence the stability of E7 proteins differentially across HPV types[34,70,76]. Thus, potential 18S rRNA – insert duplexes could modulate the rate of E7 translation initiation and elongation in HR-HPV and result in differences in stability of HR-HPV E7 proteins compared to LR-HPV, and thus could affect the interactions between E7 and various protein partners.

Given the lack of additional major genomic determinants consistently differentiating between HR-HPV and LR-HPV, it seems most likely that the insert in the E6-E7 transcript is the primary cause behind the emergence of oncogenic HPV and makes a complex contribution to the oncogenicity of the HR-HPV types in which both E6 and E7 oncoproteins are involved.

Methods

Multiple sequence alignment and phylogenetic analysis

Multiple alignments of nucleotide and amino acid sequences that were obtained from PaVE (pave.niaid.nih.gov[43]) were generated using MAFFT v7.407[77] with default parameters. Maximum likelihood phylogenetic analysis was performed using the resulting alignments and PhyML 3.1 software[78], with the Bayesian information criterion, NNI tree improvement and an LRT SH-like likelihood method for support estimation.

SVM applied to nucleotide sequences

To apply Support Vector Machines (SVM[79], using Matlab 2018b *fitsvm* function) to the nucleotide sequences, we first encoded the data with numbers, where each nucleotide is coded as '1' and each gap as '0'. We searched for alignment regions with deletions or insertions in multiple HPV strains that are surrounded by high confidence alignment regions (alignment regions of length > 15 bp containing less than 5% of gaps in each position) because these are most likely to contain relevant differences within conserved genomic regions. Within these regions, we then trained the SVM to classify alpha-HPV strains based on their oncogenic potential. The performance of the SVM was evaluated by leave-one-out cross validation, and regions with the overall balanced accuracy >0.75 (the average of the accuracy for positive and negative classes) were selected for further analysis.

SVM applied to amino acid sequences

To apply SVM (using Matlab 2018b *fitsvm* function) to the amino acid sequences of E6 and E7 proteins, we first encoded the amino acids with numbers using the frequencies of each amino acid in each protein and the BLOSUM62 matrix[44]. For each position, the most frequent amino acid was identified, and the amino acids in each protein were encoded by their BLOSUM62 distances from the most frequent amino acid in the respective position. We then trained the SVM to classify alpha-HPV strains based on their oncogenic potential using the coded protein sequences. Positions surrounded by high-confidence alignment regions (length > 5 amino acids and containing less than 5% of gaps in each position) were selected for further analysis. For these positions, we evaluated the performance by leave-one-out cross validation, and regions with the overall balanced accuracy >0.75 were selected for further estimation of significance using a permutation test.

Estimation of the significance of the identified regions using permutations

To estimate the significance of the identified regions, i.e. to determine whether similar differences could be observed by chance, we performed a permutation test (*Extended data*[62]), controlling for the topology of the reconstructed phylogenetic trees. To this end, the labels were randomly permuted while maintaining unified labels for clades with high similarity. For each identified region, the likelihood of obtaining an equivalent or higher performance for the length of the region within the respective protein was evaluated by calculating an empirical P-value. We consider regions with permutation P-value <0.05.

Analysis of RNA folding and RNA-RNA duplexes

The most stable secondary structures were predicted for all HR-HPV strains and their free energy values were calculated using the Vienna package[60]. Afold and Mfold were applied for prediction of optimal and sub-optimal mRNA structures[80,81]. Target opening free energy was estimated for motifs 1 and 2 using optimal and sub-optimal RNA structures, as described previously[82]. The sequence fold variants with the lowest secondary-structure free energy are presented in the *Extended data:* Figure S4[45]. The formation of intermolecular mRNA-rRNA duplexes and hybridization affinity of the E6-E7 inserts to ribosomal RNA were evaluated using the Hybrid software with default parameters, with a ΔG threshold of ≤–10 kcal/mol[83,84]. The human 18S rRNA 2D structure was downloaded from http://apollo.chemistry.gatech.edu/RibosomeGallery/;.

Data availability

Underlying data

Nucleotide sequences of HPV and amino acid sequences of HPV E6 and E7 proteins are available from PaVE database **pave.niaid.nih.gov**[43]

Extended data

Harvard Dataverse: A unique insert in the genomes of in high-risk human papillomaviruses with a predicted dual role in conferring oncogenic risk, https://doi.org/10.7910/DVN/FUGEUB[45].

This project contains the following extended data:

- Table 1: Complete nucleotide sequences and the amino acid sequences of HPV E1, E2, E4, E5, E6, E7, L1 and L2 proteins. This is the only source data that was required and employed for the analysis reported in this work.
- Figure S1: Maximum likelihood trees obtained with alignments of E6, E7, E1, E2, E4, E5, L1 and L2 amino acid sequences of alpha-HPV strains.
- Figure S2: The balanced accuracy (y-axis) obtained from a leave-one-out cross validation for predicting risk category (HR vs LR) of alpha-HPV strains using BLOSUM62 coding of amino acid sequences, of different positions (x-axis) E6 (A) and E7 (B). Zero-accuracy was assigned to regions surrounded with low confidence alignment.
- Figure S3: Boxplots showing the distributions of the identity fraction of each nucleotide (NN) and amino acid (AA) in the genome and protein sequences of the insertion (not considering gaps for both NN and AA). The individual identity fractions of each position are overlaid.
- Figure S4: RNA fold secondary structure prediction of HR HPV strains 16 (A), 18 (B), 45 (C) and 31 (D). The nucleotides of the first motif are marked in red, and of the second motif in purple. E7 AUG is noted in red font.

Data are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

Zenodo: Permutation test controlling for HPV strains, http://doi.org/10.5281/zenodo.3242231[85]. Apache License, Version 2.0.

Acknowledgements

The authors would like to Itamar Sela, Alejandro A. Schaffer and Kira S. Makarova for helpful discussion and comments on the manuscript.

References

1. Walboomers JM, Jacobs MV, Manos MM, *et al.*: **Human papillomavirus is a necessary cause of invasive cervical cancer worldwide.** *J Pathol.* 1999; **189**(1): 12–9.
PubMed Abstract | Publisher Full Text

2. Bosch FX, de Sanjosé S: **Chapter 1: Human papillomavirus and cervical cancer--burden and assessment of causality.** *J Natl Cancer Inst Monogr.* 2003; (31): 3–13.
PubMed Abstract | Publisher Full Text

3. Bosch FX, Lorincz A, Muñoz N, *et al.*: **The causal relation between human papillomavirus and cervical cancer.** *J Clin Pathol.* 2002; **55**(4): 244–65.
PubMed Abstract | Free Full Text

4. Bzhalava D, Eklund C, Dillner J: **International standardization and classification of human papillomavirus types.** *Virology.* 2015; **476**: 341–344.
PubMed Abstract | Publisher Full Text

5. Bernard HU, Burk RD, Chen Z, *et al.*: **Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments.** *Virology.* 2010; **401**(1): 70–9.
PubMed Abstract | Publisher Full Text | Free Full Text

6. de Villiers EM, Fauquet C, Broker TR, *et al.*: **Classification of papillomaviruses.** *Virology.* 2004; **324**(1): 17–27.
PubMed Abstract | Publisher Full Text

7. Muñoz N, Bosch FX, de Sanjosé S, *et al.*: **Epidemiologic classification of human papillomavirus types associated with cervical cancer.** *N Engl J Med.* 2003; **348**(6): 518–27.
PubMed Abstract | Publisher Full Text

8. Smith JS, Lindsay L, Hoots B, *et al.*: **Human papillomavirus type distribution in invasive cervical cancer and high-grade cervical lesions: a meta-analysis update.** *Int J Cancer.* 2007; **121**(3): 621–32.
PubMed Abstract | Publisher Full Text

9. Schiffman M, Herrero R, Desalle R, *et al.*: **The carcinogenicity of human papillomavirus types reflects viral evolution.** *Virology.* 2005; **337**(1): 76–84.
PubMed Abstract | Publisher Full Text

10. Narechania A, Chen Z, DeSalle R, *et al.*: **Phylogenetic incongruence among oncogenic genital alpha human papillomaviruses.** *J Virol.* 2005; **79**(24): 15503–10.
PubMed Abstract | Publisher Full Text | Free Full Text

11. García-Vallvé S, Alonso A, Bravo IG: **Papillomaviruses: different genes have different histories.** *Trends Microbiol.* 2005; **13**(11): 514–21.
PubMed Abstract | Publisher Full Text

12. McLaughlin-Drubin ME, Münger K: **Oncogenic activities of human papillomaviruses.** *Virus Res.* 2009; **143**(2): 195–208.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Ghittoni R, Accardi R, Hasan U, *et al.*: **The biological properties of E6 and E7 oncoproteins from human papillomaviruses.** *Virus Genes.* 2010; **40**(1): 1–13.
PubMed Abstract | Publisher Full Text

14. Münger K, Howley P, Di Maio D: **Human papillomavirus E6 and E7 oncogenes.** In: *The Papillomaviruses.* 2007; 197–252.
Publisher Full Text

15. Moody CA, Laimins LA: **Human papillomavirus oncoproteins: pathways to transformation.** *Nat Rev Cancer.* 2010; **10**(8): 550–60.
PubMed Abstract | Publisher Full Text

16. Yim EK, Park JS: **The role of HPV E6 and E7 oncoproteins in HPV-associated cervical carcinogenesis.** *Cancer Res Treat.* 2005; **37**(6): 319–24.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Münger K, Scheffner M, Huibregtse JM, *et al.*: **Interactions of HPV E6 and E7 oncoproteins with tumour suppressor gene products.** *Cancer Surv.* 1992; **12**: 197–217.
PubMed Abstract

18. Klingelhutz AJ, Roman A: **Cellular transformation by human papillomaviruses: lessons learned by comparing high- and low-risk viruses.** *Virology.* 2012; **424**(2): 77–98.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Rozenblatt-Rosen O, Deo RC, Padi M, *et al.*: **Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins.** *Nature.* 2012; **487**(7408): 491–5.
PubMed Abstract | Publisher Full Text | Free Full Text

20. White EA, Kramer RE, Tan MJ, *et al.*: **Comprehensive analysis of host cellular interactions with human papillomavirus E6 proteins identifies new E6 binding partners and reflects viral diversity.** *J Virol.* 2012; **86**(24): 13174–86.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Barbosa MS, Vass WC, Lowy DR, *et al.*: ***In vitro* biological activities of the E6 and E7 genes vary among human papillomaviruses of different oncogenic potential.** *J Virol.* 1991; **65**(1): 292–8.
PubMed Abstract | Free Full Text

22. Halbert CL, Demers GW, Galloway DA: **The E6 and E7 genes of human papillomavirus type 6 have weak immortalizing activity in human epithelial cells.** *J Virol.* 1992; **66**(4): 2125–34.
PubMed Abstract | Free Full Text

23. Thomas M, Banks L: **Human papillomavirus (HPV) E6 interactions with Bak are**

conserved amongst E6 proteins from high and low risk HPV types. *J Gen Virol.* 1999; **80**(Pt 6): 1513–7.
PubMed Abstract | Publisher Full Text

24. Schmitt A, Harry JB, Rapp B, *et al.*: **Comparison of the properties of the E6 and E7 genes of low- and high-risk cutaneous papillomaviruses reveals strongly transforming and high Rb-binding activity for the E7 protein of the low-risk human papillomavirus type 1.** *J Virol.* 1994; **68**(11): 7051–9.
PubMed Abstract | Free Full Text

25. Zhang B, Chen W, Roman A: **The E7 proteins of low- and high-risk human papillomaviruses share the ability to target the pRB family member p130 for degradation.** *Proc Natl Acad Sci U S A.* 2006; **103**(2): 437–42.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Ciccolini F, Di Pasquale G, Carlotti F, *et al.*: **Functional studies of E7 proteins from different HPV types.** *Oncogene.* 1994; **9**(9): 2633–8.
PubMed Abstract

27. Band V, Dalal S, Delmolino L, *et al.*: **Enhanced degradation of p53 protein in HPV-6 and BPV-1 E6-immortalized human mammary epithelial cells.** *EMBO J.* 1993; **12**(5): 1847–52.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Thomas M, Narayan N, Pim D, *et al.*: **Human papillomaviruses, cervical cancer and cell polarity.** *Oncogene.* 2008; **27**(55): 7018–30.
PubMed Abstract | Publisher Full Text

29. Glaunsinger BA, Lee SS, Thomas M, *et al.*: **Interactions of the PDZ-protein MAGI-1 with adenovirus E4-ORF1 and high-risk papillomavirus E6 oncoproteins.** *Oncogene.* 2000; **19**(46): 5270–80.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Nguyen ML, Nguyen MM, Lee D, *et al.*: **The PDZ ligand domain of the human papillomavirus type 16 E6 protein is required for E6's induction of epithelial hyperplasia *in vivo*.** *J Virol.* 2003; **77**(12): 6957–64.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Pim D, Bergant M, Boon SS, *et al.*: **Human papillomaviruses and the specificity of PDZ domain targeting.** *FEBS J.* 2012; **279**(19): 3530–7.
PubMed Abstract | Publisher Full Text

32. Lee SS, Weiss RS, Javier RT: **Binding of human virus oncoproteins to hDlg/SAP97, a mammalian homolog of the *Drosophila* discs large tumor suppressor protein.** *Proc Natl Acad Sci U S A.* 1997; **94**(13): 6670–5.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Kiyono T, Hiraiwa A, Fujita M, *et al.*: **Binding of high-risk human papillomavirus E6 oncoproteins to the human homologue of the *Drosophila* discs large tumor suppressor protein.** *Proc Natl Acad Sci U S A.* 1997; **94**(21): 11612–6.
PubMed Abstract | Publisher Full Text | Free Full Text

34. Ajiro M, Zheng ZM: **E6^E7, a novel splice isoform protein of human papillomavirus 16, stabilizes viral E6 and E7 oncoproteins via HSP90 and GRP78.** *mBio.* 2015; **6**(1): e02068–14.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Johansson C, Schwartz S: **Regulation of human papillomavirus gene expression by splicing and polyadenylation.** *Nat Rev Microbiol.* 2013; **11**(4): 239–51.
PubMed Abstract | Publisher Full Text

36. Ajiro M, Jia R, Zhang L, *et al.*: **Intron definition and a branch site adenosine at nt 385 control RNA splicing of HPV16 E6*I and E7 expression.** *PLoS One.* 2012; **7**(10): e46412.
PubMed Abstract | Publisher Full Text | Free Full Text

37. Tang S, Tao M, McCoy JP Jr, *et al.*: **The E7 oncoprotein is translated from spliced E6*I transcripts in high-risk human papillomavirus type 16- or type 18-positive cervical cancer cell lines via translation reinitiation.** *J Virol.* 2006; **80**(9): 4249–63.
PubMed Abstract | Publisher Full Text | Free Full Text

38. Zhao J, Lee EE, Kim J, *et al.*: **Transforming activity of an oncoprotein-encoding circular RNA from human papillomavirus.** *Nat Commun.* 2019; **10**(1): 2300.
PubMed Abstract | Publisher Full Text | Free Full Text

39. Shenvi CL, Dong KC, Friedman EM, *et al.*: **Accessibility of 18S rRNA in human 40S subunits and 80S ribosomes at physiological magnesium ion concentrations--implications for the study of ribosome dynamics.** *RNA.* 2005; **11**(12): 1898–908.
PubMed Abstract | Publisher Full Text | Free Full Text

40. Spahn CM, Jan E, Mulder A, *et al.*: **Cryo-EM visualization of a viral internal ribosome entry site bound to human ribosomes: the IRES functions as an RNA-based translation factor.** *Cell.* 2004; **118**(4): 465–75.
PubMed Abstract | Publisher Full Text

41. Spahn CM, Kieft JS, Grassucci RA, *et al.*: **Hepatitis C virus IRES RNA-induced changes in the conformation of the 40s ribosomal subunit.** *Science.* 2001; **291**(5510): 1959–62.
PubMed Abstract | Publisher Full Text

42. Gritsenko AA, Weingarten-Gabbay S, Elias-Kirma S, *et al.*: **Sequence features of viral and human Internal Ribosome Entry Sites predictive of their activity.** *PLoS Comput Biol.* 2017; **13**(9): e1005734.
PubMed Abstract | Publisher Full Text | Free Full Text

43. Van Doorslaer K, Li Z, Xirasagar S, *et al.*: **The Papillomavirus Episteme: a major update to the papillomavirus sequence database.** *Nucleic Acids Res.* 2017; **45**(D1): D499–D506.
PubMed Abstract | Publisher Full Text | Free Full Text

44. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks.** *Proc Natl Acad Sci U S A.* 1992; **89**(22): 10915–9.
PubMed Abstract | Publisher Full Text | Free Full Text

45. Auslander N: **A unique insert in the genomes of in high-risk human papillomaviruses with a predicted dual role in conferring oncogenic risk**. Harvard Dataverse, V1, UNF:6:xr7J2/tBIpkOW0NJ2yoqLA== [fileUNF], 2019.
http://www.doi.org/10.7910/DVN/FUGEUB

46. Nakagawa S, Huibregtse JM: **Human scribble (Vartul) is targeted for ubiquitin-mediated degradation by the high-risk papillomavirus E6 proteins and the E6AP ubiquitin-protein ligase.** *Mol Cell Biol.* 2000; **20**(21): 8244–53.
PubMed Abstract | Publisher Full Text | Free Full Text

47. Thomas M, Laura R, Hepner K, *et al.*: **Oncogenic human papillomavirus E6 proteins target the MAGI-2 and MAGI-3 proteins for degradation.** *Oncogene.* 2002; **21**(33): 5088–96.
PubMed Abstract | Publisher Full Text

48. Lee SS, Glaunsinger B, Mantovani F, *et al.*: **Multi-PDZ domain protein MUPP1 is a cellular target for both adenovirus E4-ORF1 and high-risk papillomavirus type 18 E6 oncoproteins.** *J Virol.* 2000; **74**(20): 9680–93.
PubMed Abstract | Publisher Full Text | Free Full Text

49. Lee C, Laimins LA: **Role of the PDZ domain-binding motif of the oncoprotein E6 in the pathogenesis of human papillomavirus type 31.** *J Virol.* 2004; **78**(22): 12366–77.
PubMed Abstract | Publisher Full Text | Free Full Text

50. Riley RR, Duensing S, Brake T, *et al.*: **Dissection of human papillomavirus E6 and E7 function in transgenic mouse models of cervical carcinogenesis.** *Cancer Res.* 2003; **63**(16): 4862–71.
PubMed Abstract

51. Luttermann C, Meyers G: **The importance of inter- and intramolecular base pairing for translation reinitiation on a eukaryotic bicistronic mRNA.** *Genes Dev.* 2009; **23**(3): 331–4.
PubMed Abstract | Publisher Full Text | Free Full Text

52. Powell ML, Napthine S, Jackson RJ, *et al.*: **Characterization of the termination-reinitiation strategy employed in the expression of influenza B virus BM2 protein.** *RNA.* 2008; **14**(11): 2394–406.
PubMed Abstract | Publisher Full Text | Free Full Text

53. Horvath CM, Williams MA, Lamb RA: **Eukaryotic coupled translation of tandem cistrons: identification of the influenza B virus BM2 polypeptide.** *EMBO J.* 1990; **9**(8): 2639–47.
PubMed Abstract | Publisher Full Text | Free Full Text

54. Ahmadian G, Randhawa JS, Easton AJ: **Expression of the ORF-2 protein of the human respiratory syncytial virus M2 gene is initiated by a ribosomal termination-dependent reinitiation mechanism.** *EMBO J.* 2002; **19**(11): 2681–9.
PubMed Abstract | Publisher Full Text | Free Full Text

55. Gould PS, Easton AJ: **Coupled translation of the second open reading frame of M2 mRNA is sequence dependent and differs significantly within the subfamily *Pneumovirinae*.** *J Virol.* 2007; **81**(16): 8488–96.
PubMed Abstract | Publisher Full Text | Free Full Text

56. Meyers G: **Translation of the minor capsid protein of a calicivirus is initiated by a novel termination-dependent reinitiation mechanism.** *J Biol Chem.* 2003; **278**(36): 34051–60.
PubMed Abstract | Publisher Full Text

57. Luttermann C, Meyers G: **A bipartite sequence motif induces translation reinitiation in feline calicivirus RNA.** *J Biol Chem.* 2007; **282**(10): 7056–65.
PubMed Abstract | Publisher Full Text

58. Meyers G: **Characterization of the sequence element directing translation reinitiation in RNA of the calicivirus rabbit hemorrhagic disease virus.** *J Virol.* 2007; **81**(18): 9623–32.
PubMed Abstract | Publisher Full Text | Free Full Text

59. Powell ML, Brown TD, Brierley I: **Translational termination-re-initiation in viral systems.** *Biochem Soc Trans.* 2008; **36**(Pt 4): 717–22.
PubMed Abstract | Publisher Full Text

60. Lorenz R, Bernhart SH, Höner Zu Siederdissen C, *et al.*: **ViennaRNA Package 2.0.** *Algorithms Mol Biol.* 2011; **6**(1): 26.
PubMed Abstract | Publisher Full Text | Free Full Text

61. Watson RA, Thomas M, Banks L, *et al.*: **Activity of the human papillomavirus E6 PDZ-binding motif correlates with an enhanced morphological transformation of immortalized human keratinocytes.** *J Cell Sci.* 2003; **116**(Pt 24): 4925–34.
PubMed Abstract | Publisher Full Text

62. James CD, Roberts S: **Viral Interactions with PDZ Domain-Containing Proteins-An Oncogenic Trait?** *Pathogens.* 2016; **5**(1): pii: E8.
PubMed Abstract | Publisher Full Text | Free Full Text

63. Javier RT, Rice AP: **Emerging theme: cellular PDZ proteins as common targets of pathogenic viruses.** *J Virol.* 2011; **85**(22): 11544–56.
PubMed Abstract | Publisher Full Text | Free Full Text

64. Thierry F: **Transcriptional regulation of the papillomavirus oncogenes by cellular and viral transcription factors in cervical carcinoma.** *Virology.* 2009; **384**(2): 375–9.
PubMed Abstract | Publisher Full Text

65. Gloss B, Bernard HU: **The E6/E7 promoter of human papillomavirus type 16 is activated in the absence of E2 proteins by a sequence-aberrant Sp1 distal element.** *J Virol.* 1990; **64**(11): 5577–84.
PubMed Abstract | Free Full Text

66. Smotkin D, Prokoph H, Wettstein FO: **Oncogenic and nononcogenic human genital papillomaviruses generate the E7 mRNA by different mechanisms.**

J Virol. 1989; **63**(3): 1441–7.
PubMed Abstract | Free Full Text

67. Zheng ZM, Baker CC: **Papillomavirus genome structure, expression, and post-transcriptional regulation.** *Front Biosci.* 2006; **11**: 2286–302.
PubMed Abstract | Publisher Full Text | Free Full Text

68. Tan TM, Gloss B, Bernard HU, *et al.*: **Mechanism of translation of the bicistronic mRNA encoding human papillomavirus type 16 E6-E7 genes.** *J Gen Virol.* 1994; **75**(Pt 10): 2663–70.
PubMed Abstract | Publisher Full Text

69. Zheng ZM, Tao M, Yamanegi K, *et al.*: **Splicing of a cap-proximal human Papillomavirus 16 E6E7 intron promotes E7 expression, but can be restrained by distance of the intron from its RNA 5' cap.** *J Mol Biol.* 2004; **337**(5): 1091–108.
PubMed Abstract | Publisher Full Text

70. Brant AC, Menezes AN, Felix SP, *et al.*: **Characterization of HPV integration, viral gene expression and *E6E7* alternative transcripts by RNA-Seq: A descriptive study in invasive cervical cancer.** *Genomics.* 2018; pii: S0888-7543(18)30658-X.
PubMed Abstract | Publisher Full Text

71. Brant AC, Majerciak V, Moreira MAM, *et al.*: **HPV18 Utilizes Two Alternative Branch Sites for E6*I Splicing to Produce E7 Protein.** *Virol Sin.* 2019; **34**(2): 211–221.
PubMed Abstract | Publisher Full Text | Free Full Text

72. Stacey SN, Jordan D, Snijders PJ, *et al.*: **Translation of the human papillomavirus type 16 E7 oncoprotein from bicistronic mRNA is independent of splicing events within the E6 open reading frame.** *J Virol.* 1995; **69**(11): 7023–31.
PubMed Abstract | Free Full Text

73. del Moral-Hernández O, López-Urrutia E, Bonilla-Moreno R, *et al.*: **The HPV-16 E7 oncoprotein is expressed mainly from the unspliced E6/E7 transcript in cervical carcinoma C33-A cells.** *Arch Virol.* 2010; **155**(12): 1959–70.
PubMed Abstract | Publisher Full Text

74. Mirabello L, Yeager M, Yu K, *et al.*: **HPV16 E7 Genetic Conservation Is Critical to Carcinogenesis.** *Cell.* 2017; **170**(6): 1164–1174.e6.
PubMed Abstract | Publisher Full Text | Free Full Text

75. Roman A, Munger K: **The papillomavirus E7 proteins.** *Virology.* 2013; **445**(1–2): 138–168.
PubMed Abstract | Publisher Full Text | Free Full Text

76. Alonso LG, García-Alai MM, Nadra AD, *et al.*: **High-risk (HPV16) human papillomavirus E7 oncoprotein is highly stable and extended, with conformational transitions that could explain its multiple cellular binding partners.** *Biochemistry.* 2002; **41**(33): 10510–8.
PubMed Abstract | Publisher Full Text

77. Katoh K, Misawa K, Kuma K, *et al.*: **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.** *Nucleic Acids Res.* 2002; **30**(14): 3059–66.
PubMed Abstract | Publisher Full Text | Free Full Text

78. Guindon S, Lethiec F, Duroux P, *et al.*: **PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference.** *Nucleic Acids Res.* 2005; **33**(Web Server issue): W557–9.
PubMed Abstract | Publisher Full Text | Free Full Text

79. Cortes C, Vapnik V: **Support-Vector Networks.** *Mach Learn.* 1995; **20**(3): 273–97.
Publisher Full Text

80. Ogurtsov AY, Shabalina SA, Kondrashov AS, *et al.*: **Analysis of internal loops within the RNA secondary structure in almost quadratic time.** *Bioinformatics.* 2006; **22**(11): 1317–24.
PubMed Abstract | Publisher Full Text

81. Mathews DH: **Revolutions in RNA secondary structure prediction.** *J Mol Biol.* 2006; **359**(3): 526–32.
PubMed Abstract | Publisher Full Text

82. Faure G, Ogurtsov AY, Shabalina SA, *et al.*: **Role of mRNA structure in the control of protein folding.** *Nucleic Acids Res.* 2016; **44**(22): 10898–10911.
PubMed Abstract | Publisher Full Text | Free Full Text

83. Matveeva OV, Shabalina SA: **Intermolecular mRNA-rRNA hybridization and the distribution of potential interaction regions in murine 18S rRNA.** *Nucleic Acids Res.* 1993; **21**(4): 1007–11.
PubMed Abstract | Publisher Full Text | Free Full Text

84. Ogurtsov AY, Mariño-Ramírez L, Johnson GR, *et al.*: **Expression patterns of protein kinases correlate with gene architecture and evolutionary rates.** *PLoS One.* 2008; **3**(10): e3599.
PubMed Abstract | Publisher Full Text | Free Full Text

85. Auslander N, Wolf IY, Shabalina AS, *et al.*: **Permutation test controlling for HPV strains.** *Zenodo.* 2019.
http://www.doi.org/10.5281/zenodo.3242231

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 2

Reviewer Report 16 October 2019

https://doi.org/10.5256/f1000research.22707.r54507

✓ **Alison McBride**
DNA Tumor Virus Section, National Institute of Allergy and Infectious Diseases (NIAID), Bethesda, MD, USA

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 08 October 2019

https://doi.org/10.5256/f1000research.22707.r54508

✓ **Elizabeth A. White**
Department of Otorhinolaryngology: Head and Neck Surgery, University of Pennsylvania Perelman School of Medicine, Philadelphia, USA

I approve the manuscript in its revised form.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: HPV E6 and E7, protein-protein interactions, transformation

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 06 August 2019

https://doi.org/10.5256/f1000research.21480.r51570

✓ **Elizabeth A. White**
Department of Otorhinolaryngology: Head and Neck Surgery, University of Pennsylvania Perelman School of Medicine, Philadelphia, USA

In this study, Auslander and colleagues compare nucleotide and amino acid sequences for a subset of genus alpha HPV genotypes. They observe that high-risk HPV genomes share a short sequence insert at the 3' end of the E6 ORF. The insertion adds a PDZ binding motif to the C-terminus of high-risk HPV E6 proteins and alters the location of the E6 termination codon relative to the E7 initiation codon.

This observation is consistent with previous findings. It has been appreciated for some time that high-risk HPV E6 and E7 are transcribed from a polycistronic mRNA whereas low-risk E6 and E7 are transcribed from separate promoters. In addition, a frequent splicing event occurs within the E6 gene in the bicistronic high-risk HPV early mRNA. This report adds to those observations. It proposes that the 3' high-risk specific insert is another feature that differs between high- and low-risk HPV and that it might drive differences in expression of high-risk versus low-risk HPV oncoproteins. Understanding the differences between oncogenic and non-oncogenic HPV is an area of intensive research and new contributions in this area are potentially significant. This report makes a useful contribution to the literature.

Weaknesses of the manuscript are that the current literature is not cited and that other features of high-risk HPV E6 that might also account for their oncogenic activities are not discussed. Although comparing nucleotide sequence differences is informative, for HPV E6 this comparison does not completely reflect the biology of the proteins. There are high-risk HPV E6-specific protein binding partners other than PDZ domain proteins. Several of these are listed in a useful 2012 review[1]; others were identified by proteomic analyses from several groups[23]. For example, the authors do not mention that TP53 binding and degradation is a feature of high-risk HPV E6 not shared by low-risk HPV E6. It also appears that not all of the high-risk HPV E6 interact with the same subset of cellular PDZ proteins. It is unclear how these diverse HPV E6-PDZ protein interactions might account for the shared oncogenic features of high-risk HPV E6, and this point is not discussed in the manuscript.

It is established in the HPV literature that small differences in HPV oncoprotein amino acid sequence result in significant differences in interactions with host cellular proteins. This is beautifully illustrated by the structural studies of Gilles Trave and colleagues, who have determined that subtle differences in E6 enable a range of specific interactions with cellular LxxLL-containing proteins. Other recent studies highlight high-risk HPV-specific activities of the E7 oncoprotein. In light of findings like these, this manuscript seems to overstate the claim that the PDZ binding motif is the 'most notable molecular feature' distinguishing high- from low-risk HPV. The potential importance of the insert sequence to E6/E7 translation regulation is high and should be tested; the discussion of the importance of the PDZ binding motif should be tempered and put in context with other recent findings.

Additional points:

1. Mirabello and colleagues recently reported an analysis of sequence variants from >5000 HPV16-positive cervical samples[4]. HPV16 E6 sequences exhibited variation across the length of the ORF that was similar in high-grade versus control lesions. A possible interpretation of this finding is that the protein sequence in the C-terminus of E6 is not important for oncogenic transformation. How does this fit into the authors' model? It would be useful to include a discussion of these data.

2. The phylogenetic trees in Figure 1B might be easier to interpret if they were presented in a linear format. It is difficult to determine where the boundaries between the high-risk/low-risk groupings overlap with the major branch points of the tree.

3. HPV diversity is much greater that that reflected solely in genus alpha. It would be interesting to know whether HPV from other genera have also acquired sequences in this region that might provide some or all of the activities suggested by the authors.

References

1. Klingelhutz AJ, Roman A: Cellular transformation by human papillomaviruses: lessons learned by comparing high- and low-risk viruses.*Virology*. 2012; **424** (2): 77-98 PubMed Abstract | Publisher Full Text

2. Rozenblatt-Rosen O, Deo RC, Padi M, Adelmant G, Calderwood MA, Rolland T, Grace M, Dricot A, Askenazi M, Tavares M, Pevzner SJ, Abderazzaq F, Byrdsong D, Carvunis AR, Chen AA, Cheng J, Correll M, Duarte M, Fan C, Feltkamp MC, Ficarro SB, Franchi R, Garg BK, Gulbahce N, Hao T, Holthaus AM, James R, Korkhin A, Litovchick L, Mar JC, Pak TR, Rabello S, Rubio R, Shen Y, Singh S, Spangle JM, Tasan M, Wanamaker S, Webber JT, Roecklein-Canfield J, Johannsen E, Barabási AL, Beroukhim R, Kieff E, Cusick ME, Hill DE, Münger K, Marto JA, Quackenbush J, Roth FP, DeCaprio JA, Vidal M: Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins.*Nature*. 2012; **487** (7408): 491-5 PubMed Abstract | Publisher Full Text

3. White EA, Kramer RE, Tan MJ, Hayes SD, Harper JW, Howley PM: Comprehensive analysis of host cellular interactions with human papillomavirus E6 proteins identifies new E6 binding partners and reflects viral diversity.*J Virol*. 2012; **86** (24): 13174-86 PubMed Abstract | Publisher Full Text

4. Mirabello L, Yeager M, Yu K, Clifford GM, Xiao Y, Zhu B, Cullen M, Boland JF, Wentzensen N, Nelson CW, Raine-Bennett T, Chen Z, Bass S, Song L, Yang Q, Steinberg M, Burdett L, Dean M, Roberson D, Mitchell J, Lorey T, Franceschi S, Castle PE, Walker J, Zuna R, Kreimer AR, Beachler DC, Hildesheim A, Gonzalez P, Porras C, Burk RD, Schiffman M: HPV16 E7 Genetic Conservation Is Critical to Carcinogenesis.*Cell*. 2017; **170** (6): 1164-1174.e6 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?
I cannot comment. A qualified statistician is required.

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: HPV E6 and E7, protein-protein interactions, transformation

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response (*Member of the F1000 Faculty* and *F1000Research Advisory Board Member*) 13 Sep 2019
Eugene Koonin, National Institutes of Health, USA, Bethesda, USA

We appreciate this highly constructive and helpful review in response to which the following modifications have been made too the article:

1. Many references to the current literature were added, and several other features of HR-HPV E6 are now discussed.

2. The suggested references to HR-HPV E6 specific binding partners have been included.

3. The discussion on the importance of the PDZ binding motif in distinguishing HR from LR HPV has been tempered, and the context for its potential relevance has been clarified.

4. We now discuss the recent analysis of 5570 cervical HPV16 genomes (Mirabello and colleagues) and have substantially expanded the discussion of the effect of the insert on E7 production.

5. Figure 1B has been updated, phylogenetic trees are now presented in a linear format.

In addition, although the referee have suggested a statistical review of the permutation test that was used in this work to assess the significance of the results, we are confident that this is unnecessary because the statistical technique we used is simple and standard.

Competing Interests: I declare no competing interests

Reviewer Report 31 July 2019

https://doi.org/10.5256/f1000research.21480.r51173

✓ **Alison McBride**
DNA Tumor Virus Section, National Institute of Allergy and Infectious Diseases (NIAID), Bethesda, MD, USA

Auslander and colleagues present a comparative sequence analysis of alpha-papillomavirus genomes. They identify and analyze a short 30-60 nucleotide sequence insert between the E6 and E7 open-reading frames that encodes the PDZ domain in the high-oncogenic risk types and propose that this region also contains sequences that enable coupled termination and reinitiation to facilitate translation of the E6 and E7 proteins. This is an intriguing observation as HPVs are associated with ~5% human cancers and understanding how the viral oncoproteins are expressed is crucial. The proposed hypothesis should be testable and indicates that experiments in which the E6 PDZ domain is mutated in the background of the viral genome should be interpreted carefully.

A weakness of the manuscript is that current literature and data sources are not used/and or cited and the data-set used seems incomplete. For example, there are currently 198 officially numbered HPV types, and 442 HPV types in total. A curated data-set of genome sequences for all papillomavirus types sequenced to date is available at PaVE. There are also many recent publications that describe analyses of PV evolution, oncogenicity and E6 PDZ domains that should be cited.

There are sequences available for 64 alpha HPV types that have been officially named by the HPV Reference Centre, yet only the genomic sequences of 44 types are listed in Figure S1. The data-set contains many isolates for some HPV types (e.g. 299 isolates of HPV16) yet almost no representatives of the species alpha 2, 4 and 8. The E6 protein of HPV40 (alpha-8) has been proposed to contain an ancestral alpha PDZ domain and so the genomes of the alpha-8 species should be closely examined/discussed. Nevertheless, a nucleotide alignment of all 64 alpha-PV nucleotide sequences from PaVE does support the authors' conclusions that only HR species-5, 6 7, 9 and 11 contain an insert separating the E6 and E7 ORFs.

Minor issues:

1. In Figure 1A, alpha-11 is listed in the key for both LR and HR.

2. In Figure 2, the resolution of the sequence images should be improved. The legend for 2B should make it clear that these are just the C-terminal sequences of the E6 polypeptides. In 2C, the symbols and colored blocks should be described.

3. In Figure 3A, the color/bold identification of the motifs and TAA/AUG are not easily visualized in the pdf. Perhaps highlight TAA/AUG by underlines or boxes. Labelling the groups of different HPV species along the left would also be helpful. In 3B, the labelling of the 18S rRNA hairpin (orange text) is confusing. What is 26es7?

4. In Figure S4: clarify in the legend that this is the E6-E7 RNA sequence for each HPV type shown.

5. The reference for IRES (54), I think should be 53.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
I cannot comment. A qualified statistician is required.

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: I founded and oversee the NIAID papillomavirus bioinformatics site https://pave.niaid.nih.gov/ described in the review

Reviewer Expertise: HPV replication, genomics, epigenetics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response (*Member of the F1000 Faculty* and *F1000Research Advisory Board Member*) 13 Sep 2019
Eugene Koonin, National Institutes of Health, USA, Bethesda, USA

We appreciate the constructive and very helpful review in response to which the following changes have been made to the manuscript:

1. We now mention the most recent number of HPV types that have been formally recognized (198), and the reference supporting this has been updated.

2. We added numerous references to recent publications that describe analyses of HPV evolution, oncogenicity, oncoprotein interactions and E6 interactions with PDZ domains.

3. We now include all 64 alpha HPV types from PaVE, and all analyses and figures 1-3 include those HPV strains.

4. Resolution of the new figures has been improved.

5. The legend to Figure 2B has been modified as suggested.

6. The legend to Figure 2C describes all symbols and colored blocks.

7. Figure 3A has been updated to show the motifs clearly and arrange the strains by the order in the phylogenetic tree.

8. Figure 3B has been updated, confusing labelling of 18S rRNA removed.

Competing Interests: I declare no competing interests.

Author Response (*Member of the F1000 Faculty* and *F1000Research Advisory Board Member*) 13 Sep 2019

Eugene Koonin, National Institutes of Health, USA, Bethesda, USA

Although the referee has suggested a statistical review of the permutation test that was used in this work to assess the significance of the results, we are confident that this is unnecessary because the statistical technique we used is simple and standard.

Competing Interests: I declare no competing interests.