HUMAN PAPILLOMAVIRUS TYPE 16 SUB-LINEAGES AND INTEGRATION IN CANCER

Robert Jackson[1,2,*,#], **Alejandro Ortigas-Vásquez**[2,3], **Ingeborg Zehbe**[2,4,5]

[1]Biotechnology Program, Lakehead University; [2]Probe Development & Biomarker Exploration, Thunder Bay Regional Health Research Institute; [3]Bioinformatics Program; [4]Department of Biology; [5]Northern Ontario School of Medicine, Lakehead University, Thunder Bay, Canada

[*]**Corresponding author:** Robert Jackson, PhD (e-mail: rjackso1@lakeheadu.ca)

[#]**Affiliation history:** Robert Jackson was affiliated with Lakehead University and the Thunder Bay Regional Health Research Institute at the time the study was performed and is currently affiliated with the University of Arizona.

KEYWORDS

The Cancer Genome Atlas (TCGA), Human papillomavirus type 16 (HPV16), Sub-lineage, Variants, Integration, Cervical cancer, Head and neck cancer

SHORT TITLE

HPV16 sub-lineage integration

ABSTRACT

Our lab has been intrigued by the fact that viral genomes often take on the role of mobile elements to perpetuate their existence in a complex organism's genome. Multiple DNA viruses such as Epstein-Barr virus, hepatitis B virus, and human papillomavirus (HPV) can invade their host genome, as "genomic parasites". We have been investigating HPV type 16 (HPV16), which is a prominent human tumour virus. In our recent *in vitro* work using 3D organoids, a common variant of HPV16's coding region elicited early integration into the host genome compared to the prototype. Next-generation sequencing (NGS) data confirmed a transcriptomic profile of increased proliferation and chromosomal instability—both hallmarks of cancer. Epidemiologically, this variant is associated with a high cervical cancer incidence. To substantiate *in vitro* findings and test variant-specific integration across HPV16-related cancers, we employed NGS data from population-derived clinical samples in The Cancer Genome Atlas (TCGA)-curated database. Data were analyzed for HPV16 positivity, sub-lineage, and viral-host integration using a bioinformatic pipeline of open-source tools, including HPVDetector. Here, we report the initial analysis for a subset of 37 cervical cancer cases comprised of HPV16 positive and negative samples as well as different sub-lineage and integration states. We do not presently find an association between sub-lineage and integration, but these determinations are confounded by discrepancies between DNA- and RNA-level data and the small sample size. We conclude by discussing challenges and future directions for analyzing the remaining TCGA cases—a requirement to increase statistical power. The overall goal is to investigate integration patterns quantitatively and qualitatively as well as to confirm or refute our initial hypothesis based on clinical sample data. Dissecting integration patterns of HPV sub-lineage genomes may uncover viral-host interactions evolutionarily related to transposable elements—a group of molecules increasingly identified in a wide array of functions within humans. The overall understanding of risk factors related to HPV sub-lineage could be relevant for prognostic and cancer treatment efficiency biomarkers.

INTRODUCTION

The co-evolutionary interplay between pathogens and their hosts involves a plethora of intriguing biological phenomena, including genomic integration of pathogen DNA into host cells. While cellular integration of viral genomic material is a means of propagation for families such as *Retroviridae* (*e.g.*, human immunodeficiency virus, HIV), it can also occur during persistent infections with DNA viruses and is often associated with virally induced cancers. One such family of common DNA tumour viruses, *Papillomaviridae*, is responsible for causing ~5% of all human cancers worldwide [Ghittoni *et al*., 2015]. There are 660 types of papillomaviruses (PVs) discovered as of June 2020 based on the PaVE reference genome database [PaVE: https://pave.niaid.nih.gov, Van Doorslaer *et al*., 2013; 2017a]. The number of PV types, including the non-human PVs (currently 215) is expected to continue rising due to high-throughput genomics and sampling [Van Doorslaer and Dillner, 2019]. The majority of PVs currently identified are the human papillomaviruses (HPVs) [Van Doorslaer *et al*., 2017b] and these types (445 to date) differ in their propensity for inducing cancer as well as their epithelial tropism, with “high-risk” types being the primary cause of malignancies: predominantly via sexually-transmitted persistent HPV type 16 (HPV16) infection of anogenital and oropharyngeal mucosa [Ndiaye *et al*., 2014]. Although this and other HPV genomes are considered “small”, only ~7.9 kb containing nine open reading frames (ORFs) and a variety of alternatively spliced and multi-cistronic variant transcripts [Graham & Faizo, 2017], it encodes potent oncoproteins such as E6 and E7 which drive host cells to acquire the hallmarks of cancer [Hanahan and Weinberg, 2000; 2011; Mesri *et al*., 2014]. The combined E6 and E7 actions disable protective apoptotic mechanisms and promote host cell proliferation as outlined by Hoppe-Seyler and colleagues [2018]—a phenomenon that could potentially be exploited therapeutically at an RNA [Togtema *et al*., 2018] or protein-level [Togtema *et al*., 2019]. Of particular relevance for this study, the HPV-augmented environment can permit chromosomal instability (an enabling characteristic among the hallmarks

of cancer) and eventual integration of HPV16 sequences into human DNA via double-strand break (DSB) repair events [Winder *et al*., 2007], yielding unique integration signatures found in next-generation sequencing (NGS) data from human carcinomas [Holmes *et al*., 2016].

HPV16 sub-lineages (variants) via epidemiological and lab-based studies have been reported to infer differing cancer risk [reviewed by Mirabello *et al*., 2018 and references within; Clifford *et al*., 2019]. HPV16 variant designations were originally based on their geographical region of origin and have since been updated to lineage (1.0 to 10.0% whole viral genome sequence difference) and sub-lineage (0.5 to 1.0% difference) nomenclature. The European "prototype" sub-lineage (EP; now A1 according to new nomenclature) was the first HPV16 genome published [Seedorf *et al*., 1985]. The remaining sub-lineages include two additional European groups (E; now A2 and A3), one Asian (As; now A4), eight African (Af-1a, -1b and -2; now B1-4 and C1-4), two North-American (NA; now D1 and D4) and two Asian-American (AA-1 and -2; now D3 and D2, respectively) [Burk *et al*., 2013; Mirabello *et al*., 2018]. We have put our efforts toward comparing the EP and AA sub-lineage variants, which differ by only three amino acid changes at residues 14 (Q to H), 78 (H to Y) and 83 (L to V) in the major transforming protein E6 [Richard *et al*., 2010; Niccoli *et al*., 2012; Jackson *et al*., 2014; 2016; Togtema *et al*., 2015; Cuninghame *et al*., 2017]. Epidemiological studies revealed that the AA sub-lineage infers a higher risk for dysplasia and an earlier onset of invasive tumours than EP [Xi *et al*., 1997; Berumen *et al*., 2001]. Our functional assays showed that AAE6 has a greater transforming, migratory, and invasive potential than EPE6 when the respective E6 gene was retrovirally transduced into primary human keratinocytes in long-term *in vitro* immortalization studies [Richard *et al*., 2010; Niccoli *et al*., 2012; Togtema *et al*., 2015]. These observations may be due, at least in part, to an AAE6-mediated altered metabolic phenotype reminiscent of the Warburg effect [Richard *et al*., 2010; Cuninghame *et al*., 2017]. Consistent with the hypothesis that AAE6 is more oncogenic than EPE6, our recent experimental work revealed that AAE6 in the context of the full-length viral

genome is more prone to host genome integration than is EPE6 [Jackson *et al*., 2014; 2016]. At that time, one study mentioned HPV16 sub-lineage variants and integration but did not find a difference ($P = 0.28$, two-tailed Fisher's exact test) between EPE6 (3 episomal and 20 integrated cases) and the E-T350G variant (6 episomal and 16 integrated cases, responsible for one of the residue changes also found in AAE6: L83V) [Xu *et al*., 2013]. Notably, however, the lone tumour sample that contained the AA variant was in integrated form.

Based on our hypothesis that the propensity for HPV16 integration may vary between sub-lineages, we set out to do a search of clinical samples with available genomic and transcriptomic data reported in databases. Two landmark publications of The Cancer Genome Atlas (TCGA) reported on cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) [TCGA Research Network, 2017] and head and neck squamous cell carcinoma (HNSC) [TCGA Research Network, 2015]. More recently, Cantalupo *et al*. [2018] reported further analysis of additional TCGA data, including additional cancer sets that have become available in recent years. They also used a clearly defined, rules-based approach to filtering and detection: first extracting high-quality non-human reads followed by removal of non-relevant viral families (bacteriophage, plant viruses, insect viruses) and artifacts, with 10 or more high-quality alignments required for virus detection. Prior to the primary TCGA data analysis papers, there was a publication about DNA viruses in TCGA RNA-Seq data across a variety of human cancers [Khoury *et al*., 2013]. An additional literature search found several more related papers: HPV and HNSC comprehensive analysis [Castellsagué *et al*., 2016] and an HPV16-focused follow-up TCGA-HNSC analysis [Nulton *et al*., 2017]. In the latter, three different groupings of integration status were reported: episomal, integrated, and human-viral episomal hybrids, challenging integration data by Parfenov *et al*. [2014]. The Parfenov *et al*. study disclosed HPV typing and sub-lineage analyses of the 35 HPV+ HNSC cases with 29 containing HPV16: one of which was an AA variant. Interestingly, integration with 16 breakpoints in the human genome (*i.e.*, ~ 4-fold more compared to EP) was

shown. In the Nulton *et al*. study, for the same case, a new phenomenon of human-viral hybrids in the host genomic DNA subsequently found in episomal form was reported [Nulton *et al*., 2017]. Recently, Zapatka *et al*. [2018] further analyzed TCGA data to determine viral associations with cancer, including the impact of integrations on host copy number variations. Most recently, Lou *et al.* [2020] report that integration rates indeed differ between sub-lineages in cervical cancers. In a Guatemalan population with a 24%, 31%, and 45% prevalence of D2, D3, and A1 sub-lineages, respectively, they found that D2 had a higher rate of integration than D3 (76% vs 44%), coupled with an earlier age of cancer diagnosis. Interestingly, they also found that A1 had an equally high rate of integration (78%) as D2, which opposes our specific hypothesis (at least, within this specific population). It is noteworthy that A1 is phylogenetically distant from D2/D3, with ~2% genome difference (>150 single-nucleotide polymorphisms, SNPs, and insertions/deletions, indels), whereas D2 and D3 differ from each other by only ~0.5% (<40 SNPs and indels, and hence it remains open what the biological meaning for these findings may be.

Our overall objective is to screen TCGA for all HPV16-containing samples, verify their sub-lineage genotypes, then detect and characterize instances of viral integration along with a complete picture of its impact on the host genome and identify potential contributing factors in the surrounding molecular landscape. The most pertinent research question for us is whether there are specific integration mechanisms associated with different HPV16 sub-lineage genotypes: is the AA variant (D2 and D3) more frequently associated with integration than the EP variant (A1) and/or are there unique integration patterns with potential functional relevance? Such a finding would resonate with our variant-specific integration hypothesis [Jackson *et al*., 2014; 2016] and provide evidence for a novel mechanism in differing variant tumourigenesis. Our approach of HPV16 sub-lineage genotyping and integration detection, along with an attempt to reconcile and verify existing data, has not been previously addressed for TCGA data. We analyzed a subset of 37 cases, including DNA- and RNA-level data, but do not find an association between HPV16

sub-lineage and integration. However, while a variety of bioinformatics pipelines exist for sub-lineage genotyping and integration detection, we find they are susceptible to biological (*e.g.*, low copies and incomplete viral coverage) and technical (*e.g.*, inadequate read depth) hurdles, which are important considerations for future analyses.

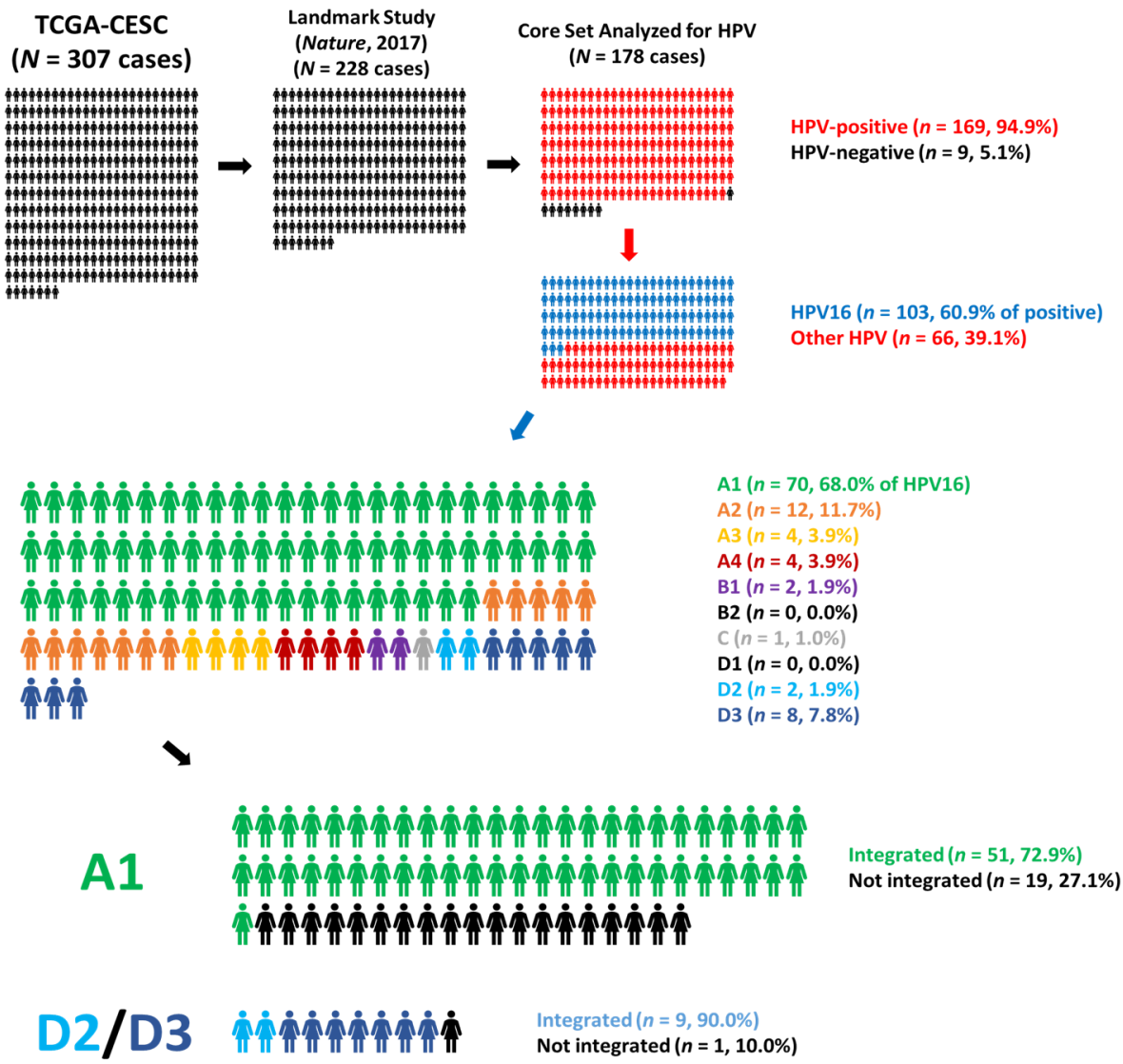
RESULTS & DISCUSSION

HPV16 D2/D3 integration data are limited for cervical and head and neck cancers

When seeking to test the relationship between HPV16 sub-lineage and integration propensity we started by compiling a sample summary of previously published and available data from TCGA [**Table 1**]. Next, we performed a preliminary sub-lineage and integration analysis for available CESC sample data [**Figure 1**] as it was stratified into sub-lineage for the A lineage, whereas HNSC data were not further stratified between A1, A2, and A3. Overall, while the proportion of integrated cases was 1.23x higher in D2/D3-containing CESC samples (9/10, 90.0% vs 51/70, 72.9% for A1), the sample size was very low rendering this initial analysis underpowered and the finding as not statistically significant (P = 0.242, X^2 test).

Table 1. Sample summary breakdown for each TCGA case set. Sample genotypes for cancer cases from the Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (CESC) as well as Head and Neck Squamous Cell Carcinoma (HNSC) datasets. CESC HPV16 data are based on TCGA Research Network [2017] (using the sub-lineages identified at the time, which were only 10 of the current 16). *HNSC HPV16 data are based on Parfenov *et al*. [2014], where the A sub-lineage was grouped as "EUR" genotypes without specifying A1, A2 or A3.

Sample genotype	**CESC (*n* = 178)**	**HNSC (*n* = 279)**
Total HPV16 cases	103 (57.9% of CESC cases)	29 (10.4% of HNSC cases)
A1	70 (68.0% of HPV16 cases)	
A2	12 (11.7%)	22 (75.9% of HPV16 cases)*
A3	4 (3.9%)	
A4	4 (3.9%)	2 (6.9%)
B1	2 (1.9%)	3 (10.3%)
B2	–	–
C	1 (1.0%)	1 (3.4%)
D1	–	–
D2	2 (1.9%)	–
D3	8 (7.8%)	1 (3.4%)



	Integrated	Not Integrated	
A1	51	19	70
D2/D3	9	1	10
	60	20	80

Initial X^2 test = 1.3714, P = 0.241567 (not significant)
1.23x (+23%) higher integration proportion in D2/D3 vs A1 cases

Figure 1. Preliminary analysis of CESC samples. Previously available HPV16 sub-lineage and integration data for the CESC dataset, published by the TCGA Research Network [2017], were analyzed to determine whether there was a higher proportion of integrated samples

corresponding to D2/D3 (which have viral genomes that differ from each other by ~0.5%, but have identical E6 amino-acid sequences) vs the A1 sub-lineage (which differs ~2% from D2/D3 viral genomes). While these sub-lineage classifications were based on 10 reference genomes, the set has since been expanded to 16 sub-lineages, as described in Mirabello *et al*. [2018].

The untapped potential of TCGA

To expand upon these preliminary results, we are using our own HPV16 detection and sub-lineage analyses to verify sub-lineage calls by prior studies, resolve discrepancies, and further analyze the "unknown" samples. From the National Cancer Institute's Genomic Data Commons (GDC) Data Portal, there are currently 835 relevant CESC and HNSC cases available for analysis, including 2,568 binary alignment map (BAM) files (27.84 TB total, an average of 33.34 GB/case, including all whole-exome sequencing (WXS) and RNA-Seq sequencing read files and excluding miRNA-Seq files as they are not primarily relevant for HPV16 genotyping). There are 307 CESC cases (924 BAM files, 10.44 TB) and 528 HNSC cases (1,644 BAM files, 17.41 TB), with both sets including samples from primary tumours, metastases, normal blood, and normal solid tissues. It appears the TCGA legacy data portal also contains whole-genome sequencing (WGS) files for these cancer sets. Based on the TCGA landmark studies [TCGA Research Network, 2015; 2017], we expect the majority of CESC samples are HPV positive, whereas a smaller proportion of HNSC samples are HPV positive. While HPV16 is expected to be the predominant type, multi-type infections and co-infections with other non-commensal microorganisms are also possible. As with most nucleic-acid based techniques, cross-contamination between samples is a potential source of confusion, especially for adjacent cases where sequence libraries could be contaminated [Cantalupo *et al*., 2015].

The number of CESC and HNSC samples that were previously marked as being positive for HPV16 differ between studies, likely due to differences in analysis pipelines (*e.g.*, data types

analyzed, thresholding, and filtering for contamination) [TCGA Research Network, 2017; Cantalupo *et al*., 2018]. While both RNA-Seq and WXS were used for integration analysis in the landmark CESC study, it seems that the authors mostly if not totally relied on RNA-Seq [TCGA Research Network, 2017]. Moreover, for integration status analysis, an in-house tool was used, which cannot be directly compared with other published and available tools such as ViFi [Nguyen *et al*., 2018], HPVDetector [Chandrani *et al*., 2015] or VirTect [Xia *et al*., 2019] (and others detailed in the **METHODS** section). This poses major challenges to a current, valid analysis of sub-lineage association with integration based on these previous publications.

HPV16 positivity quantification reveals discrepancy between RNA-Seq and WXS samples

We analyzed both RNA-Seq and WXS CESC samples independently to detect the presence of HPV16 in a subset of 37 cases representing all available D2 and D3 cases identified in the landmark study, and a similar-sized subset of A cases, a B case, as well as negatives to determine how robust our methods are [**Figure 2A**]. As expected, control sample SiHa was a low-positive (5.51 VRPM) and CaSki was a high-positive (502.66 VRPM). TCGA RNA-Seq samples had a range of 0.00 to 482.08 VRPM, whereas WXS samples ranged from 0.00 to 9264.87 VRPM. In total, 23/37 (62.2%) of the cases we analyzed were HPV16 positive based on RNA-Seq and used for subsequent sub-lineage and integration analysis. However, for the same cases, DNA and RNA level HPV16 quantification were often inconsistent, with only 15/37 (40.5%) cases yielding consistent HPV16 quantification (negative, low-positive, or high-positive).

To resolve these discrepancies, we compared our data for both RNA-Seq and WXS samples to data from the landmark CESC study [TCGA Research Network, 2017], summarized in [**Table 2**]. We found that positivity results for RNA-Seq samples were the most consistent with previous findings, as all cases that had been reported as negative for HPV16 were also found negative in our analysis at the RNA level, but not necessarily at the DNA level, where we found 7

WXS samples above the low-positive threshold (2 of which are even above the high-positive threshold). Interestingly, 9/10 (90.0%) of the D lineage samples were positive at both the DNA and RNA level, whereas only 4/12 (33.3%) of the A lineage samples analyzed were positive at both DNA and RNA levels. These discrepant findings could be due to biological or technical phenomena, such as insufficient read depth, too few viral reads (small amount of viral DNA in a subset of cells, relative to host DNA), or no/weak transcription of integrated viral sequences (in the cases where RNA-Seq signal is lacking). These issues could be further compounded by computational artefacts and potential contamination viral sequences, which makes it critical (yet challenging) to develop tools that successfully distinguish contaminating from authentic viral sequences [Cantalupo *et al*., 2018].

Overall, for this subset of 37 cases there was a very weak positive correlation between DNA and RNA HPV16 levels (R^2 = 0.09) [**Figure 2B**]. Lineage-specific HPV16 levels were compared for DNA and RNA, with RNA-levels similar between A and D lineages, but DNA-levels were highest in the D lineage [**Figure 2C**].

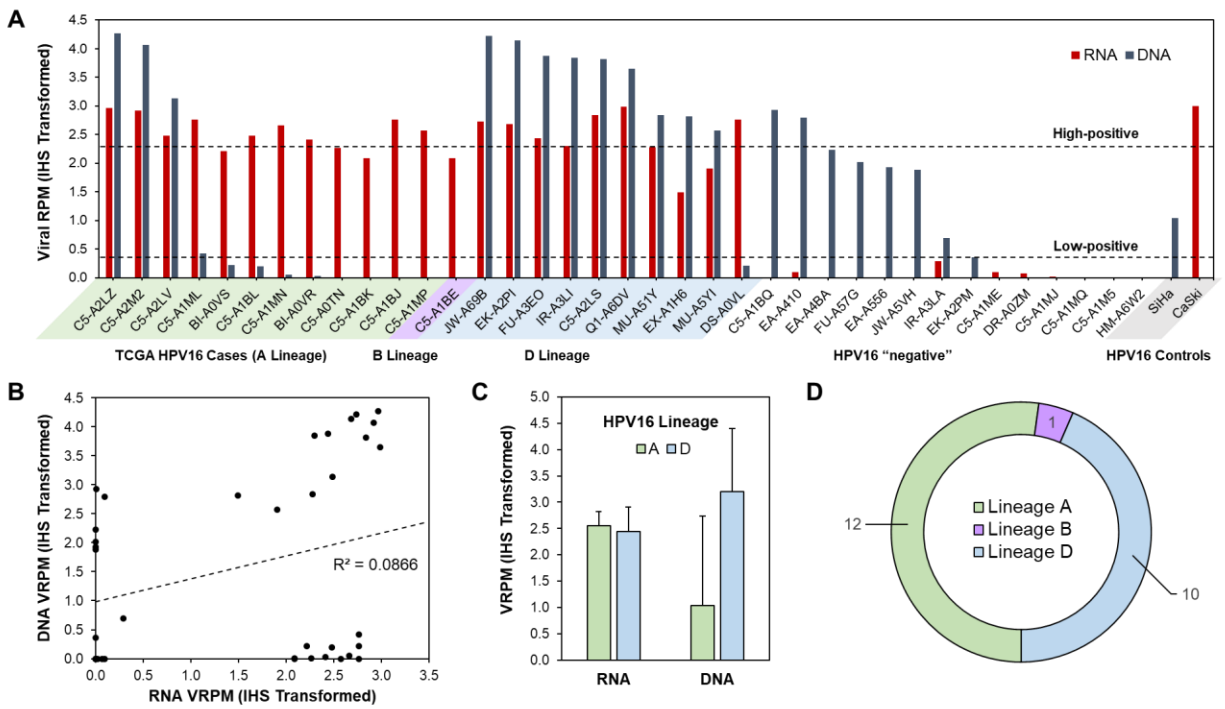


Figure 2. HPV16 positivity in RNA-Seq and WXS TCGA samples. (A) A total of 74 samples were analyzed to determine whether they were positive for HPV16, with 37 corresponding to RNA-Seq and 37 corresponding to WXS samples. Samples with less than 0.99 viral reads per million (RPM) (marked by the lowest dashed line) were classified as negative. Positive samples were then subdivided based on strength of positivity, with RPM values below 99 (marked by the highest dashed line) indicating low positivity and those above 99 indicating high positivity. TCGA case IDs are colour-coded based on their CESC landmark study lineage determination (A = green, B = purple, and D = blue). Cervical cancer control cell lines, SiHa and CaSki (shaded grey), were included as low- and high-positive controls, respectively. Values are inverse hyperbolic sine (IHS) transformed. **(B)** Correlation plot of RNA vs DNA-level HPV16 VRPM. **(C)** Mean positivity for HPV16 lineage A vs D. Error bars represent standard deviation. **(D)** HPV16 lineage pie chart, for the 23/37 positive cases analyzed.

Table 2. Summary comparison table for HPV16 positivity, lineage, and integration. Landmark CESC study results [TCGA Research Network, 2017] were compared to our current findings.

Case ID	HPV16 Positivity		Lineage		Integration Status	
	Landmark	*Current*	*Landmark*	*Current*	*Landmark*	*Current*
TCGA-DS-A0VL	Positive	High	D2	D	Integrated	Integrated
TCGA-C5-A2LS	Positive	High	D3	D	Integrated	Episomal
TCGA-C5-A1BE	Positive	Low	B1	B	Integrated	Integrated
TCGA-BI-A0VS	Positive	Low	A2	A	Integrated	Integrated
TCGA-BI-A0VR	Positive	High	A1	A	Episomal	Episomal
TCGA-C5-A1MQ	Negative	Negative	-	-	-	-
TCGA-C5-A1BQ	Negative	Negative	-	-	-	-
TCGA-C5-A1ML	Positive	High	A2	A	Episomal	Episomal
TCGA-C5-A2LZ	Positive	High	A2	A	Episomal	Episomal
TCGA-C5-A2M2	Positive	High	A2	A	Episomal	Episomal
TCGA-IR-A3LI	Positive	Low	D2	D	Integrated	Episomal
TCGA-EK-A2PI	Positive	High	D3	D	Integrated	Integrated
TCGA-EX-A1H6	Positive	Low	D3	D	Integrated	Integrated
TCGA-FU-A3EO	Positive	High	D3	D	Episomal	Episomal
TCGA-JW-A69B	Positive	High	D3	D	Integrated	Episomal
TCGA-MU-A51Y	Positive	Low	D3	D	Integrated	Integrated
TCGA-MU-A5YI	Positive	Low	D3	D	Integrated	Integrated
TCGA-Q1-A6DV	Positive	High	D3	D	Integrated	Episomal
TCGA-C5-A0TN	Positive	Low	A1	A	Integrated	Integrated
TCGA-C5-A1BJ	Positive	High	A1	A	Integrated	Integrated
TCGA-C5-A1BK	Positive	Low	A1	A	Integrated	Integrated
TCGA-C5-A1BL	Positive	High	A1	A	Integrated	Integrated
TCGA-C5-A1M5	Negative	Negative	-	-	-	-
TCGA-C5-A1ME	Negative	Negative	-	-	-	-
TCGA-C5-A1MJ	Negative	Negative	-	-	-	-
TCGA-C5-A1MN	Positive	High	A1	A	Integrated	Integrated
TCGA-C5-A1MP	Positive	High	A1	A	Integrated	Integrated
TCGA-C5-A2LV	Positive	High	A1	A	Integrated	Integrated
TCGA-DR-A0ZM	Negative	Negative	-	-	-	-
TCGA-EA-A410	Negative	Negative	-	-	-	-
TCGA-EA-A4BA	Negative	Negative	-	-	-	-
TCGA-EA-A556	Negative	Negative	-	-	-	-
TCGA-EK-A2PM	Negative	Negative	-	-	-	-
TCGA-FU-A57G	Negative	Negative	-	-	-	-
TCGA-HM-A6W2	Negative	Negative	-	-	-	-
TCGA-IR-A3LA	Negative	Negative	-	-	-	-
TCGA-JW-A5VH	Negative	Negative	-	-	-	-

HPV16 lineage, but not sub-lineage, is consistent between RNA-Seq and WXS

Early genotyping results showed high consistency among RNA-Seq and WXS samples at the lineage level, with all sample pairs with matching positivity calls also possessing matching lineage calls. However, of these cases, only 7 matched at the sub-lineage level (1 match for D2 and 6 matches for D3). No exact matches were found at the sub-lineage level for samples of the A or B lineages. This could be due to the high degree of sequence similarity between members of the A lineage as opposed to those belonging to the D lineage, making the sub-lineage determination dependent on a higher degree of difference between references. In the case of WXS sample TCGA-IR-A3LA, our genotyping algorithm was not capable of distinguishing between sub-lineages A2, A3 and A4, indicating equal identity for all three sequences based on mapping statistics. The final lineage distribution of the 23 cases positive for HPV16 can be seen in **Figure 2D**. For our controls, SiHa and CaSki were both within A lineage, with SiHa identified as A1/A3 and CaSki as sub-lineage A2. Overall, our results are consistent with others reporting that sub-lineage identification is challenging, with difficulty in confidently assigning a single sub-lineage among highly similar members of the same lineage [Dawson *et al*., 2019].

An important consideration is that accurate sub-lineage identification can be negatively impacted by integration itself, due to truncation of viral genomic regions which may decrease the amount of genomic space available for alignment. Also, it is possible that some integration events may render the viral genome transcriptionally inactive, but still tumourigenic due to insertional mutagenesis (*e.g.*, disrupting a tumour suppressor). In this theoretical example, DNA-level high-throughput sequencing (e.g., WGS and WXS) would be require to resolve the integration site, but low sequencing depth and a low abundance of virus relative to host DNA could yield false-negatives for viral positivity, sub-lineage identification, and integration. This is a limitation with TCGA datasets. Viral sequence capture enrichment could address this concern, so long as probes are designed to target across the entire viral genome and flanking host regions.

The proportion of HPV16 integrated reads does not differ between lineages

Initial raw output indicated the presence of viral integration sites in all samples that had been found positive for HPV16 in earlier steps of the pipeline, suggesting a higher than anticipated false-positive rate. While it is possible these could represent a low proportion of truly integrated reads, we decided that by considering the sequencing depth of each sample we could obtain a more reliable metric in the form of percentage of integrated reads [**Figure 3A**]. Samples were then filtered based on this value and compared to findings from the landmark CESC study [TCGA Research Network, 2017]. Once again, results from RNA-Seq samples were the most consistent with previous reports, indicating presence of integrated HPV16 [**Figure 3A**]. In this case, the proportion of integrated reads was higher in RNA-Seq vs WXS samples. The HPV16 positive cervical carcinoma controls resulted in positive integration for SiHa, while CaSki was considered negative with the current methodology. Clearly, CaSki's tandemly repeated full genomes pose a problem for detection and represent a false-negative, highlighting the importance of complimentary techniques and a careful biological interpretation and analysis. Mean integrated reads (%) in HPV16 lineages A vs D were compared and indicate no difference at the RNA or DNA level **[Figure 3B]**, yet a high amount of variability among the HPV16 positive samples.

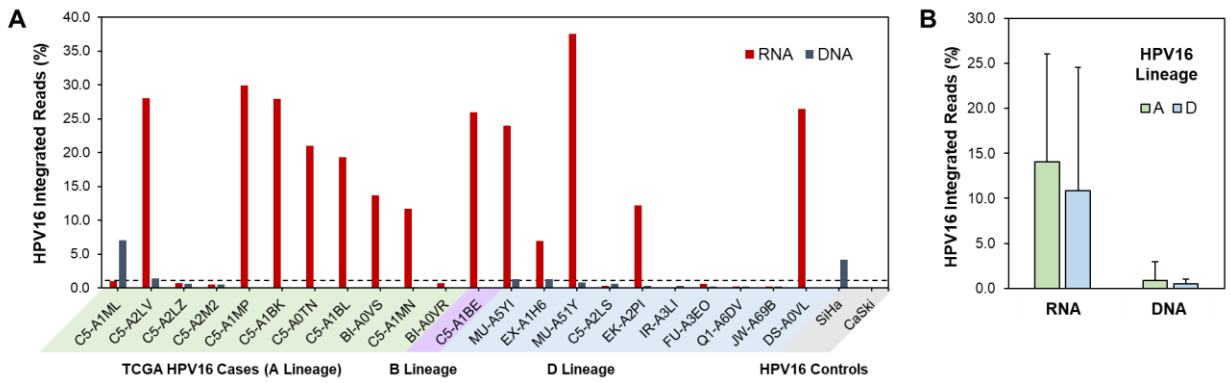


Figure 3. Integrated HPV16 reads in RNA-Seq and WXS samples. (A) For each of the HPV16 positive samples, the total number of integrated reads reported by HPVDetector were converted to a percentage of the total number of reads. Samples with 1.14% or more integrated reads (marked by the dashed line) were classified as positive for HPV16 integration. Samples below this threshold were classified as predominately containing the episomal form of HPV16. SiHa and CaSki controls were also included, as they are both known to have integrated HPV16. **(B)** Proportion of integrated reads for HPV16 lineages A and D. The value of percentage integrated reads for RNA-Seq and WXS samples, error bars represent standard deviation.

Present challenges and future directions

While we have addressed relevant literature, acquired access to TCGA sequencing data, designed and setup a working algorithm, and in a subset of samples carried out analyses for the HPV16 genotype, sub-lineages, and integration, continued efforts (and analysis of remaining TCGA cases) are required to complete our work. Do we reject or fail to reject our hypothesis that D2/D3 is more strongly associated with integration than A1, and do we have sufficient statistical power for this conclusion, or is sample size still too small? Besides the general hypothesis (based on counts/proportion data, which presently leads us to reject that D2/D3 sub-lineages are associated with integration), are there any differences in integration patterns that are unique? Even with TCGA datasets, a major limitation in answering these questions is the small number of

D2/D3 compared to A1 samples, and it may be necessary to extend these analyses beyond the well-curated TCGA database, but more broadly into other databases (*e.g*., the Sequence Read Archive, SRA). While additional samples could possibly be found and incorporated into our analyses, these data may lack sufficient metadata to be easily searchable and screened (*e.g*., experimental vs clinical data). Since the needs of a specific HPV-based analyses may not have been considered *a priori* when data were collected for those samples, there may be significant variability in data formats, read lengths, depths, and quality. Specifically, whether sequence capture or enrichment for HPV sequences was performed would greatly affect the probability of detecting HPV sequences [Holmes *et al*., 2016; Jackson *et al*., 2016]. Sequencing errors could further confound variant identification, requiring careful mitigation [Stewart *et al*., 2018]. Overall, while these challenges exist, potentially relevant datasets deserve analysis and could be useful contributions toward our research question.

Additional work is also required to determine the factors involved in a HPV16 sub-lineage-specific genomic instability and integration risk, including the role of defective DNA damage repair pathways [Seiwert *et al*., 2015; Ratnaparkhe *et al*., 2018]. It is worthwhile to explore relationships between E6 splice variants and genomic stability [Olmedo-Nieva *et al*., 2018]. Researchers continue to explore novel mechanisms surrounding HPV16 integration, such as “super-enhancer-like elements” [Warburton *et al*., 2018] and interruption of tumour suppressor genes [Zhao *et al*., 2016]. Future research could continue exploring these relationships, such as determining the role of non-coding genome elements in genomic instability (*e.g*., lncRNAs) [Munschauer *et al*., 2018], the similarities of viral-human integration with mobile elements and human fusion genes [Imielinski and Ladanyi, 2018], as well as evolutionary perspectives [Petrie *et al*., 2018]. Finally, it remains to be seen how clinically-relevant HPV16 sub-lineage genotyping could be, and whether this could lead to improved care and outcome for patients. One strategy could be to compute individual risk scores based on the co-factors affecting the risk of disease progression

[Bastarache *et al*., 2018]. New screening techniques could also be useful, such as urine sampling [Van Keer *et al*., 2018] followed by high-throughput sequencing and HPV genotyping, as well as assessing methylation signatures [Cook *et al*., 2018].

Full characterization of the genomic landscape surrounding detected HPV16 integration sites within human chromosomes is required to comprehensively address our research question *qualitatively*, including viral and host nucleotide positions, sequence overlap/microhomologies, functional annotations of viral and host features, such as structure and gene proximity (inside/outside, exon/intron), promoters, enhancers, transcription factor binding sites, repeat elements, and proximity to fragile sites. “Virtual ChIP-seq” [Karimzadeh and Hoffman, 2018] could be a useful tool to predict nearby transcription factor binding sites (*e.g.*, CTCF [Doolittle-Hall *et al*., 2015; Paris *et al*., 2015]). As described previously [Jackson *et al*., 2016], the region surrounding a viral integration site can be scanned for repeat elements, which may be prone to rearrangements, using the UCSC human genome browser RepeatMasker track. Another important aspect for qualitative assessment and interpretation will be to visualize and compare the A1 vs D2/D3 cases; example visualizations exist in the literature [Akagi *et al*., 2014; Tang *et al*., 2013; Holmes *et al*., 2016; Jackson *et al*., 2016; Warburton *et al*., 2018; Lagström *et al*., 2019].

CONCLUDING REMARKS

From a basic science perspective, big data analysis may aid in unraveling the mechanisms by which viruses can become integrated into their hosts, akin to mobile elements, and as a result lead to persistence of those sequences with consequences for the host genome. In the case of HPV16, a small number of changes in its DNA may be responsible for differential host genomic instability and integration propensity, as evidenced by a difference in the rate of integration between HPV16 D2 and D3 sub-lineages [Lou *et al*., 2020]. Additional bioinformatics analysis of this study’s and other *ex vivo* data, coupled with *in vitro* experimental work, is required to further

test the HPV integration hypothesis and the impact of other co-factors. Hypothesis-free exploration of these data will also be important to clarify the gaps in our existing knowledge and identify fascinating phenomena and molecular mechanisms to test in future experiments [Pipas, 2019]. From an applied science and clinical perspective, HPV16 sub-lineage genotyping can be used for preventative screening and routine monitoring, and next-generation sequencing specifically could be used on blood to detect integrated HPV sequences in circulating tumour DNA [Holmes *et al*., 2016] as a tumour-specific diagnostic biomarker and for monitoring residual disease and recurrence after treatments.

METHODS

Data access and acquisition

Access to controlled sample data from TCGA, using the database of Genotypes and Phenotypes (dbGaP), was acquired via application to the electronic Research Administration (eRA) commons. We registered our host institution, Lakehead University, as a new organization and received a Data Universal Numbering System (DUNS) identifier for access. Our proposal title was “Characterization of human papillomavirus type 16 integration sites in cervical and head and neck tumour biopsies”, with approval allowing access to TCGA Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (TCGA-CESC) as well as Head and Neck Squamous Cell Carcinoma (TCGA-HNSC) files for analysis. After receiving access to the data (a multi-step administrative process, outlined in **Figure 4A**) TCGA case files were transferred via their universally unique identifier (UUID) using the command-line implementation of the Genomics Data Commons Transfer Tool v1.4.0 and analyzed on the Compute Canada Cedar cluster, which uses the Secure Shell (SSH) encryption protocol for secure data transmission.

High-throughput sequencing data from HPV16-positive cervical carcinoma-derived cell lines SiHa [Friedl *et al*., 1970] and CaSki [Pattillo *et al*., 1977] were used as controls. For SiHa,

which contains low-copy number integrated HPV16, we used the same sample used in the HPVDetector paper [Chandrani *et al*., 2015], a whole-genome sequence from the DDBJ (DNA Data Bank of Japan) Sequence Read Archive (DRA) run SRR1609142, from study SRP048769. For CaSki, which contains a high-copy number of tandemly repeated HPV16 genome integrations, we downloaded RNA-Seq data from the Sequence Read Archive (SRA) run SRR9924750, from project PRJNA484129 [Sen *et al*., 2020].

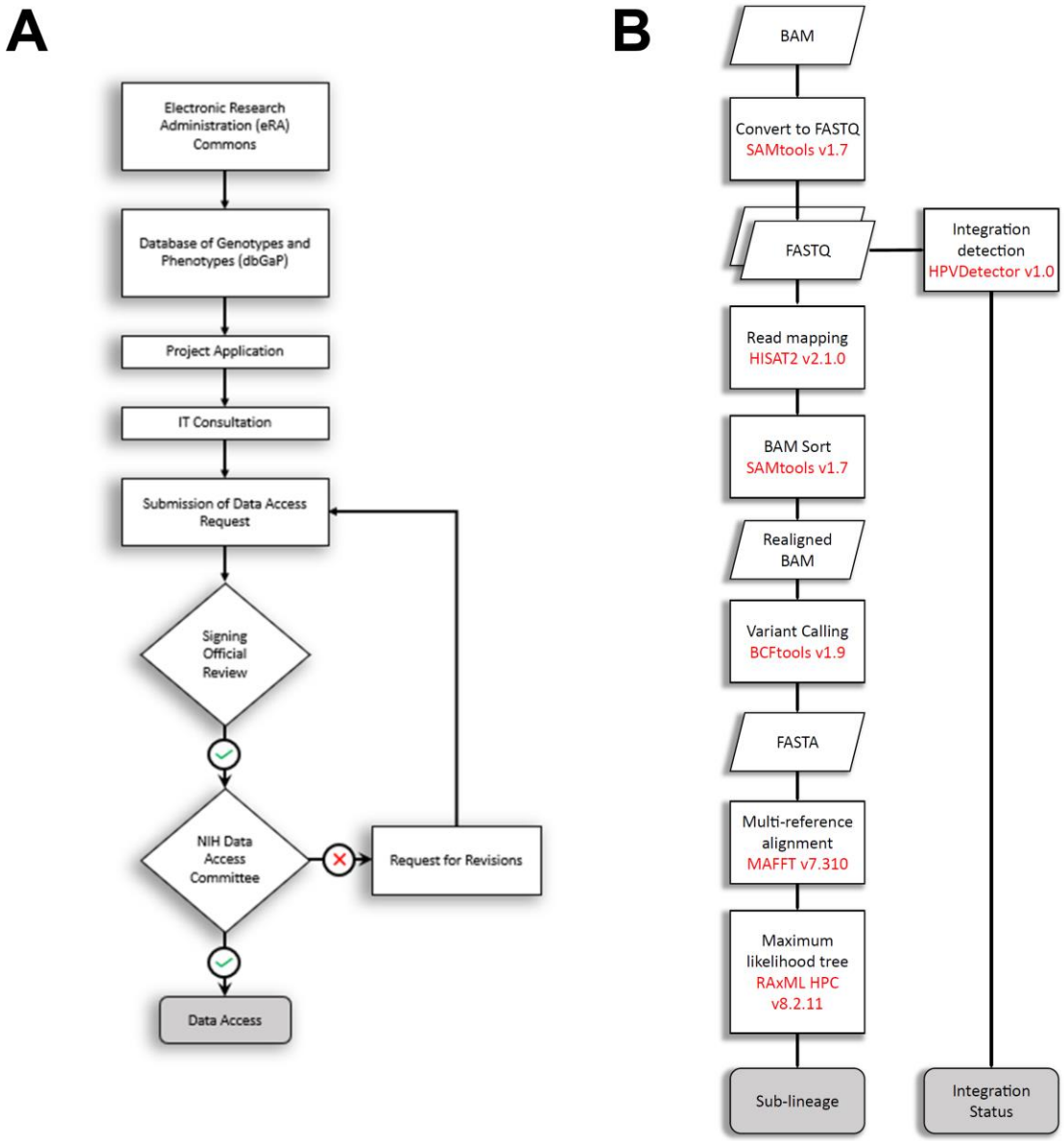


Figure 4. Access and analysis of TCGA data. (A) Access to controlled sample genomic and transcriptomic data from TCGA required a multi-step authorization process before being able to access relevant datasets. This administrative process, which requires annual renewal and project revision, takes approximately three months initially. **(B)** Our analytical workflow involved the following computational steps: TCGA high-throughput sequence read data (WXS or RNA-Seq) were first accessed and converted from pre-aligned BAM/SAM format to FASTQ read format. Data were then subjected to HPV16 sub-lineage determination and integration analysis.

HPV16 positivity and sub-lineage genotyping analysis

While HPV16-positive cervical cancer (TCGA-CESC) samples had associated sub-lineages already identified (albeit, with 10 of the 16 current sub-lineage reference genomes [TCGA Research Network, 2017]), sub-lineage genotyping was not immediately available for the head and neck (TCGA-HNSC) cases [TCGA Research Network, 2015]. It was thus necessary to develop a pipeline for ascertaining the sub-lineages of unknown identified samples (*i.e.*, those that were identified as HPV16 positive by Cantalupo *et al*., [2018] as well) in addition to confirming their status as HPV16 positive through an analysis of our own. Our pipeline follows an analytical workflow starting with the pre-aligned BAM files from TCGA [**Figure 4B**]. First, BAM files were converted into FASTQ files using SAMtools v1.7 [Li *et al*., 2009]. The reads were then mapped to the reference genomes of each HPV16 sub-lineage [**Table 3**] using HISAT2 v2.1.0 [Kim *et al*., 2015]. From the sixteen separate alignments performed, the one with the highest number of mapped reads was used for quantifying the presence of HPV16 by calculating the number of viral reads per million sequencing reads (RPM = 10^6 x number of mapped reads/total reads, which accounts for sequence depth variation between sample libraries). RPM ranges for positivity analysis were based on a prior study which sought reasonable cutoffs for HPV detection in cancer tissues [Montgomery *et al*., 2016]. If RPM < 0.99, the sample was marked as negative across all categories and the analysis was terminated. Samples with RPM between 0.99 and 99 were marked as having “low” positivity for HPV16, while samples of RPM > 99 were marked as having “high” positivity. Given their large range, all RPM values were subjected to an inverse hyperbolic sine (IHS) transformation [Burbidge *et al*, 1988] following $IHS(x) = \log_{10}(x + (x^2 + 1)^{1/2})$, an alternative to traditional logarithmic transformation that allows values lower than 1 to have a plottable log value and which has seen recent use when analyzing biological data [Haines *et al*., 2020]. SAMtools v1.7 was then used to sort and prepare the files corresponding to these samples for conversion to VCF format. BCFtools v1.9 [Li, 2011] was used for variant calling and generating

the consensus sequence. The new consensus sequences along with all reference sequences were aligned against each other using MAFFT v7.310 [Katoh *et al*., 2009] with default parameters. Maximum likelihood trees were inferred using RAxML HPC v8.2.11 [Stamatakis, 2014], under the general time-reversible nucleotide model with gamma-distributed rate heterogeneity and invariant sites (GTRGAMMAI). Sub-lineages were determined from Newick tree files using ETE Toolkit v3.0.0b34 [Huerta-Cepas *et al*., 2010] based on topological distance. Each sample was incorporated into the reference tree [**Figure 5**] and assigned a sub-lineage corresponding to the nearest neighbour.

Table 3. HPV16 sub-lineage reference genomes used for genotyping. Derived from reference data summarized in Burk *et al*. [2013], Chen *et al*. [2018], and Mirabello *et al*. [2018].

HPV16 sub-lineage	GenBank #	Other ID	Original nomenclature
A1	K02718	NC_001526.4 (latest)	European Prototype (EP)
A2	AF536179	w0122	European (E)
A3	HQ644236	AS411	European (E)
A4	AF534061	w0724	Asian (As)
B1	AF536180	w0236	African-1a (Af-1a)
B2	KU053915	IARC1100085NI	African-1 (Af-1)
B3	HQ644298	Z109	African-1b (Af-1b)
B4	KU053914	IARC907912AL	African-1 (Af-1)
C1	AF472509	R460	African-2 (Af-2)
C2	HQ644244	IARC240211GU	African-2 (Af-2)
C3	KU053920	IARC040105BE	African-2 (Af-2)
C4	KU053925	IARC304612MO	African-2 (Af-2)
D1	HQ644257	Qv00512	North-American-1 (NA1)
D2	AY686579	Qv15321	Asian-American (AA) 2
D3	AF402678	Qv00995	Asian-American (AA) 1
D4	KU053931	IARC903812AL	North-American-2 (NA2)

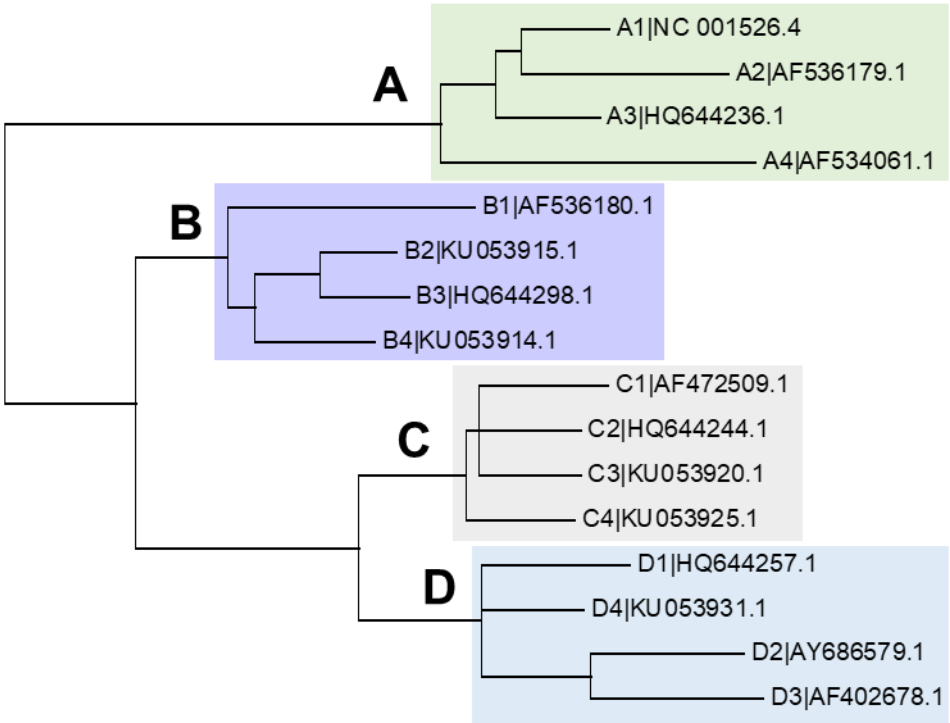


Figure 5. HPV16 sub-lineage reference tree for phylogenetic analyses. HPV16 phylogenetic maximum likelihood tree used as a reference for sub-lineage determination. Branches correspond to four lineages (A, B, C, and D), each containing four sub-lineage genomes (1 to 4). The tree was built with FigTree v1.4.4 using standard settings.

Viral-human integration detection and characterization

Several open-source software tools for integration analysis are available for WGS, WXS, and whole-transcriptome sequencing (RNA-Seq): *e.g.*, VirusSeq [Chen *et al*., 2013], ViralFusionSeq [Li *et al*., 2013], VirusFinder [Wang *et al*., 2013], HPVDetector v1.0 [Chandrani *et al*., 2015], ViFi [Nguyen *et al*., 2018] and VirTect [Xia *et al*., 2019]. The latter tool name, VirTect, was also used by another group for viral genome sequencing and host genes related to carcinogenesis using RNA-Seq [Khan *et al*., 2018]. HPVDetector v1.0 was chosen since it was specifically designed for detecting and annotating HPV integration events, as well its compatibility with the Compute Canada Cedar cluster environment. VirTect was also considered but is currently incompatible with available modules in Cedar, requiring an older version of Python (Python 2.7 vs 3.0). Integration analysis was performed using command-line execution of HPVDetector v1.0 [**Figure 4B**] set to “Integration” mode for all samples and “Transcriptome” or “Exome” mode depending if the sample contained WXS or RNA-Seq data, respectively. Raw output was obtained in the form of the number of integrated reads. This value was then converted to a percentage of the total number of viral reads to account for sequencing depth variation amongst selected samples and used as a metric to quantify viral-host integration. Samples with values at or above 1.14% were classified as positive for HPV16 integration, a threshold obtained through comparison to the landmark CESC study cases to maximize concordance. The remaining samples were considered as positive for the episomal form of HPV16.

DECLARATIONS

Acknowledgements

The results reported here are based upon data generated by The Cancer Genome Atlas Research Network (cancergenome.nih.gov) and we are grateful to the participants and Network members for this resource. This research was enabled in part by computational support provided

by Compute Ontario (www.computeontario.ca) and Compute Canada (www.computecanada.ca). Thank you to Darryl Willick and Dr. Wely Floriano for their assistance in helping set up and maintain required bioinformatics tools (such as the Galaxy platform) hosted at the Lakehead University High Performance Computing Centre (LUHPCC) as well as Dr. Zigui Chen who shared a HPV sub-lineage detection pipeline. We are also thankful for Bruce Bogacki's technical support at the Thunder Bay Regional Health Research Institute (TBRHRI). Additional thanks to Kathleen Roulston, Vanessa Masters, Anirudh Shahi, Dallas Nygard, and Christopher Gibb for assisting with preliminary literature review, initial data access, and early analyses. Preliminary work related to this study was presented as a poster in Heidelberg, Germany (11-14 Oct 2017) at The Mobile Genome: Genetic and Physiological Impacts of Transposable Elements conference and talks in Thunder Bay, Canada at the Health & Information Technology (HIT) Research Group Workshop (12 Oct 2018) and the Health & Information Technology & Biotechnology & Allied Sciences (HITBASS) Symposium (27 Oct 2019). This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) grants to IZ (#RGPIN-2015-03855) and RJ (CGS-D#454402-2014). The funding body had no role in study design, data collection, data analysis and interpretation, or preparation of the manuscript. We declare no conflicts of interest.

Authors' contributions

The study was initially conceived of and designed by IZ and RJ. Data access was managed by RJ, IZ, and AOV. Methods were developed, optimized, and analyses performed by AOV with input from RJ and IZ. Biological interpretation, literature searches and manuscript writing were led by RJ and IZ, with input from AOV, while all authors were involved in providing critical revision and feedback.

REFERENCES

Akagi K, Li J, Broutian TR, Padilla-Nash H, Xiao W, Jiang B, *et al*. Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. *Genome Res*. **2014**; 24:185-99. DOI: 10.1101/gr.164806.113

Bastarache L, Hughey JJ, Hebbring S, Marlo J, Zhao W, Ho WT, *et al*. Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. *Science* **2018**; 359:1233-9. DOI: 10.1126/science.aal4043

Berumen J, Ordoñez RM, Lazcano E, Salmeron J, Galvan SC, Estrada RA, *et al*. Asian American variant of human papillomavirus 16 and risk for cervical cancer: a case–control study. *J Natl Cancer Inst*. **2001**; 93:1325-30. DOI: 10.1093/jnci/93.17.1325

Burbidge JB, Magee L, Robb AL. Alternative transformations to handle extreme values of the dependent variable. *J Am Stat Assoc*. **1988**; 83:123-7. DOI: 10.1080/01621459.1988.10478575

Burk RD, Harari A, Chen Z. Human papillomavirus genome variants. *Virology*. **2013**; 445:232-43. DOI: 10.1016/j.virol.2013.07.018

Cantalupo PG, Katz JP, Pipas JM. HeLa nucleic acid contamination in The Cancer Genome Atlas leads to the misidentification of HPV18. *J Virol*. **2015**; 89:4051-7. DOI: 10.1128/JVI.03365-14

Cantalupo PG, Katz JP, Pipas JM. Viral sequences in human cancer. *Virology*. **2018**; 513:208-16. DOI: 10.1016/j.virol.2017.10.017

Castellsagué X, Alemany L, Quer M, Halec G, Quirós B, Tous S, *et al*. HPV involvement in head and neck cancers: comprehensive assessment of biomarkers in 3680 patients. *J Natl Cancer Inst*. **2016**; 108:djv403. DOI: 10.1093/jnci/djv403

Chandrani P, Kulkarni V, Iyer P, Upadhyay P, Chaubal R, Das P, *et al*. NGS-based approach to determine the presence of HPV and their sites of integration in human cancer genome. *Br J Cancer* **2015**; 112:1958-65. DOI: 10.1038/bjc.2015.121

Chen Y, Yao H, Thompson EJ, Tannir NM, Weinstein JN, Su X. VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. *Bioinformatics.* **2013**; 29:266-267. DOI: 10.1093/bioinformatics/bts665

Chen Z, DeSalle R, Schiffman M, Herrero R, Wood CE, Ruiz JC, *et al*. Niche adaptation and viral transmission of human papillomaviruses from archaic hominins to modern humans. *PLoS Pathog*. **2018**; 14:e1007352. DOI: 10.1371/journal.ppat.1007352

Clifford GM, Tenet V, Georges D, Alemany L, Pavón MA, Chen Z. Human Papillomavirus 16 sub-lineage dispersal and cervical cancer risk worldwide: whole viral genome sequences from 7116 HPV16-positive women. *Papillomavirus Res*. **2019**; 7:67-74. DOI: 10.1016/j.pvr.2019.02.001

Cook DA, Krajden M, Brentnall AR, Gondara L, Chan T, Law JH, *et al*. Evaluation of a validated methylation triage signature for human papillomavirus positive women in the HPV FOCAL cervical cancer screening trial. *Int J Cancer* **2018**: 9 Nov 2018. DOI: 10.1002/ijc.31976

Cuninghame S, Jackson R, Lees SJ, Zehbe I. Two common variants of human papillomavirus type 16 E6 differentially deregulate sugar metabolism and hypoxia signalling in permissive human keratinocytes. *J Gen Virol*. **2017**; 98:2310-9. DOI: 10.1099/jgv.0.000905

Dawson ET, Wagner S, Roberson D, Yeager M, Boland J, Garrison E, *et al*. Viral coinfection analysis using a MinHash toolkit. *BMC Bioinformatics* **2019**; 20:389. DOI: 10.1186/s12859-019-2918-y

Doolittle-Hall J, Cunningham Glasspoole D, Seaman W, Webster-Cyriaque J. Meta-analysis of DNA tumor-viral integration site selection indicates a role for repeats, gene expression and epigenetics. *Cancers* **2015**; 7:2217-2235. DOI: 10.3390/cancers7040887

Friedl F, Kimura I, Osato T, Ito Y. Studies on a new human cell line (SiHa) derived from carcinoma of uterus. I. Its establishment and morphology. *Proc Soc Exp Biol Med.* **1970**; 135:543-5. DOI: 10.3181/00379727-135-35091a

Ghittoni R, Accardi R, Chiocca S, Tommasino M. Role of human papillomaviruses in carcinogenesis. *Ecancermedicalscience*. **2015**; 9:526. DOI: 10.3332/ecancer.2015.526

Graham SV, Faizo AA. Control of human papillomavirus gene expression by alternative splicing. *Virus Res*. **2017**; 231:83-95. DOI: 10.1016/j.virusres.2016.11.016

Haines SR, Siegel JA, Dannemiller KC. Modeling microbial growth in carpet dust exposed to diurnal variations in relative humidity using the “Time-of-Wetness” framework. *Indoor Air*. **2020**. DOI: 10.1111/ina.12686

Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. **2011**; 144:646-74. DOI: 10.1016/j.cell.2011.02.013

Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell*. **2000**; 100:57-70. DOI: 10.1016/S0092-8674(00)81683-9

Holmes A, Lameiras S, Jeannot E, Marie Y, Castera L, Sastre-Garau X, *et al*. Mechanistic signatures of HPV insertions in cervical carcinomas. *Genome Med*. **2016**; 1:16004. DOI: 10.1038/npjgenmed.2016.4

Hoppe-Seyler K, Bossler F, Braun JA, Herrmann AL, Hoppe-Seyler F. The HPV E6/E7 oncogenes: key factors for viral carcinogenesis and therapeutic targets. *Trends Microbiol*. **2018**; 26:158-68. DOI: 10.1016/j.tim.2017.07.007

Huerta-Cepas J, Dopazo J, Gabaldón T. ETE: a python Environment for Tree Exploration. BMC Bioinformatics **2010**; 11:24. DOI: 10.1186/1471-2105-11-24

Imielinski M, Ladanyi M. Fusion oncogenes—genetic musical chairs. *Science* **2018**; 361:848-9. DOI: 10.1126/science.aau8231

Jackson R, Rosa BA, Lameiras S, Cuninghame S, Bernard J, Floriano WB, *et al*. Functional variants of human papillomavirus type 16 demonstrate host genome integration and transcriptional alterations corresponding to their unique cancer epidemiology. *BMC Genomics*. **2016**; 17:851. DOI: 10.1186/s12864-016-3203-3

Jackson R, Togtema M, Lambert PF, Zehbe I. Tumourigenesis driven by the human papillomavirus type 16 Asian-American E6 variant in a three-dimensional keratinocyte model. *PLoS One* **2014**; 9:e101540. DOI: 10.1371/journal.pone.0101540

Karimzadeh M, Hoffman MM. Virtual ChIP-seq: Predicting transcription factor binding by learning from the transcriptome. *bioRxiv* **2018**:168419. DOI: 10.1101/168419

Katoh K, Asimenos G, Toh H. Multiple alignment of DNA sequences with MAFFT. *Methods Mol Biol* **2009**; 537:39-64. DOI: 10.1007/978-1-59745-251-9_3

Khan A, Liu Q, Chen X, Stucky A, Sedghizadeh PP, Adelpour D, *et al*. Detection of human papillomavirus in cases of head and neck squamous cell carcinoma by RNA-seq and VirTect. *Mol Oncol***. 2019**; 13:829-39. DOI: 10.1002/1878-0261.12435

Khoury JD, Tannir NM, Williams MD, Chen Y, Yao H, Zhang J, *et al*. Landscape of DNA virus associations across human malignant cancers: analysis of 3,775 cases using RNA-Seq. *J Virol*. **2013**; 87:8916-26. DOI: 10.1128/JVI.00340-13

Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **2015**; 12:357-60. DOI: 10.1038/nmeth.3317

Lagström S, Umu SU, Lepistö M, Ellonen P, Meisal R, Christiansen IK, *et al*. TaME-seq: an efficient sequencing approach for characterisation of HPV genomic variability and chromosomal integration. *Sci Rep*. **2019**; 9:524. DOI: 10.1038/s41598-018-36669-6

LeConte BA, Szaniszlo P, Fennewald SM, Lou DI, Qiu S, Chen NW, *et al*. Differences in the viral genome between HPV-positive cervical and oropharyngeal cancer. *PLoS One* **2018**; 13:e0203403. DOI: 10.1371/journal.pone.0203403

Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **2011**; 27:2987-2993. DOI: 10.1093/bioinformatics/btr509

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics* **2009**; 25:2078-9. DOI: 10.1093/bioinformatics/btp352

Li JW, Wan R, Yu CS, Co NN, Wong N, Chan TF. ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution. *Bioinformatics* **2013**; 29:649-51. DOI: 10.1093/bioinformatics/btt011

Lou H, Boland JF, Torres-Gonzalez E, Albanez A, Zhou W, Steinberg M, *et al*. Human papillomavirus 16 positive cervical cancer in Guatemala: The D2 and D3 sublineages differ in integration rate and age of diagnosis. *bioRxiv* 2020. DOI: 10.1101/2020.01.07.897546

Mesri EA, Feitelson MA, Münger K. Human viral oncogenesis: a cancer hallmarks analysis. *Cell Host Microbe*. **2014**; 15:266-82. DOI: 10.1016/j.chom.2014.02.011

Mirabello L, Clarke MA, Nelson CW, Dean M, Wentzensen N, Yeager M, *et al*. The intersection of HPV epidemiology, genomics and mechanistic studies of HPV-mediated carcinogenesis. *Viruses* **2018**; 10:80. DOI: 10.3390/v10020080

Montgomery ND, Parker JS, Eberhard DA, Patel NM, Weck KE, Sharpless NE, *et al*. Identification of Human Papillomavirus Infection in Cancer Tissue by Targeted Next Generation Sequencing. *Appl Immunohistochem Mol Morphol*. **2016**; 24:490-5. DOI: 10.1097/PAI.0000000000000215

Munschauer M, Nguyen CT, Sirokman K, Hartigan CR, Hogstrom L, Engreitz JM, *et al*. The NORAD lncRNA assembles a topoisomerase complex critical for genome stability. *Nature* **2018**; 561:132-136. DOI: 10.1038/s41586-018-0453-z

Ndiaye C, Mena M, Alemany L, Arbyn M, Castellsagué X, Laporte L, *et al*. HPV DNA, E6/E7 mRNA, and p16^{INK4a} detection in head and neck cancers: a systematic review and meta-analysis. *Lancet Oncol*. **2014**; 15:1319-1331. DOI: 10.1016/S1470-2045(14)70471-1

Nguyen N-PD, Deshpande V, Luebeck J, Mischel PS, Bafna V. ViFi: accurate detection of viral integration and mRNA fusion reveals indiscriminate and unregulated transcription in proximal genomic regions in cervical cancer. *Nucleic Acids Res*. **2018**; 46:3309-25. DOI: 10.1093/nar/gky180

Niccoli S, Abraham S, Richard C, Zehbe I. The Asian-American E6 variant protein of human papillomavirus 16 alone is sufficient to promote immortalization, transformation, and migration of primary human foreskin keratinocytes. *J Virol*. **2012**; 86:12384-96. DOI: 10.1128/JVI.01512-12

Nulton TJ, Olex AL, Dozmorov M, Morgan IM, Windle B. Analysis of The Cancer Genome Atlas sequencing data reveals novel properties of the human papillomavirus 16 genome in head and neck squamous cell carcinoma. *Oncotarget* **2017**; 8:17684-99. DOI: 10.18632/oncotarget.15179

Olmedo-Nieva L, Muñoz-Bello J, Contreras-Paredes A, Lizano M. The role of E6 spliced isoforms (E6*) in human papillomavirus-induced carcinogenesis. *Viruses* **2018**; 10:45. DOI: 10.3390/v10010045

Parfenov M, Pedamallu CS, Gehlenborg N, Freeman SS, Danilova L, Bristow CA, *et al*. Characterization of HPV and host genome interactions in primary head and neck cancers. *Proc Natl Acad Sci USA*. **2014**; 111:15544-15549. DOI: 10.1073/pnas.1416074111

Paris C, Pentland I, Groves I, Roberts DC, Powis SJ, Coleman N, *et al*. CCCTC-binding factor recruitment to the early region of the human papillomavirus 18 genome regulates viral oncogene expression. *J Virol*. **2015**; 89:4770-4785. DOI: 10.1128/JVI.00097-15

Pattillo RA, Hussa RO, Story MT, Ruckert AC, Shalaby MR, Mattingly RF. Tumor antigen and human chorionic gonadotropin in CaSki cells: a new epidermoid cervical cancer cell line. *Science* **1977**; 196:1456-8. DOI: 10.1126/science.867042

Petrie KL, Palmer ND, Johnson DT, Medina SJ, Yan SJ, Li V, *et al*. Destabilizing mutations encode nongenetic variation that drives evolutionary innovation. *Science* **2018**; 359:1542-1545. DOI: 10.1126/science.aar1954

Pipas JM. DNA tumour viruses and their contributions to molecular biology. *J Virol.* **2019**; JVI.01524-18. DOI: 10.1128/JVI.01524-18

Ratnaparkhe M, Wong J, Wei PC, Hlevnjak M, Kolb T, Haag D, *et al*. Defective DNA damage repair leads to frequent catastrophic genomic events in murine and human tumors. *Nat Commun*. **2018**; 9:4760. DOI: 10.1038/s41467-018-06925-4

Richard C, Lanner C, Naryzhny S, Sherman L, Lee H, Lambert PF, *et al*. The immortalizing and transforming ability of two common human papillomavirus 16 E6 variants with different prevalences in cervical cancer. *Oncogene* **2010**; 29:3435-3445. DOI: 10.1038/onc.2010.93

Seedorf K, Krämmer G, Dürst M, Suhai S, Röwekamp WG. Human papillomavirus type 16 DNA sequence. *Virology*. **1985**; 145:181-185. DOI: 10.1016/0042-6822(85)90214-4

Seiwert TY, Zuo Z, Keck MK, Khattri A, Pedamallu CS, Stricker T, *et al*. Integrative and comparative genomic analysis of HPV-positive and HPV-negative head and neck squamous cell carcinomas. *Clin Cancer Res*. **2015**; 21:632-641. DOI: 10.1158/1078-0432.CCR-13-3310

Sen P, Ghosal S, Hazra R, Arega S, Mohanty R, Kulkarni KK, *et al*. Transcriptomic analyses of gene expression by CRISPR knockout of miR-214 in cervical cancer cells. *Genomics* **2020**; 112:1490-9. DOI: 10.1016/j.ygeno.2019.08.020

Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **2014**; 30:1312-1313. DOI: 10.1093/bioinformatics/btu033

Stewart C, Leshchiner I, Hess J, Getz G. Comment on “DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification”. *Science* **2018**; 361:eaas9824. DOI: 10.1126/science.aas9824

Tang KW, Alaei-Mahabadi B, Samuelsson T, Lindh M, Larsson E. The landscape of viral expression and host gene fusion and adaptation in human cancer. *Nat Commun*. **2013**; 4:2513. DOI: 10.1038/ncomms3513

TCGA Research Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **2015**; 517:576-582. DOI: 10.1038/nature14129

TCGA Research Network. Integrated genomic and molecular characterization of cervical cancer. *Nature* **2017**; 543:378-384. DOI: 10.1038/nature21386

Togtema M, Hussack G, Dayer G, Teghtmeyer MR, Raphael S, Tanha J, *et al*. Single-domain antibodies represent novel alternatives to monoclonal antibodies as targeting agents against the human papillomavirus 16 E6 protein. *Int J Mol Sci*. **2019**; 20:2088. DOI: 10.3390/ijms20092088

Togtema M, Jackson R, Grochowski J, Villa PL, Mellerup M, Chattopadhyaya J, *et al*. Synthetic siRNA targeting human papillomavirus 16 E6: a perspective on *in vitro* nanotherapeutic approaches. *Nanomedicine* **2018**; 13:455-474. DOI: 10.2217/nnm-2017-0242

Togtema M, Jackson R, Richard C, Niccoli S, Zehbe I. The human papillomavirus 16 European-T350G E6 variant can immortalize but not transform keratinocytes in the absence of E7. *Virology*. **2015**; 485:274-282. DOI: 10.1016/j.virol.2015.07.025

Van Doorslaer K, Dillner J. The launch of an international animal papillomavirus reference center. *Viruses* **2019**; 11:55. DOI: 10.3390/v11010055

Van Doorslaer K, Li Z, Xirasagar S, Maes P, Kaminsky D, Liou D, *et al*. The Papillomavirus Episteme: a major update to the papillomavirus sequence database. *Nucleic Acids Res*. **2017a**; 45:D499-506. DOI: 10.1093/nar/gkw879

Van Doorslaer K, Ruoppolo V, Schmidt A, Lescroël A, Jongsomjit D, Elrod M, *et al*. Unique genome organization of non-mammalian papillomaviruses provides insights into the evolution of viral early proteins. *Virus Evol*. **2017b**; 3:vex027. DOI: 10.1093/ve/vex027

Van Doorslaer K, Tan Q, Xirasagar S, Bandaru S, Gopalan V, Mohamoud Y, *et al*. The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. *Nucleic Acids Res*. **2013**; 41:D571-578. DOI: 10.1093/nar/gks984

Van Keer S, Tjalma WA, Pattyn J, Biesmans S, Pieters Z, Van Ostade X, *et al*. Human papillomavirus genotype and viral load agreement between paired first-void urine and clinician-collected cervical samples. *Eur J Clin Microbiol Infect Dis*. **2018**; 37:859-869. DOI: 10.1007/s10096-017-3179-1

Wang Q, Jia P, Zhao Z. VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. *PLoS One*. **2013**; 8:e64465. DOI: 10.1186/s12920-018-0461-8

Warburton A, Redmond CJ, Dooley KE, Fu H, Gillison ML, Akagi K, *et al*. HPV integration hijacks and multimerizes a cellular enhancer to generate a viral-cellular super-enhancer that drives high viral oncogene expression. *PLoS Genet*. **2018**; 14:e1007179. DOI: 10.1371/journal.pgen.1007179

Winder DM, Pett MR, Foster N, Shivji MK, Herdman MT, Stanley MA, *et al*. An increase in DNA double-strand breaks, induced by Ku70 depletion, is associated with human papillomavirus 16 episome loss and de novo viral integration events. *J Pathol*. **2007**; 213:27-34. DOI: 10.1002/path.2206

Xi LF, Koutsky LA, Galloway DA, Kiviat NB, Kuypers J, Hughes JP, *et al*. Genomic variation of human papillomavirus type 16 and risk for high grade cervical intraepithelial neoplasia. *J Natl Cancer Inst*. **1997**; 89:796-802. DOI: 10.1093/jnci/89.11.796

Xia Y, Liu Y, Deng M, Xi R. Detecting virus integration sites based on multiple related sequencing data by VirTect. *BMC Med Genomics* **2019**; 12:19. DOI: 10.1186/s12920-018-0461-8

Xu B, Chotewutmontri S, Wolf S, Klos U, Schmitz M, Dürst M, *et al*. Multiplex identification of human papillomavirus 16 DNA integration sites in cervical carcinomas. *PLoS One* **2013**; 8:e66693. DOI: 10.1371/journal.pone.0066693

Zapatka M, Borozan I, Brewer DS, Iskar M, Grundhoff A, Alawi M, *et al*. The landscape of viral associations in human cancers. *bioRxiv* **2018**; 465757. DOI: 10.1101/465757

Zhao JW, Fang F, Guo Y, Zhu TL, Yu YY, Kong FF, *et al*. HPV16 integration probably contributes to cervical oncogenesis through interrupting tumor suppressor genes and inducing chromosome instability. *J Exp Clin Cancer Res*. **2016**; 35:180. DOI: 10.1186/s13046-016-0454-4