Investigating Molecular Drivers of Human Papillomavirus (HPV)-Related Oropharyngeal Cancer

by

Lisa M. Pinatti

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Cancer Biology)
in the University of Michigan
2021

Doctoral Committee:

Associate Professor David B. Lombard, Chair
Assistant Professor Chad J. Brenner
Professor Emeritus Thomas E. Carey
Professor Pierre A. Coulombe

Lisa M. Pinatti

lpinatti@umich.edu

ORCID iD: 0000-0002-4864-5698

Dedication

To the generous patients who donated their tissue for our studies.

Acknowledgements

I would first like to acknowledge my mentor Dr. Thomas Carey for being the optimistic voice driving me forward through this process. I admire your endless enthusiasm for your work and am eternally grateful for all you have invested in my training. Thank you to my dissertation committee chair Dr. David Lombard and other committee members Dr. Chad Brenner and Dr. Pierre Coulombe for providing their expertise and support through this process.

Thank you to the entire Head and Neck Oncology team for their contributions to my project and for sharing access to samples, equipment and lab space, including Dr. Chad Brenner, Dr. Heather Walline, Dr. Steven Chinn and their respective teams. I would especially like to thank Dr. Heather Walline for her mentorship during the most challenging times I faced as a graduate student. My work would not have been possible without Apurva Bhangale for her work on bioinformatic analysis and Collin Brummel on patient recruitment and data collection. Special thanks to Christine Goudsmit for being a great desk-mate in the "penthouse of science". Thanks to the undergraduate researchers who have helped me, including Nadine Jawad, Lila Peters and especially Hana Sinha for being my second pair of hands. This work would not have been possible without the efforts of the entire Head and Neck Surgery team at Michigan Medicine, including clinical research coordinators, nurses, and surgeons.

I want to acknowledge my collaborators near and far for their contributions to my work, including Dr. Christian Huebbers and Dr. Jens Peter Klussman from the University of Cologne, Dr. Ernst-Jan Speel from the University of Maastricht, Dr. George Wilson and Dr. Jan Akervall from

Beaumont Hospital, and Dr. Ryan Mills from the University of Michigan. Dr. Yifan Wang and Wenjin Gu from the Mills lab were instrumental in developing the bioinformatic analysis pipeline presented here. Additional thanks to the University of Michigan Advanced Genomics Core for their contributions to my work.

Thank you to the Cancer Biology Program for giving me this opportunity and for supporting me along the way, particularly Dr. Beth Lawlor, Dr. David Lombard, Zarinah Aquil, and Dawn Storball. I have enjoyed being a member of this graduate program and appreciated the support of the other students throughout my time here. Additional thanks to Margaret Conlon and the Kresge Hearing Research Institute (KHRI) administrative team for their help maneuvering through all the red tape of academic research. I am grateful to have been supported by the following funding sources: NIH-NCI R01 CA194536, Cancer Biology Training Grant NIH-NCI T32-CA009676, the Rackham Bernard L. Maas Fellowship and supplementary funds from the University of Michigan Department of Otolaryngology/Head and Neck Surgery.

Finally, thank you to my friends and family for their never-ending encouragement throughout this process. I owe everything I have and ever will have to my parents – thank you for always being my biggest supporters. To my brother Tommy and sister-in-law Bria, my sister Sarah and soon-to-be brother-in-law Kris, my extended family, and my lifelong friends Mary O'Neil and Mackenzie Murphy, thanks for always listening to my graduate school woes and welcoming me home with open arms. Thanks to Armaan Dandavati for always being there to cheer me on during the difficult times and to celebrate the good ones – I could not have done this without you.

Table of Contents

List of Tables

Table

List of Figures

Figure

List of Appendices

Appendix

List of Abbreviations

AJCC	American joint committee on cancer
APOT	amplification of papillomavirus transcripts
BOT	base of tongue
bp	base pair
CIN	cervical intraepithelial neoplasia
CIS	carcinoma in situ
CNA	copy number alteration
CRT	chemoradiotherapy
DIPS	detection of integrated papillomavirus sequences
DFS	disease-free survival
DOD	died of disease
DSS	disease specific survival
DM	distant metastasis
FFPE	formalin fixed paraffin embedded
HNSCC	head and neck squamous cell carcinoma
HPV	human papillomavirus
hrHPV	high-risk human papillomavirus
kb	kilobase
lncRNA	long non-coding RNA

LRF	locoregional failure
lrHPV	low-risk human papillomavirus
miRNA	microRNA
NED	no evidence of disease
NGS	next generation sequencing
OCSCC	oral cavity squamous cell carcinoma
OPSCC	oropharyngeal squamous cell carcinoma
ORF	open reading frame
OS	overall survival
PCR	polymerase chain reaction
PDX	patient derived xenograft
PFS	progression free survival
qPCR	quantitative polymerase chain reaction
qRT-PCR	quantitative reverse transcription polymerase chain reaction
RT	radiotherapy
RT-PCR	reverse transcription polymerase chain reaction
SCAF	genomic scaffolding region
TCGA	the cancer genome atlas
URR	upstream regulatory region
WES	whole exome sequencing
WGS	whole genome sequencing

Abstract

There has been a growing epidemic of human papillomavirus (HPV)-induced oropharyngeal squamous cell carcinoma (OPSCC) for the last few decades. This is an underappreciated disease that can have a devastating effect on the lives of otherwise healthy young patients. Patients with HPV-positive OPSCC generally have a good prognosis, but still 20-30% fail to respond to therapy or later recur for unknown reasons. Our hypothesis is that the type of interactions between viral DNA and the human host DNA may determine a patient's disease progression. HPV can remain its circular episomal shape, but often linearizes and integrates into the human genome, either into intergenic loci or into genes. The process of HPV integration is of particular interest as a potential driver of HPV-positive OPSCC because it is thought to be a marker of disease progression in cervical cancer and is reported in a large proportion of head and neck tumors with estimates ranging from 50-70%.

Integration can lead to large structural variations, disrupt cellular genes and alter gene expression both locally and genome wide, but the exact effects of HPV integration on OPSCC progression are unclear. Previous studies that assessed survival differences between HPV integration positive versus negative patients demonstrated mixed results, so we aimed to clarify whether this process impacts patient outcomes. Using a polymerase chain reaction (PCR)-based approach, we found that HPV-positive, integration-positive patients had higher levels of the HPV oncogenes E6/E7 and a survival advantage over HPV-positive, integration-negative patients. The underlying mechanism for this improved outcome is unclear, but this work provides evidence that HPV integration could serve as a prognostic biomarker.

We also utilized this methodology to investigate the clonal nature of HPV integration events, as it is clear that this process affects cell biology given the survival differences we discovered, but how cells containing these viral-human fusions may be selected for during tumor evolution is unclear. We explored the clonality of integration events in bilateral HPV+ tonsillar tumors and found evidence that these tumors often form as a result of clonal expansion from one tonsil to another given that we found shared integration sites across samples. These results indicate that integration events provide a survival advantage to tumor cells which are then selected for and expanded such that they are able to metastasize elsewhere.

Finally, to overcome limitations of previous integration calling methodologies, we optimized a new targeted capture sequencing and analysis pipeline called SearcHPV. Through integrated analysis of HPV+ models by SearcHPV and genome-wide linked read sequencing, we demonstrated that HPV integration sites were found not only adjacent to known cancer-related genes such as *TP63* and *MYC*, but also near regions of large structural variation in the tumor genome. Further, analysis of SearcHPV-assembled junction contigs demonstrated that the tool can be used to accurately identify viral-host junction sequences and showed that viral integration occurs through a variety of DNA repair mechanisms including non-homologous end joining, alternative end joining and microhomology mediated repair. Together, these studies highlight HPV genomic integration as an important contributor to cancer progression, and with new tools available, we believe the field is now primed to make major advances in the understanding of HPV-driven pathogenesis, some of which may lead to the development of novel biomarkers and/or treatment paradigms.

Chapter 1 Introduction

Human Papillomavirus

Discovery

Human papillomaviruses (HPV) are double-stranded, small DNA viruses transmitted through sexual contact that infect human epithelium in anogenital and oral mucosa. There are more than 200 HPV types described with varying epithelial tropism and associated conditions [1]. Infection with one or more HPV strains is nearly ubiquitous in sexually active persons. These infections are usually asymptomatic and clear spontaneously. However, persistent infections may lead to a variety of HPV-mediated diseases, including skin warts, genital warts, precancers, cancers of cervix, anus, penis, vulva, vagina, and head and neck-particularly cancers of the oropharynx [2].

The investigation of human papillomavirus as a potential cause of warts and cancers has been ongoing for centuries. The infectious nature of skin and genital warts has been described dating back to the ancient Greek and Roman cultures, but their viral nature wasn't clear until the mid-1900s [3]. An Italian physician Rigoni-Stern noted in 1842 that cervical cancer occurred frequently in married women and prostitutes but almost never in nuns, leading him to conclude development of cervical cancer was related to an infection acquired during sexual contact [4]. Studies of rabbit papillomaviruses in cottontail and domestic rabbits in the mid-1900s demonstrated the carcinogenic potential of these viruses and helped inspire interest in studying human papillomaviruses in the context of cancer [5-7]. The first description of the double-stranded circular DNA structure of HPV was published in 1965 [8, 9]. zur Hausen began investigating the potential role of HPV in cervical cancer in 1972 based on anecdotal evidence of malignant conversion of genital warts into squamous carcinomas, and later, he and others described novel HPV types in genital warts and laryngeal papillomas (HPV6 and HPV11) as well as cervical cancers (HPV16 and HPV18) [10-12]. A large-scale epidemiological study examining the association of cervical cancer and HPV in nearly 10,000 samples was first published in 1987 [13]. During this same time period, it was noted that women with cervical cancer had an increased risk for development of oral cancer, suggesting that HPV may also play an important role in head and neck cancer [14]. Loning and de Villiers described in 1985 the first reports of specific HPV types in oropharyngeal cancers [15, 16]. Larger studies examining the rates of HPV positivity in oropharynx cancer have since been published, estimating an HPV-positivity rate of between 25-60% [17-20].

Association with Disease

Human papillomaviruses are part of the *Papillomaviridae* family and are grouped into six genera based on homology in the L1 open reading frame (ORF), which encodes for the viral capsid protein that coats the virus [21]. These genera are known as alpha, beta, gamma, delta, mu and nu; each member of a given genus has at least 40% homology. These genera are further grouped by HPV species, which share 60-70% L1 homology, as well as common features and epithelial tropism. HPVs are then distinguished by type, of which there are over 200 currently classified. Each HPV type is at least 10% different in L1 sequence from its closest neighbors.

The alpha-papillomaviruses infect human mucosal and cutaneous epithelium and contain HPV types that are implicated in both benign and malignant lesions [22]. A subset of 14 mucosal HPV types within this genus is considered high-risk (hrHPV) based on their frequent association with various cancers and transforming ability in model systems. This includes HPV types 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66 and 68 [23]. These hrHPV types have been reported as the causative factor not only in cervical and oropharynx cancers but in other cancer types as well, including penile, anal, vulvar, and vaginal cancers [2]. hrHPV has also been reported in cases of esophageal [24, 25], colon [26, 27], bladder [28], prostate [29], ovarian [30, 31], breast [32], and lung cancers [33], but the role of HPV in these cancer types is unclear and is not currently recognized as causative by the International Agency for Research on Cancer [23]. hrHPV types 16 and 18 are known to cause the majority of cervical and oropharyngeal cancers [34, 35]. Among oropharynx cancers studied at the University of Michigan, roughly 90% are associated with HPV16 and 10% are associated with other hrHPV types [36]. Despite their association with cancer, the majority of individuals infected with a hrHPV type will not develop cancer and will clear the infection [37].

Another subset of HPV types is considered potentially high-risk based on association with cancers, but evidence demonstrating carcinogenicity is lacking, for example HPV types 26 and 53 [38]. Additional HPV types are considered low-risk (lrHPV) because they are primarily associated with benign lesions, genital warts and laryngeal papillomas rather than invasive neoplasias (HPV 6, 11, 42, 43, 44) [22]. However, lrHPVs can sometimes induce non-melanoma skin cancers in immunocompromised individuals, including those with severe combined immunodeficiency (SCID), epidermodysplasia verruciformis (EV), or organ transplant recipients [39, 40]. Rare cases of larynx, lung, nasopharynx, and sinonasal cancer after genomic integration of lrHPV types 6 and 11 have also been reported [41-44].

HPV Biology

HPV16 is involved in the majority of HPV+ head and neck cancers. The HPV16 genome is a 7.9 kilobase (kb) circular genome organized into an upstream regulatory region (URR), 6 early region genes (E1, E2, E4, E5, E6, E7) and 2 late region genes (L1, L2) **(Figure 1.1)**. The process of an HPV infection begins with microtraumas in the epithelium that allow HPV virions to enter and infect the basal epithelium [45]. In the oral cavity, HPV virions most frequently infect the basal cells in tonsillar crypts [46]. Once within the cell, HPV is dependent on the host cell for replication. The viral proteins manipulate cellular pathways for viral genome replication/amplification and coordinate these processes with the cellular differentiation pathway, timing viral capsid production with the later stages of epithelial differentiation. As the differentiated cells reach the surface, HPV episomes are packaged and released.

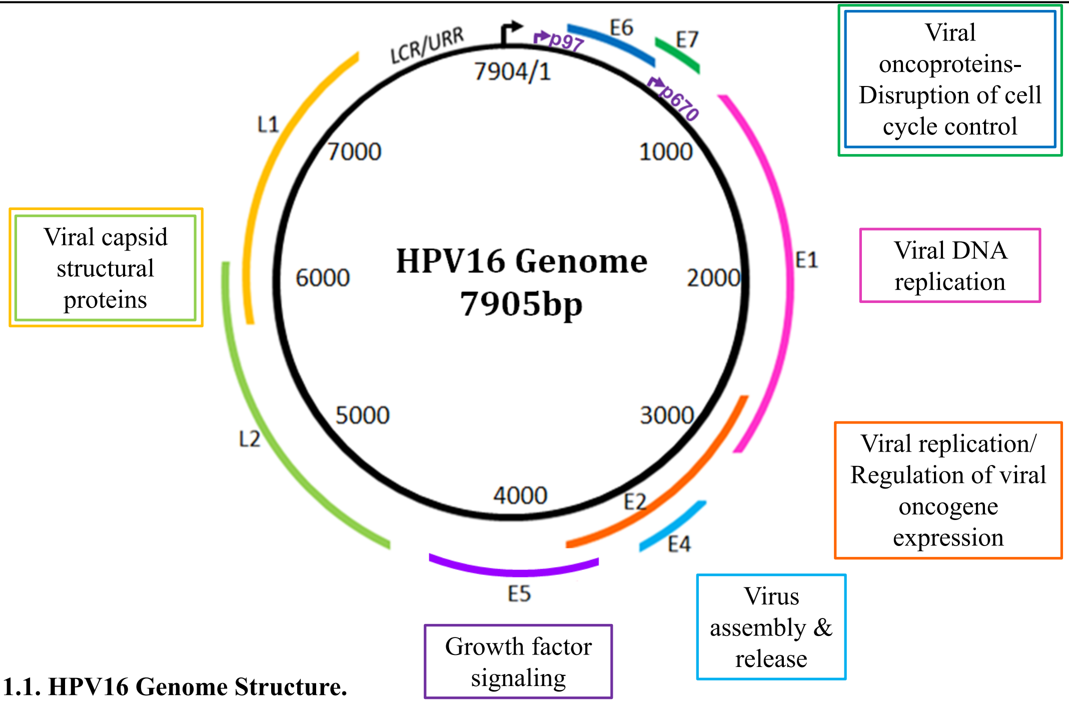


Figure 1.1. HPV16 Genome Structure.

Early in an infection, the HPV oncoproteins E6 and E7 induce host cell replication by blocking the function of key cell cycle regulators, TP53 and RB1 **(Figure 1.2)**. The HPV oncoprotein E6 recruits the cellular E3 ubiquitin ligase, E6 associated protein (E6AP), and binds TP53, leading to TP53 polyubiquitination and subsequent degradation by the 26S proteasome [47, 48]. The destruction of TP53 results in failure of cell cycle arrest and apoptosis, contributing to unrestricted host cell growth and proliferation. The E6 gene can be expressed as full length or alternatively spliced forms, referred to as E6*I, E6*II, or E6*III. These alternate transcripts are thought to be drivers of oncogenic transformation and are expressed at higher levels than full length E6 in tumors samples and cell lines [49].

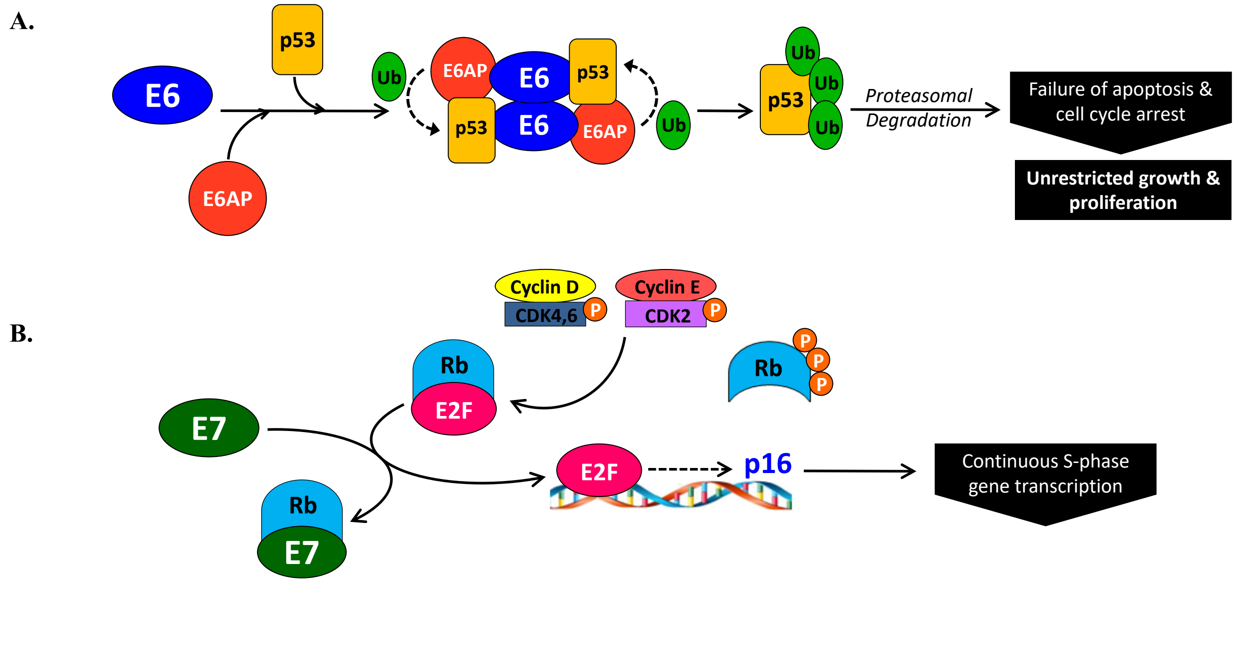


Figure 1.2. Function of HPV oncoproteins to block cell cycle regulators (A) p53 and (B) Rb.

HPV E7 protein sequesters and disrupts the function of the cell cycle regulator RB. In normal cells, RB binds the transcription factor E2F, preventing cell cycle progression. When cell growth signaling occurs, expression of cyclin D1 is initiated. Cyclin D1 activates cyclin dependent kinase (CDK)4/6, leading to monophosphorylation of RB. CDK2 is then activated by Cyclin E and further phosphorylates RB, releasing E2F and initiating transcription of cell cycle entry genes. E2F also activates transcription of p16INK4a (CDKN2A, an inhibitor of cyclin dependent kinase 4 and the off signal for RB phosphorylation), shutting off RB phosphorylation. Ubiquitous phosphatase activity dephosphorylates RB, which re-sequesters E2F and stops cell cycle entry. In the presence of HPV, E7 binds to the pocket of RB, disrupting the interaction with E2F. When E2F is liberated, it leads to continual transcription of S-phase genes, driven by other cell cycle cyclin-CDK complexes. p16INK4a is also inappropriately transcribed and expressed, making it a useful surrogate histological marker of HPV infection. The binding of E7 to RB leads to unscheduled continuous cell cycle entry, progression and cellular proliferation [50].

The hrHPV E1 and E2 genes also play an essential role in viral replication; E2 recruits E1, which acts as a viral DNA helicase, to the replication origin in the URR [51]. This allows for the replication of hundreds of viral copies. Additionally, through its recruitment of cellular factors to the URR, E2 acts a transcriptional repressor of E6 and E7; this negative feedback loop allows for the coordination of the viral life cycle with the cellular differentiation process [51]. The later expression of L1 and L2 in the uppermost layers of the epithelium allows for packaging, assembly and release of the virus [52]. In most cases, the infection will become latent within 1-2 years, but some people will have a persistent HPV infection which can lead to the development of precancers or cancers. Persistent HPV infection that leads to carcinoma is characterized by high expression of E6 and E7 and frequent loss of E2, leading to unregulated expression of E6 and E7, which promotes genomic instability, oncogenic transformation, and clonal expansion [53].

Oral Infection Epidemiology

Although HPV is one of the most common sexually transmitted infections, persistent oral HPV infections are relatively rare in the normal population. Data from the National Health and Nutrition Examination Survey showed that the prevalence of oral HPV infection in the United States is approximately 4-6% [46, 54]. There are many factors associated with increased risk of oral HPV infection, including age, sex, race, vaccination status, number of sexual partners, and current smoking habits. Oral HPV infections in women peak in prevalence around ages 25 and 55, following a bimodal pattern. In men, risk of infection increases with age, peaking at age 60. Men are more likely to be infected, and higher numbers of sexual partners and smoking are associated with increased risk of infection. Race also plays a factor; white women are less likely to have an oral infection than women of other races, and black men were more likely to have an infection. Asian people of both sexes living in the U.S. have lower infections than white men and women. Men and women who receive the quadrivalent HPV vaccine are protected against both oral and genital infections of HPV 6, 11, 16 and 18 but not necessarily all HPV types [54].

Most HPV infections are naturally cleared, but in women, it has been well established that persistence of a genital HPV infection is a significant risk factor for developing cervical squamous cell carcinoma [55, 56]. However, the factors that contribute to persistent oral infection and the natural history of oral HPV infections leading to cancer have still not been well characterized.

Head and Neck Squamous Cell Carcinoma

Etiology

Head and neck squamous cell carcinomas (HNSCC) are a heterogeneous group of cancers that arise in the mucosal lining of the upper aerodigestive tract. As a whole, HNSCC accounts for 650,000 cancer cases and 330,000 deaths annually worldwide [57]. In the United States, HNSCC accounts for approximately 4% of all cancers and presents more frequently in men over the age of 50 [58]. These cancers can arise anywhere in the squamous epithelium in the head and neck region, including oral cavity, pharynx (divided into nasopharynx, oropharynx and hypopharynx), larynx, paranasal sinuses/nasal cavity and salivary glands. The classical risk factors for the development of HNSCC are smoking, excessive alcohol consumption and HPV infection.

HNSCC can be subdivided by HPV status; HPV-positive and HPV-negative oropharynx tumors take different clinical courses and have different outcomes **(Table 1.1)**. HPV-positive cancer is more likely to develop in the oropharynx than anywhere else in the head and neck region, and it is suggested that this is due to the architecture of tonsillar crypts in the oropharynx, which act as a reservoir for HPV. However, there are cases of HPV+ HNSCC arising in other anatomical sites within the head and neck. The palatine and lingual tonsils are the most common sites of origin for HPV-induced oropharyngeal cancer (OPSCC) [18, 59]. HPV-positive OPSCC patients tend to be diagnosed at a younger age and have fewer overall health problems than patients with HPV-negative disease [60]. Patients with HPV-positive OPSCC have a survival advantage over those with HPV-negative OPSCC regardless of treatment modality [18, 61-63]. HPV-positive OPSCC patients tend to respond better to chemoradiation therapy [64] and have enhanced radiosensitivity [65]. However, patients with HPV-positive oral cavity squamous cell carcinoma (OCSCC) do not have the same survival advantage over HPV-negative OCSCC patients [66]; some studies suggest they have a worse prognosis [67]. A meta-analysis by Ragin showed no survival difference between HPV-positive and HPV-negative patients with cancer at non-oropharyngeal sites of the head and neck, including oral cavity [68]. The source of the discrepancy in outcome between the oropharynx and oral cavity is not entirely clear, but differences in immune response from site to site may be an important factor [69].

	HPV+ HNSCC	**HPV- HNSCC**
Anatomic Site	Oropharynx	No predilection
Age	Younger	Older
Survival	Improved at oropharynx, but not other sites	Poor
Incidence	Increasing	Decreasing
Risk factors	Persistent HPV infection	Tobacco and alcohol use
Mutational burden	Relatively low; frequent *PIK3CA* mutations	Highly mutated; frequent *TP53, PIK3CA,* *EGFR, NOTCH* mutations
Treatment	Surgery + CRT; de-escalation trials ongoing	Surgery + CRT

Table 1.1. Differences between HPV+ and HPV- HNSCC.

Abbreviations: CRT, chemoradiotherapy.

The incidence of HPV-associated head and neck cancers has been increasing rapidly over the past few decades [70, 71]. Three out of four new oropharynx tumors are HPV-related [72]. This is in contrast to HPV-negative head and neck cancers, which have been declining in incidence primarily due to public health efforts to decrease smoking. The incidence of cervical cancer has also been declining due to improved screening and detection; in 2009 there were more incident cases of oropharynx cancer than cervical cancer in the US [73]. Therefore, HPV+ OPSCC represents an increasingly concerning public health threat. Although the overall five-year survival rate is relatively high for HPV+ OPSCC (~80%), there is still a subset of patients who fail to respond to therapy or later recur for unknown reasons [74].

Genomic landscape

HNSCC as a whole has been studied extensively in large scale studies like The Cancer Genome Atlas (TCGA). TCGA mutational analysis shows frequent mutations and alterations in genes in critical cellular pathways including cell cycle control (*TP53, CDKN2A, CCND1*), tumor cell survival (*PIK3CA, PTEN*), growth signaling (*EGFR*), WNT signaling (*FAT1, AJUBA, NOTCH1*) and epigenetic regulation (*KMT2D, NSD1*) [75, 76]. However, the vast majority of HNSCC tumors in the TCGA are HPV-negative, which are considered a clinically distinct entity and have different biology than HPV-positive tumors. Only about seventy HPV-positive tumors are currently represented in the TCGA, and the majority of these samples are from large, aggressive tumors [77]. In general, these HPV+OPSCC tumors do not contain *TP53* mutations and have a relatively low somatic mutational burden. The most frequently altered gene in HPV+ HNSCC is *PIK3CA*; many patients have activating mutations or amplifications at its locus on 3q26 [76].

In a small cohort of HPV+ OPSCCs, Zhang et al. reported that HPV+ HNSCCs could be separated into two groups with distinct gene expression signatures, one enriched for mesenchymal and immunological response genes (HPV-IMU) and the other enriched for keratinocyte differentiation genes (HPV-KRT) [78]. HPV-IMU tumors were enriched for chromosome 16q losses, and HPV-KRT tumors were enriched for chromosome 3q copy number alterations (CNAs) and activating mutations in *PIK3CA*. There was no significant difference in survival between the two groups, but their sample size was likely not large enough to power the analysis. Due to underrepresentation in sequencing studies, further investigation into the somatic mutations seen in HPV+ samples needs to be done in order to assess potential drivers of carcinogenesis in this subset of patients, especially in patients with lower grade disease.

Treatment

Although it has long been recognized that HPV-positive and HPV-negative HNSCC are distinct clinical and biological entities, the current treatment protocols mostly do not differ based on HPV status. Patients with advanced OPSCC, regardless of HPV status, are treated with primary surgery and adjuvant chemoradiotherapy (CRT) or definitive concurrent cisplatin-based CRT [79]. These treatments are known to cause acute side effects, including mucositis and loss of taste, as well as more serious long-term health problems like dysphagia, xerostomia, hearing loss, neck muscle fibrosis, trismus, and osteoradionecrosis of the jaw [80]. Given that HPV+ patients on average are younger and have longer life expectancies, these toxicities can seriously damage their quality of life for decades. Therefore, there has been a big push in the field to de-intensify the therapy for HPV+ disease with the goal of improving quality of life and reducing treatment-induced harm while maintaining the survival rates seen with the current standard of care. Several clinical trials aiming to deintensify therapy for these patients have already been completed with many more currently underway, but the current challenge is to stratify patients into the appropriate risk group as the clinical and molecular markers for poor prognosis are not yet entirely understood [81-84].

HPV Genomic Integration

In cervical SCC

HPV typically persists in cells as a circular episome but can also linearize and become integrated into the host genome. It has been of great interest to understand the implications of integration and to determine whether it is involved in tumor formation. HPV is commonly found integrated into the host genome in cervical cancer [85, 86]. Integration of HPV is characteristic of cervical lesion progression but may not be required for tumor formation [85, 87]. Early studies investigating the role of integration in cervical lesions showed that integration is a stochastic process or favors a preference for common fragile sites, regions of microhomology, highly transcriptionally active regions, or near microRNAs (miRNAs) [87-89]. There were few reported examples of integration into genes that led to a disruption of gene expression, and in general, integration was not presumed to have any major impact on gene expression. Only one study during this time period reported effects on gene expression; they showed integration near the *cMYC* locus on chromosome 8q24 led to overexpression of *cMYC* [90].

Later studies, however, showed that integration of HPV might represent an additional oncogenic mechanism through direct alteration of cancer-related gene expression. One study showed that the majority of integration events occur in known or predicted genes or near miRNAs, which have major roles in regulation of cellular processes [91]. Tian et al. recently demonstrated frequent integration into non-coding genes known as long intergenic non-protein coding RNAs (lincRNAs) [92]. Hu et al. showed that integration events occur in genomic hotspot regions and may function to inactivate or activate genes that favor clonal expansion [89]. Bodelon and colleagues analyzed over 1200 integration events in cervical cancers and reported that integration occurred most frequently at three loci: 3q28, 8q24.21, and 13q22.1 [93]. These regions all are gene-rich and contain important tumor suppressors, including *TP63, TPRG1, cMYC, KLF5* and

KLF12. They also reported that integration into genes occurs more often than expected by chance and may lead to functional alteration of important genes. Using an advanced technique of HPV-capture sequencing, Holmes et al. was able to distinguish five different HPV structures in cervical cancers: episomal (EPI), single integration in either a colinear (2J-COL) or noncolinear (2J-NL) fashion associated with chromosome deletion or amplifications respectively, and multiple integrations either clustered in one locus (MJ-CL) or scattered at different loci (MJ-SC) [94]. In their cohort (n=72), they reported a relatively even distribution of each structure (29% EPI, 24% 2J-NL, 17% MJ-CL, 16% 2J-COL, 11% MJ-SC), indicating that cervical cancers most frequently have at least one HPV integration site. Integration events in cervical cancer have been better described than in head and neck cancer, but still much is not understood about the role integration plays on the progression from dysplasia to invasive carcinoma.

In HNSCC

Like in cervical cancers, there is no consensus sequence or one location HPV integration is known to target in oral and oropharyngeal cancers. Integration breakpoints have been reported throughout the cellular genome. In HNSCC cell lines, Akagi reported that HPV insertional breakpoints were found at regions of genomic amplification or deletion and demonstrated an association of insertional breakpoints with structural variation, including chromosomal translocations, deletions/insertions, and rearrangements [95]. Walline investigated nine HNSCC cell lines and found integration in all cell lines throughout the cellular genome, eight of which had integration into cancer-related genes [96]. Parfenov and colleagues analyzed the genomic landscape of thirty-five HPV-positive HNSCCs in TCGA, including both OPSCCs and OCSCCs, by whole genome sequencing (WGS) and found over one hundred integration sites in 25 of the tumors [97]. Integration into a known gene was seen in 54% of the events and 17% integrated within 20kb of a gene. Nulton et al. expanded upon this previous study to include 72 HPV-positive HNSCCs in the TCGA and reported that 23% of tumors showed HPV integration into the genome consistent with a partial deletion of E1/E2 and 44% of tumors showed episomal HPV [98]. They hypothesized in the remaining 33% of tumors, HPV had integrated and then was excised as a viral hybrid episome that can replicate autonomously. Another recent study reported a higher frequency of HPV integration in HNSCC (71%) and reported structural changes in the human genome near the integration site in some cases but not all [99].

In other anogenital cancers

HPV integration in other anogenital cancers has not been as widely studied as in cervical or HNSCC, and therefore it is unclear if they follow a similar mechanism for HPV-driven carcinogenesis. A small cohort of penile cancers was examined for HPV integration; 73% of samples had integrated HPV and 27% had episomal HPV [100]. Frequent HPV integration has also been reported in vulvar squamous cell carcinoma [101]. In 2019, Morel et al. reported the integration status in 93 anal squamous cell carcinomas determined by an HPV-capture sequencing method [102]. Similar to Holmes et al., they separated tumors into 5 categories: EPI, 2J-COL, 2J-NL, MJ-CL, MJ-SC but the signatures they reported differed significantly. There was a much higher proportion of tumors with episomal HPV (45% vs 29%), and the most common form of integrated HPV was multiple junctions scattered across different loci (27% vs 11%). Interestingly, four patients showed integration into the cellular gene *NFIX*, and each of these patients had a complete response to therapy and a longer overall survival than the other patients in the study.

HPV integration detection methods

Whether integration of HPV is required for malignant transformation in oral/oropharyngeal cancers is not clear. The wide variety of techniques used to detect integration events makes it challenging to compare results of different studies. Some methods try to establish the physical state of the virus as episomal, integrated or a mix of both within a given sample. The most frequently used method of this type is measuring the ratio of E2/E6 gene expression. This method is based on the hypothesis that during integration, the E2 gene is disrupted, leading to increased levels of E6. A ratio is made comparing the expression levels of the E2 and E6 genes as measured by quantitative reverse transcription polymerase chain reaction (qRT-PCR), assuming that a ratio of one means HPV is episomal and a ratio of less than one means HPV is integrated. This method is not as effective as others because it is based on the assumption that E2 is always disrupted and E6 is always increased during integration, which has been shown to not be true in all cases [89, 97, 103]. A newer method was recently developed to distinguish the episomal state from the integrated state using exonuclease V, which can only digest linear DNA [104]. A sample is digested and then quantitative polymerase chain reaction (qPCR) is performed to the E6 region of HPV; if HPV is still present in the sample, that indicates HPV was not in a linear state and was therefore in an episome. This method is useful to characterize the physical state of HPV but gives no information about the location or number of HPV integration sites.

There are many other methods used to detect and characterize HPV integration sites. The most commonly used methods are Detection of Integrated Papillomavirus Sequences (DIPS-PCR), Amplification of Papillomavirus Oncogene Transcripts (APOT), whole-genome sequencing (WGS), whole-exome sequencing (WES), and RNA-seq. DIPS-PCR and APOT are polymerase chain reaction (PCR)-based methods used to detect fusions at the DNA and RNA level respectively. These two methods are technically simpler and cheaper options than larger-scale sequencing methods like WGS or RNA-seq but may be unable to detect all integration sites and complicated structural changes within samples. Therefore, WGS and RNA-seq may better reflect the true complexity of viral integration using tiling of paired ends across the genome, but these methods have limitations of their own. Due to the rare nature of integration events in the context of the entire genome, these methods may not have enough depth to fully characterize an integration site. To overcome this sensitivity problem, other groups have begun using HPV-capture technology before sequencing to enrich for these sites and to get deeper reads of these regions of interest. While this vastly improves the sensitivity, it does not help overcome the limitation of all next-generation sequencing methods of short-reads; these integration events are vastly complicated and short reads limit our ability to generate assemblies that are large enough to capture this complexity. Newer long-range technologies such as linked read sequencing, or PacBio and Nanopore long-range sequencing systems might be able to generate this type of data required to examine these large-scale rearrangements.

Integration Mechanism

The exact mechanism of HPV integration into the host genome is not known. In most models, both the viral and cellular genomes undergo breakage, allowing for fusion between the two. Some groups assert that fusion occurs as a result of cellular repair mechanisms, including non-homologous end joining and homologous recombination [105]. However, others have criticized these proposed mechanisms because small numbers of breakpoints are seen even when many copies of HPV are present, which argues against random breakpoints [95]. There are two main mechanisms that have been described: direct integration into the genome or looping integration **(Figure 1.3)**. Direct integration into the host genome can result in insertion of a single copy or multiple concatemerized copies of HPV; in either case, both the HPV and host genomes undergo deletions of the flanking sequences [106]. Alternatively, Akagi developed a looping model for focal genomic instability to explain the genomic structural variations seen in HNSCC cell lines using a chromosomal mapping technique to determine the DNA structure surrounding integration sites [95]. In this model, both the host and viral genomes are nicked, the viral genome is inserted, and a circular piece of DNA containing both is transiently formed, resulting in rolling circle amplification. This amplification leads to concatemer formation characterized by amplified segments of genomic sequence flanked by HPV segments. This is consistent with reports from patient tumors with focal copy number elevation at sites of HPV integration [75, 97]. Looping amplification can also result in the creation of extrachromosomal HPV-human fusion episomes, as has been proposed by Nulton et al [98].

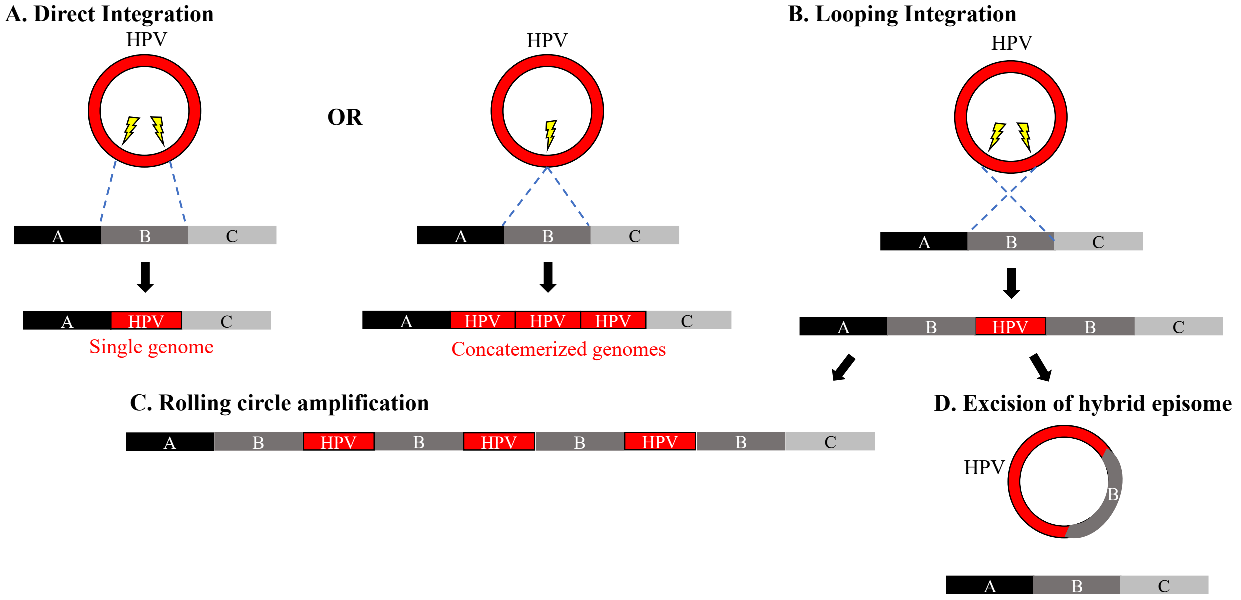


Figure 1.3. Mechanisms of HPV integration. A) Two types of direct integration B) Looping integration resulting in C) Rolling circle amplification or D) Excision of an HPV-human episome. Adapted from Groves 2018. Copyright permission received on April 27, 2020, license number 4817180283904.

Role in oncogenesis

Integration has been thought to promote oncogenesis through the dysregulation of the oncoproteins E6 and E7, resulting in increased cellular proliferation and genetic instability [107]. Dysregulation of E6 and E7 gives the cells a selective growth advantage and allows for oncogenic progression. Multiple events have been described that result in dysregulation of E6/E7, including (1) disruption of E2 or its binding sites, (2) disruption of E1, (3) formation of stable viral-host transcripts, or (4) generation of a viral super-enhancer from repeats of regulatory elements [108].

E2 is responsible for regulation of E6/E7, so disruption of the E2 gene or its binding sites allows for unregulated E6/E7 transcription. E2 can be disrupted at the genomic level or at the transcriptome level through integration-induced gene fracture and loss of expression of E2 itself or the upstream genes (E6, E7, E1). Methylation of the E2 binding sites in the URR can also lead to increased E6/E7 because the E2 protein is no longer able to bind and recruit cellular factors to repress their transcription. High levels of methylation at these sites has been reported frequently in HPV16+ cervical carcinomas [109, 110].

When E1 is disrupted, lack of replicative functions can induce DNA damage and growth arrest, promoting focal instability at the site of integration [111]. E1 has been reported to be the viral gene most likely to be involved in integration breakpoints [97, 112]. In a group of cervical carcinomas, Brant et al. reported that the donor splicing site in E1 was recurrently involved in viral-cellular fusion transcripts, even when the integration junction occurred at a different position at the DNA level [113]. Chimeric transcripts were formed as a result of splicing between the viral donor site with a nearby acceptor splicing site in the human genome, resulting in disruption of E1 and E2 expression.

It has been shown that integration can generate hybrid E6/E7 viral-host fusion transcripts, which are often more stable than viral E6/E7 transcripts due to loss of viral AU-rich elements in the 3’ UTR [107]. Ehrig et al. cloned episomal-derived viral transcripts and a small subset of viral-cellular fusion transcripts and compared their stability; they reported that the E6/E7 transcripts derived from episomal HPV were less stable than the fusion transcripts [114]. This increased stability may contribute to sustained higher expression of E6 and E7.

Dooley recently showed that tandemly integrated copies of the HPV16 genome can generate a super-enhancer-like element that can drive transcription of E6/E7 [115]. Super-enhancers are clusters of traditional enhancers that are associated with the expression of oncogenes; binding of transcription factors and chromatin regulators like Brd4 are enriched at super-enhancers. Brd4 is an epigenetic regulator that recognizes acetylated lysine residues and recruits transcriptional complexes. This study reported that Brd4 activates viral transcription at these tandem integrated HPV sites, and treatment with an inhibitor of Brd4 resulted in decreased E6/E7 transcription and inhibited cellular proliferation.

However, both Parfenov and Olthof reported that there are tumors with HPV integration that do not have increased levels of E6/E7 [97, 116]. Olthof reported that there was no significant difference in E2, E6 or E7 levels between integrated versus non-integrated tumors. This suggests that increased E6/E7 is not always the main driver of oncogenesis.

Effect on cellular gene expression

Integration has traditionally been thought of as promoting oncogenesis through sustained expression of E6 and E7. However, integration has more recently been shown to have effects on cellular gene expression, which may represent an additional oncogenic mechanism in the development of HNSCC. Parfenov saw increases in somatic DNA copy number of the integrated region and reported that gene disruption occurs by integration through several key mechanisms: tumor suppressor loss of function, enhanced oncogene expression, and rearrangements that lead to altered gene expression.

Loss of function of a tumor suppressor occurs when HPV integration into a gene results in deletion of gene regions and generates truncated transcripts, as well as host-viral fusion transcripts. Parfenov reported integration into *RAD51*, resulting in a twenty-eight-fold amplification extrachromosomally, leading to alternate transcripts being generated and likely non-functional

RAD51 protein. They also reported integration into *ETS2*, which led to deletion of exons 7 and 8. The overall expression of the gene was unaffected but transcription of exons 7 and 8 was decreased, likely leading to a truncated protein.

HPV integration upstream of an oncogene can lead to oncogene overexpression via amplification of the nearby downstream region, leading to elevated transcripts. Parfenov reported viral integration upstream of *NR4A2*, leading to a 250-fold amplification of the downstream region and subsequent overexpression of *NR4A2*. NR4A2 is a transcription factor that is overexpressed in a wide variety of human cancers [117]. Parfenov also reported interchromosomal translocation of chromosomes 3 and 13, which caused overexpression of key oncogenes *KLF5*, *TP63*, and *TPRG1*.

Walline characterized integration sites of eight HPV-positive HNSCC cell lines (seven HPV16 and one HPV18) by DIPS-PCR [96]. Integration into cancer-related genes was detected in all of the HPV16 cell lines. The HPV18 cell line, UM-SCC-105, had two integration events but both were intergenic. In UM-SCC-104, viral integration of HPV16 E1 into the tumor suppressor *DCC* was detected. When the transcripts of the *DCC* gene were interrogated, no transcripts were generated. This demonstrates an example of viral integration leading to disruption of a tumor suppressor, potentially providing a growth advantage for those cells. In UM-SCC-47, integration into *TP63* resulted in the generation of a hybrid viral-host fusion transcript between exon 14 of *TP63* and HPV16 E2, which resulted in a truncated ΔNTP63 protein as shown by Western blot. The other cell lines did not exhibit viral-host fusion transcripts, potentially due to integration in frame into introns that were subsequently spliced out.

Akagi investigated whether the rearrangements resulting from integration generated cell-virus fusion transcripts and altered cellular gene expression. In all ten HNSCC cell lines analyzed and in one primary tumor, they found virus-host fusion transcript expression, which frequently confirmed the rearrangements described by WGS. They also reported multiple examples of gene disruption at sites of integration. In UD-SCC-2, HPV integration led to deletions and rearrangements of segment of *DIAPH2*, which resulted in viral-fusion transcripts but no native transcripts or functional protein. In UM-SCC-47, they reported aberrant *TP63* expression due to HPV integration-mediated amplification, leading to viral-host transcripts and a truncated TP63 protein. They saw additional examples of gene disruption, including amplification of the oncogenes *FOXE1* and *PIM1* in UPCI:SCC090 cells.

Multiple groups have examined transcriptome-wide differences between integration-positive and integration-negative tumors. Huebbers et al. showed that integration-positive tumors have significantly deregulated expression of genes related to epidermal development and differentiation, hormone regulation and processing, oxidative stress and metabolic processes compared to integration-negative tumors [118]. They specifically found that integration-positive tumors had overexpression of *AKR1C1* and *AKR1C3*, which are members of the aldo-keto reductase superfamily of NADPH-dependent oxidoreductases. They further reported that HPV+ OPSCC patients with overexpression of these proteins had a significantly worse survival than those with low expression. Zhang et al. reported that HPV+ HNSCCs could be separated into two groups with distinct gene expression signatures, one enriched for mesenchymal and immunological response genes (HPV-IMU) and the other enriched for keratinocyte differentiation genes (HPV-KRT) [78]. When they analyzed the integration status of the tumors, as assessed by viral-host fusion transcripts, they noted that the HPV-KRT group was enriched for samples with HPV integration into cellular genes, suggesting that integration can alter the expression of these cellular pathways. This group went on to publish an additional study directly examining the gene expression signature patterns between tumors with and without viral-cellular transcripts [119]. Samples without fusion transcripts were enriched for genes related to the adaptive immune response, including lymphocyte and leukocyte activation and activity. Samples with fusion transcripts were enriched for genes related to ribosomal biogenesis, keratinization, and cell-cell adhesion. These different gene expression patterns suggest HPV integration alters the biology of the cells as it relates to immune response, metabolism and other critical cellular processes, the functional consequences of which have not been evaluated.

However, Olthof examined patient tumors and saw no significant effect of integration on gene expression nor were mRNA levels of disrupted genes significantly different. Even when HPV was integrated directly into a gene, the mRNA expression levels were not significantly different from a non-disrupted gene elsewhere in the genome. Either there are other expressed gene copies present that allow overall expression levels to be unchanged, or viral integration did not deregulate genes as assessed by their method.

Deregulation of miRNAs in HPV-positive HNSCCs could result from HPV integration near miRNA sites as has been shown in cervical cancer [89, 91] and HNSCC cell lines [120]. HPV-positive and HPV-negative HNSCCs have distinct miRNA expression patterns, and miRNA subsets were significantly associated with overall survival, disease-free survival, and distant metastasis in HPV-positive HNSCCs [121, 122]. Hui et al. reported 128 miRNAs that were differentially expressed between tumor and normal tissue in OPSCCs and speculated that integration of HPV into the genome near these miRNAs contribute to their deregulation. Wald et al. reported a subset of miRNAs that had altered expression in HPV16-positive HNSCC cell lines as compared to both HPV-negative HNSCC cell lines and immortalized normal keratinocytes [120]. The HPV16-positive cell lines used in this study all have been reported to contain integrated

HPV, suggesting a possible role of integration on the deregulation of miRNAs.

Effect on viral gene expression

Many studies investigating HPV integration report breakpoints throughout the viral genome, with an increased incidence in E1 [97, 112]. The effects of integration on viral gene expression are still not entirely known. Akagi reported that loss of viral segments upon integration or rearrangement contributes to non-uniform coverage of the viral genome when analyzed by RNAseq. Despite this, viral fragments containing E6 and E7 were retained and all samples had strong E6/E7 expression. Walline also reported enhanced E6/E7 expression upon integration, particularly the splice isoform E6*I, and reduced E1/E2 expression in integration-positive cell lines [96]. E6* transcripts are thought to be drivers of tumor development, so the expression of this isoform at the expense of full length E6 is significant. Despite many reports of enhanced oncoprotein expression, Parfenov reported that this does not occur in all integration-positive tumors. Although integration-negative tumors tended to have higher E2/E5 expression levels and lower E6/E7 than integration-positive tumors, this was not always the case. They reported no correlation between the presence of integration within specific HPV genes and their expression level. These results further support the view that HPV plays a larger role in oncogenesis beyond viral oncoprotein expression and subsequent disruption of the P53 and RB axes.

Clinical utility of integration status/site

In cervical cancer, it was long believed that HPV integration was a required event for a lesion to progress from low-grade to high-grade. Tian et al. recently reported HPV integration frequency increases gradually through the different stages of carcinogenesis (infected but normal epithelium < cervical intraepithelial neoplasia (CIN) 1 < CIN2 < SCC) [92]. However, other studies have reported that some cervical cancers show only HPV episomes [85, 87]. Given this, it has been unclear whether HPV integration, through all of its effects on both the human and viral genomes, has an impact on the progression of a patient's course of disease. Many studies in both the cervical and head and neck literatures have attempted to assess the relationship between HPV integration status and patient outcomes with conflicting results depending on the methods used. Discovering this relationship would help determine whether HPV integration status should be evaluated in routine clinical practice as a predictive or prognostic factor.

By evaluating cervical SCC tumors for integration using the E2/E6 or E2/E7 method, multiple groups reported that patients whose tumors contained integration had significantly worse outcomes than patients with only episomal HPV. Shin et al. showed in a cohort of 110 patients that women with episomal HPV had significantly better disease-free survival (DFS) than women with any integration events (integrated only and mix of integrated/episomal)[123]. Ibragimova et al. showed in a cohort of 140 patients that women with HPV integration had a significantly worse progression-free survival (PFS) and overall survival (OS) compared to women with episomal HPV, even when tumor stage was controlled for [124]. Women with a mix of integrated and episomal HPV had intermediate survival. Similarly, another group demonstrated that in women with stage III cervical carcinoma (n=92), those with any integrated HPV (integrated only and mix of integrated/episomal) vs episomal HPV had a 3X higher relative risk of a negative outcome, as well as worse DFS and OS [125].

Multiple groups have also tried to assess the relationship between HPV physical state and outcome in HNSCC. Lim et al. used the E2/E6 ratio assay on 179 HPV+ HNSCCs to differentiate between tumors containing episomal, integrated or both states of HPV [126]. They reported that 12% of the tumors contained episomal HPV only, 24% of the tumors contained integrated HPV only, and the remaining 64% of tumors contained both episomal and integrated HPV, but they reported there was no significant difference in outcomes between these three groups. This is in contrast to the data reported in the cervical literature.

Others have a used a similar principle but focused solely on the status of E2 as a marker for HPV physical state, as it is hypothesized E2 is disrupted/lost when HPV integration occurs. Two groups recently showed concordant results that patients with disrupted/lost E2 DNA and therefore theoretically integrated HPV had worse outcomes [127, 128]. Anayannis et al. demonstrated this in a small cohort of HNSCC patients (n=31) and specifically noted that these patients had a higher risk of locoregional failure (LRF) and lower disease-specific survival (DSS), and Nulton et al. used the HPV+HNSCC TCGA cohort to report that patients with integrated HPV have a significantly worse OS. However, Vojtechova et al. performed a similar E2 analysis on 91 HPV+ HNSCC tumors and reported 27.5% and 72.5% of patients had integrated and episomal HPV respectively, but there was no significant difference in survival between these two groups [129]. Overall, there is some conflicting results in the literature about whether the physical state of the virus can be associated with patient outcome.

There are some limitations to these prior studies; first, it has been previously established that not all cases of HPV integration show disruption or loss of E2, so the integration status based on E2/E6 ratio, E2/E7 ratio or E2 status in these studies may not be accurate and therefore tumors may be misidentified. These methods also do not take HPV copy number into account. Secondly, these studies have focused solely on the physical state of the virus (episomal vs integrated) but have not differentiated based on the locations of those integration sites (intergenic vs in genes/other genomic elements).

In a small cohort study, our group tried to assess whether integration site location has an effect on patient outcome. After observing integration of HPV16 into cancer-related genes in seven

HNSCC cell lines, six established from patients who had progressed, our group investigated integration events in HPV16-positive oropharynx tumors [130]. We hypothesized that responsive tumors are driven primarily by viral oncoprotein expression, but recurrent tumors harbor additional carcinogenic events as a result of HPV integration into cancer-related genes. We expected to see integration into cancer-related genes leading to an alteration in gene expression and potential generation of fusion transcripts in tumors that later recurred but no integration or integration only into cellular intergenic regions in responsive tumors. The integration events in HNSCC tumors from 10 patients were characterized; five were responsive after therapy and five recurred after treatment. Our results supported our hypothesis; tumors from responsive patients had integration events into mainly intergenic loci and tumors from recurrent patients had integration events into cancer-related genes. Only one of the responsive tumors had an integration event into a gene; HPV L1 was found integrated into intron 4 of *TP63* on chromosome 3q28. However, when transcript analysis of the region was performed, no fusion transcript was produced and transcripts across exons 4 and 5, spanning the integration site in intron 4, were produced and were in-frame. This suggests that *TP63* may not be disrupted by this integration, or that at least one intact copy of *TP63* remains unaltered. All other responsive tumors had only intergenic integration events.

In contrast, all five of the tumors from recurrent patients had at least one integration event into an intron of a cancer-related gene. There were seven total gene integration events detected in the five tumors, and upon transcript analysis, four of the events led to gene disruption **(Figure 1.4)**. The other three events did not produce fusion transcripts and retained intact, in-frame cellular gene exon-exon transcripts spanning each respective intronic integration site as well as exon-exon transcripts downstream of the integration site. In tumor 2049 from a recurrent patient, viral integration into *SMOC1* led to generation of a *SMOC1*-HPV E1 fusion transcript. The result of this fusion transcript was inactivation of the gene, demonstrated by the absence of intact cellular exon-exon transcripts surrounding the integration site. Tumor 0843 had integration into *SCN2A*; transcript analysis revealed a complex rearrangement that produced a fusion transcript containing *SCN2A*, HPV L1, and fragments of chromosomes 2q34 and 1q32. This integration event failed to yield intact *SCN2A* exon-exon transcripts downstream of the integration site, suggesting gene disruption. A third tumor, 2238, had two integration events that each resulted in gene disruption. In this tumor, HPV L1 was integrated into *NF1A* and E2 integrated into *SEMA6D*. Neither of these integration events produced fusion transcripts, but disruption of both genes was evident from the lack of cellular exon-exon transcripts spanning the integration sites. This demonstrates that generation of viral/cellular fusion transcripts is not required for cellular gene disruption to occur. All of the tumors, including those from responsive patients, displayed strong E6/E7 gene expression; E6*I was the highest expressed viral gene in eight of the ten tumors [130]. This study, although small, aligns with the work of others that have shown HPV integration may be correlated with worse outcomes. Following this study, Koneva et al. assessed integration status by searching for fusion transcripts within RNA-seq data from the TCGA cohort plus samples from the University of Michigan; they reported that patients with HPV integration had a significantly worse outcome that mirrored HPV-negative disease as compared to patients without HPV integration [119]. Taken together, these results suggest that there are multiple mechanisms leading to integration-mediated cellular gene disruption and that viral integration events can alter gene expression in the host cell. Furthermore, the consequence of these alterations in cellular gene expression may mediate additional carcinogenic mechanisms leading to a more aggressive tumor phenotype.

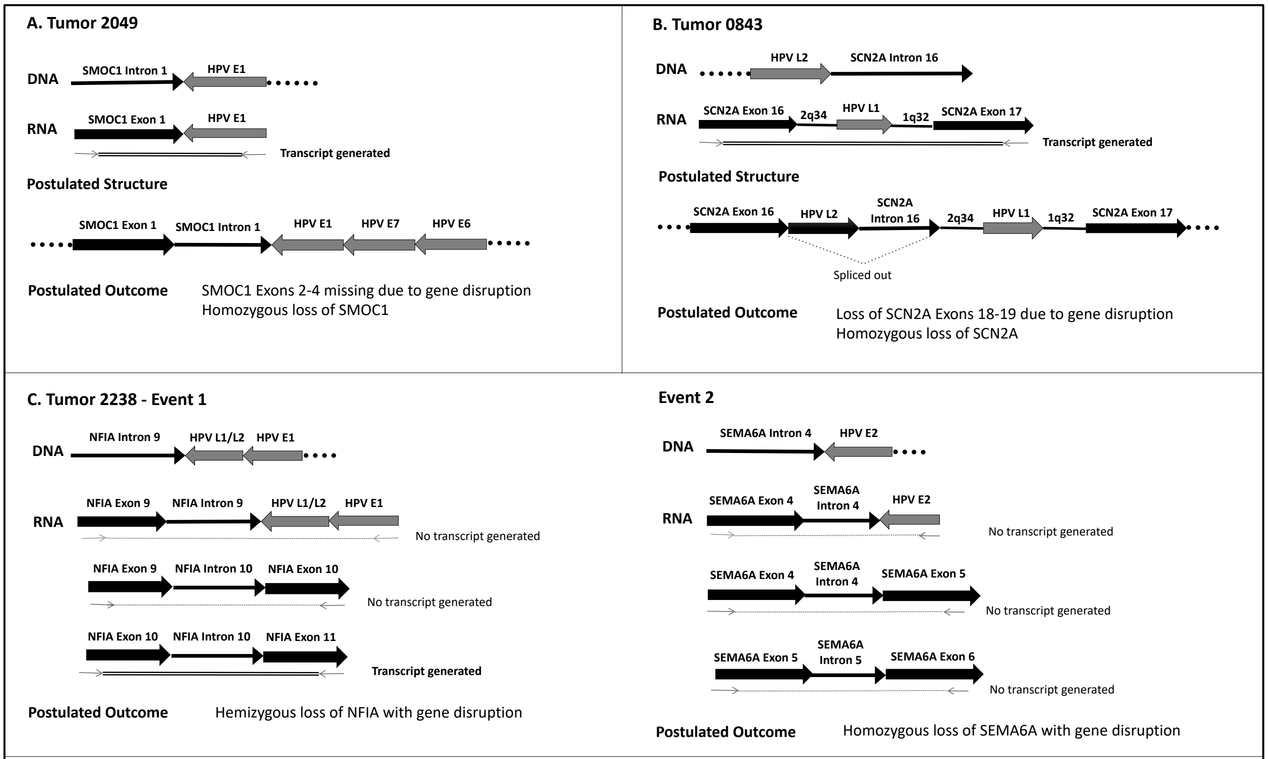


Figure 1.4: Gene disruption by integration seen in OPSCC tumors. For each event, integration is shown at DNA level and RNA transcript level, with a postulated structure of the full integration site. A) Tumor 2049 showing integration into *SMOC1* B) Tumor 0843 showing integration into *SCN2A* and intergenic loci C) Tumor 2238 events 1 (integration into *NFIA*) and event 2 (*SEMA6A*).

Conclusions and Thesis Aims

There is evidence that HPV integration is implicated in oral/oropharyngeal cancer oncogenesis, but its exact role remains largely unknown. A variety of mechanisms of integration, and their effects on both the viral and cellular genome, and likely outcomes are summarized in **Figure 1.5**. Integration of HPV into the host genome may lead to increased expression of viral oncoproteins, and recent data suggest that viral integration contributes to alterations in host cell gene expression and generation of viral-host fusion transcripts. It is unclear whether integration is required for oncogenesis or if it is consistently associated with a more aggressive, treatment-resistant phenotype. Our work has shown that tumors from patients with recurrent disease are more likely to exhibit integration into cancer-related cellular genes than those from patients who respond to treatment, which contain integration events primarily at intergenic sites. Therefore, the work carried out in this dissertation aims to further investigate integration-mediated alterations of the cellular genome, production of viral-host fusion transcripts, and the subsequent effects that contribute to oncogenesis and tumor progression. Additionally, we aimed to carry out a correlative study on the outcome and survival of patients based on HPV integration status and site in the hopes of establishing the feasibility of developing viral integration evaluation as a clinically relevant predictive or prognostic indicator.

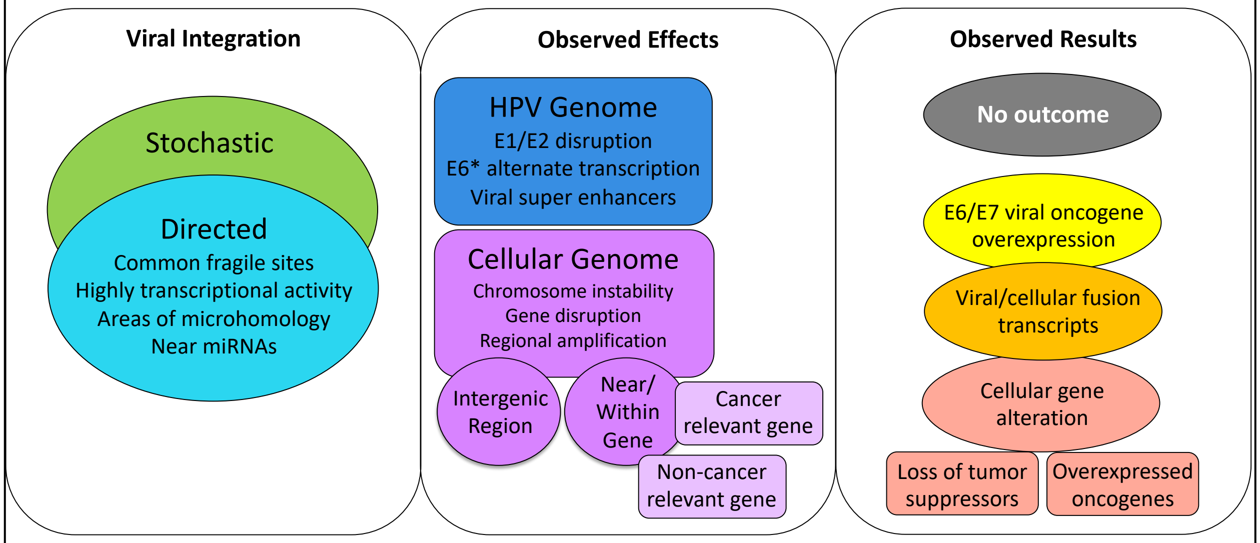


Figure 1.5. HPV integration into the human genome: effects and potential outcomes.

References

1. NIAID. *Papillomavirus Episteme*. [cited 2020; Available from: https://pave.niaid.nih.gov/.
2. Centers for Disease Control, P. *Centers for Disease Control and Prevention. Sexually Transmitted Diseases Treatment Guidelines: HPV-Associated Cancers and Precancers*. 2015 [cited 2017; Available from: https://www.cdc.gov/std/tg2015/hpv-cancer.htm.
3. Bafverstedt, B., *Condylomata acuminata--past and present.* Acta Derm Venereol, 1967. **47**(5): p. 376-81.
4. Rigoni-Stern, *Fatti statistici relative alle mallatie cacrosi che servirono de base alla poche cose dette dal dott.* Gironale service propr. pathol. trap. ser., 1842. **2**: p. 507-517.
5. Rous, P. and J.G. Kidd, *The Carcinogenic Effect of a Papilloma Virus on the Tarred Skin of Rabbits : I. Description of the Phenomenon.* J Exp Med, 1938. **67**(3): p. 399-428.
6. Rous P., B.J.W., *Carcinomatous changes in the virus-induced papillomas of the skin of the rabbit.* Proc Soc Exp Biol Med, 1934. **32**: p. 578-580.

7. Rous P., F.W.F., *The effect of chemical carcinogens on virus-induced rabbit papillomas.* J Exp Med, 1944. **79**: p. 511-537.
8. Crawford, L.V., *A study of human papilloma virus DNA.* J Mol Biol, 1965. **13**(2): p. 362-72.
9. Klug, A. and J.T. Finch, *Structure of Viruses of the Papilloma-Polyoma Type. I. Human Wart Virus.* J Mol Biol, 1965. **11**: p. 403-23.
10. Gissmann, L., et al., *Presence of human papillomavirus in genital tumors.* J Invest Dermatol, 1984. **83**(1 Suppl): p. 26s-28s.
11. Gissmann, L., et al., *Molecular cloning and characterization of human papilloma virus DNA derived from a laryngeal papilloma.* J Virol, 1982. **44**(1): p. 393-400.
12. Gissmann, L. and H. zur Hausen, *Partial characterization of viral DNA from human genital warts (Condylomata acuminata).* Int J Cancer, 1980. **25**(5): p. 605-9.
13. de Villiers, E.M., et al., *Human papillomavirus infections in women with and without abnormal cervical cytology.* Lancet, 1987. **2**(8561): p. 703-6.
14. Newell, G.R., E.T. Krementz, and J.D. Roberts, *Excess occurrence of cancer of the oral cavity, lung, and bladder following cancer of the cervix.* Cancer, 1975. **36**(6): p. 2155-8.
15. de Villiers, E.M., et al., *Papillomavirus DNA in human tongue carcinomas.* Int J Cancer, 1985. **36**(5): p. 575-8.
16. Loning, T., et al., *Analysis of oral papillomas, leukoplakias, and invasive carcinomas for human papillomavirus type related DNA.* J Invest Dermatol, 1985. **84**(5): p. 417-20.
17. Gillison, M.L., et al., *Distinct risk factor profiles for human papillomavirus type 16-positive and human papillomavirus type 16-negative head and neck cancers.* J Natl Cancer Inst, 2008. **100**(6): p. 407-20.
18. Gillison, M.L., et al., *Evidence for a causal association between human papillomavirus and a subset of head and neck cancers.* J Natl Cancer Inst, 2000. **92**(9): p. 709-20.
19. Psyrri, A. and D. DiMaio, *Human papillomavirus in cervical and head-and-neck cancer.* Nat Clin Pract Oncol, 2008. **5**(1): p. 24-31.
20. Weinberger, P.M., et al., *Molecular classification identifies a subset of human papillomavirus--associated oropharyngeal cancers with favorable prognosis.* J Clin Oncol, 2006. **24**(5): p. 736-47.
21. de Villiers, E.M., et al., *Classification of papillomaviruses.* Virology, 2004. **324**(1): p. 17-27.
22. Bernard, H.U., et al., *Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments.* Virology, 2010. **401**(1): p. 70-9.
23. Humans, I.W.G.o.t.E.o.C.R.t., *Biological agents. Volume 100 B. A review of human carcinogens.* IARC Monogr Eval Carcinog Risks Hum, 2012. **100**(Pt B): p. 1-441.
24. Bognar, L., et al., *Prognostic role of HPV infection in esophageal squamous cell carcinoma.* Infect Agent Cancer, 2018. **13**: p. 38.
25. Kamangar, F., et al., *Human papillomavirus serology and the risk of esophageal and gastric cancers: results from a cohort in a high-risk region in China.* Int J Cancer, 2006. **119**(3): p. 579-84.
26. Bodaghi, S., et al., *Colorectal papillomavirus infection in patients with colorectal cancer.* Clin Cancer Res, 2005. **11**(8): p. 2862-7.
27. Damin, D.C., et al., *Evidence for an association of human papillomavirus infection and colorectal cancer.* Eur J Surg Oncol, 2007. **33**(5): p. 569-74.

28. Jorgensen, K.R. and J.B. Jensen, *Human papillomavirus and urinary bladder cancer revisited.* APMIS, 2020.
29. Moghoofei, M., et al., *Association between human papillomavirus infection and prostate cancer: A global systematic review and meta-analysis.* Asia Pac J Clin Oncol, 2019. **15**(5): p. e59-e67.
30. Kisseljova, N., et al., *Detection of Human Papillomavirus Prevalence in Ovarian Cancer by Different Test Systems.* Intervirology, 2020: p. 1-7.
31. Roos, P., et al., *In North America, some ovarian cancers express the oncogenes of preventable human papillomavirus HPV-18.* Sci Rep, 2015. **5**: p. 8645.
32. Khodabandehlou, N., et al., *Human papilloma virus and breast cancer: the role of inflammation and viral expressed proteins.* BMC Cancer, 2019. **19**(1): p. 61.
33. Zhai, K., J. Ding, and H.Z. Shi, *HPV and lung cancer risk: a meta-analysis.* J Clin Virol, 2015. **63**: p. 84-90.
34. Syrjanen, S., *Human papillomaviruses in head and neck carcinomas.* New England Journal of Medicine, 2007. **356**(19): p. 1993-1995.
35. zur Hausen, H., *Papillomaviruses in the causation of human cancers - a brief historical account.* Virology, 2009. **384**(2): p. 260-5.
36. Walline, H.M., et al., *High-risk human papillomavirus detection in oropharyngeal, nasopharyngeal, and, oral cavity cancers: Comparison of multiple methods.* JAMA Otolaryngology, 2013.
37. Doorbar, J., *Molecular biology of human papillomavirus infection and cervical cancer.* Clin Sci (Lond), 2006. **110**(5): p. 525-41.
38. Schiffman, M., G. Clifford, and F.M. Buonaguro, *Classification of weakly carcinogenic human papillomavirus types: addressing the limits of epidemiology at the borderline.* Infect Agent Cancer, 2009. **4**: p. 8.
39. Dubina, M. and G. Goldenberg, *Viral-associated nonmelanoma skin cancers: a review.* Am J Dermatopathol, 2009. **31**(6): p. 561-73.
40. Nindl, I., M. Gottschling, and E. Stockfleth, *Human papillomaviruses and non-melanoma skin cancer: basic virology and clinical manifestations.* Dis Markers, 2007. **23**(4): p. 247-59.
41. Gerein, V., et al., *Incidence, age at onset, and potential reasons of malignant transformation in recurrent respiratory papillomatosis patients: 20 years experience.* Otolaryngol Head Neck Surg, 2005. **132**(3): p. 392-4.
42. Huebbers, C.U., et al., *Integration of HPV6 and downregulation of AKR1C3 expression mark malignant transformation in a patient with juvenile-onset laryngeal papillomatosis.* PLoS One, 2013. **8**(2): p. e57207.
43. Reidy, P.M., et al., *Integration of human papillomavirus type 11 in recurrent respiratory papilloma-associated cancer.* Laryngoscope, 2004. **114**(11): p. 1906-9.
44. Scheel, A., et al., *Human papillomavirus infection and biomarkers in sinonasal inverted papillomas: clinical significance and molecular mechanisms.* Int Forum Allergy Rhinol, 2015. **5**(8): p. 701-7.
45. Schiller, J.T., P.M. Day, and R.C. Kines, *Current understanding of the mechanism of HPV infection.* Gynecol Oncol, 2010. **118**(1 Suppl): p. S12-7.
46. Gillison, M.L., et al., *Prevalence of oral HPV infection in the United States, 2009-2010.* JAMA, 2012. **307**(7): p. 693-703.

47. Scheffner, M., et al., *The E6 Oncoprotein Encoded by Human Papillomavirus Type-16 and Type-18 Promotes the Degradation of P53.* Cell, 1990. **63**(6): p. 1129-1136.
48. Talis, A.L., J.M. Huibregtse, and P.M. Howley, *The role of E6AP in the Regulation of p53 protein levels in human papillomavirus (HPV)-positive and HPV-negative cells.* Journal of Biological Chemistry, 1998. **273**(11): p. 6439-6445.
49. Sathish, N., et al., *Human papillomavirus 16 E6/E7 transcript and E2 gene status in patients with cervical neoplasia.* Mol Diagn, 2004. **8**(1): p. 57-64.
50. Boyer, S.N., D.E. Wazer, and V. Band, *E7 protein of human papilloma virus-16 induces degradation of retinoblastoma protein through the ubiquitin-proteasome pathway.* Cancer Research, 1996. **56**(20): p. 4620-4624.
51. McBride, A.A., *The papillomavirus E2 proteins.* Virology, 2013. **445**(1-2): p. 57-79.
52. Doorbar, J., et al., *The biology and life-cycle of human papillomaviruses.* Vaccine, 2012. **30 Suppl 5**: p. F55-70.
53. Wiest, T., et al., *Involvement of intact HPV16 E6/E7 gene expression in head and neck cancers with unaltered p53 status and perturbed pRb cell cycle control.* Oncogene, 2002. **21**(10): p. 1510-1517.
54. Brouwer, A.F., et al., *Multisite HPV infections in the United States (NHANES 2003-2014): An overview and synthesis.* Prev Med, 2019. **123**: p. 288-298.
55. Walboomers, J.M., et al., *Human papillomavirus is a necessary cause of invasive cervical cancer worldwide.* J Pathol, 1999. **189**(1): p. 12-9.
56. Bosch, F.X., et al., *The causal relation between human papillomavirus and cervical cancer.* J Clin Pathol, 2002. **55**(4): p. 244-65.
57. Bray, F., et al., *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA Cancer J Clin, 2018. **68**(6): p. 394-424.
58. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer Statistics, 2017.* CA Cancer J Clin, 2017. **67**(1): p. 7-30.
59. Paz, I.B., et al., *Human papillomavirus (HPV) in head and neck cancer. An association of HPV 16 with squamous cell carcinoma of Waldeyer's tonsillar ring.* Cancer, 1997. **79**(3): p. 595-604.
60. Gillison, M.L., et al., *Tobacco smoking and increased risk of death and progression for patients with p16-positive and p16-negative oropharyngeal cancer.* J Clin Oncol, 2012. **30**(17): p. 2102-11.
61. Licitra, L., et al., *High-risk human papillomavirus affects prognosis in patients with surgically treated oropharyngeal squamous cell carcinoma.* Journal of Clinical Oncology, 2006. **24**(36): p. 5630-5636.
62. Fakhry, C., et al., *Improved survival of patients with human papillomavirus-positive head and neck squamous cell carcinoma in a prospective clinical trial.* J Natl Cancer Inst, 2008. **100**(4): p. 261-9.
63. Worden, F.P., et al., *Chemoselection as a strategy for organ preservation in advanced oropharynx cancer: response and survival positively associated with HPV16 copy number.* J Clin Oncol, 2008. **26**(19): p. 3138-46.
64. Feng, F.Y., et al., *Intensity-modulated chemoradiotherapy aiming to reduce dysphagia in patients with oropharyngeal cancer: clinical and functional results.* J Clin Oncol, 2010. **28**(16): p. 2732-8.

65. Ziemann, F., et al., *Increased sensitivity of HPV-positive head and neck cancer cell lines to x-irradiation +/- Cisplatin due to decreased expression of E6 and E7 oncoproteins and enhanced apoptosis.* American Journal of Cancer Research, 2015. **5**(3): p. 1017-1031.
66. Fakhry, C., et al., *The prognostic role of sex, race, and human papillomavirus in oropharyngeal and nonoropharyngeal head and neck squamous cell cancer.* Cancer, 2017. **123**(9): p. 1566-1575.
67. Duray, A., et al., *Human papillomavirus DNA strongly correlates with a poorer prognosis in oral cavity carcinoma.* Laryngoscope, 2012. **122**(7): p. 1558-65.
68. Ragin, C.C. and E. Taioli, *Survival of squamous cell carcinoma of the head and neck in relation to human papillomavirus infection: review and meta-analysis.* Int J Cancer, 2007. **121**(8): p. 1813-20.
69. Duray, A., et al., *Prognosis of HPV-positive head and neck cancers: implication of smoking and immunosuppression.* Adv Cell Mol Otolaryng, 2014. **2**.
70. Chaturvedi, A.K., et al., *Incidence trends for human papillomavirus-related and -unrelated oral squamous cell carcinomas in the United States.* J Clin Oncol, 2008. **26**(4): p. 612-9.
71. Brouwer, A.F., M.C. Eisenberg, and R. Meza, *Age Effects and Temporal Trends in HPV-Related and HPV-Unrelated Oral Cancer in the United States: A Multistage Carcinogenesis Modeling Analysis.* Plos One, 2016. **11**(3).
72. Mehanna, H., et al., *Prevalence of human papillomavirus in oropharyngeal and nonoropharyngeal head and neck cancer--systematic review and meta-analysis of trends by time and region.* Head Neck, 2013. **35**(5): p. 747-55.
73. Jemal, A., et al., *Annual Report to the Nation on the Status of Cancer, 1975-2009, featuring the burden and trends in human papillomavirus(HPV)-associated cancers and HPV vaccination coverage levels.* J Natl Cancer Inst, 2013. **105**(3): p. 175-201.
74. Ang, K.K., et al., *Human papillomavirus and survival of patients with oropharyngeal cancer.* N Engl J Med, 2010. **363**(1): p. 24-35.
75. Cancer Genome Atlas, N., *Comprehensive genomic characterization of head and neck squamous cell carcinomas.* Nature, 2015. **517**(7536): p. 576-82.
76. Leemans, C.R., P.J.F. Snijders, and R.H. Brakenhoff, *The molecular landscape of head and neck cancer.* Nat Rev Cancer, 2018. **18**(5): p. 269-282.
77. Chakravarthy, A., et al., *Human Papillomavirus Drives Tumor Development Throughout the Head and Neck: Improved Prognosis Is Associated With an Immune Response Largely Restricted to the Oropharynx.* J Clin Oncol, 2016. **34**(34): p. 4132-4141.
78. Zhang, Y., et al., *Subtypes of HPV-Positive Head and Neck Cancers Are Associated with HPV Characteristics, Copy Number Alterations, PIK3CA Mutation, and Pathway Signatures.* Clin Cancer Res, 2016. **22**(18): p. 4735-45.
79. Colevas, A.D., et al., *NCCN Guidelines Insights: Head and Neck Cancers, Version 1.2018.* J Natl Compr Canc Netw, 2018. **16**(5): p. 479-490.
80. Ringash, J., *Survivorship and Quality of Life in Head and Neck Cancer.* J Clin Oncol, 2015. **33**(29): p. 3322-7.
81. Bigelow, E.O., T.Y. Seiwert, and C. Fakhry, *Deintensification of treatment for human papillomavirus-related oropharyngeal cancer: Current state and future directions.* Oral Oncol, 2020. **105**: p. 104652.
82. Marur, S., et al., *E1308: Phase II Trial of Induction Chemotherapy Followed by Reduced-Dose Radiation and Weekly Cetuximab in Patients With HPV-Associated Resectable*

Squamous Cell Carcinoma of the Oropharynx- ECOG-ACRIN Cancer Research Group. J Clin Oncol, 2017. **35**(5): p. 490-497.

83. Misiukiewicz, K., et al., *Standard of care vs reduced-dose chemoradiation after induction chemotherapy in HPV+ oropharyngeal carcinoma patients: The Quarterback trial.* Oral Oncol, 2019. **95**: p. 170-177.

84. Seiwert, T.Y., et al., *OPTIMA: a phase II dose and volume de-escalation trial for human papillomavirus-positive oropharyngeal cancer.* Ann Oncol, 2019. **30**(10): p. 1673.

85. Klaes, R., et al., *Detection of high-risk cervical intraepithelial neoplasia and cervical cancer by amplification of transcripts derived from integrated papillomavirus oncogenes.* Cancer Res, 1999. **59**(24): p. 6132-6.

86. Vinokurova, S., et al., *Type-dependent integration frequency of human papillomavirus genomes in cervical lesions.* Cancer Res, 2008. **68**(1): p. 307-13.

87. Wentzensen, N., S. Vinokurova, and M. von Knebel Doeberitz, *Systematic review of genomic integration sites of human papillomavirus genomes in epithelial dysplasia and invasive cancer of the female lower genital tract.* Cancer Res, 2004. **64**(11): p. 3878-84.

88. Ziegert, C., et al., *A comprehensive analysis of HPV integration loci in anogenital lesions combining transcript and genome-based amplification techniques.* Oncogene, 2003. **22**(25): p. 3977-84.

89. Hu, Z., et al., *Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism.* Nat Genet, 2015. **47**(2): p. 158-63.

90. Ferber, M.J., et al., *Preferential integration of human papillomavirus type 18 near the c-myc locus in cervical carcinoma.* Oncogene, 2003. **22**(46): p. 7233-42.

91. Schmitz, M., et al., *Non-random integration of the HPV genome in cervical cancer.* PLoS One, 2012. **7**(6): p. e39632.

92. Tian, R., et al., *Risk stratification of cervical lesions using capture sequencing and machine learning method based on HPV and human integrated genomic profiles.* Carcinogenesis, 2019. **40**(10): p. 1220-1228.

93. Bodelon, C., et al., *Genomic characterization of viral integration sites in HPV-related cancers.* Int J Cancer, 2016. **139**(9): p. 2001-11.

94. Holmes, A., et al., *Mechanistic signatures of HPV insertions in cervical carcinomas.* NPJ Genom Med, 2016. **1**: p. 16004.

95. Akagi, K., et al., *Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability.* Genome Res, 2014. **24**(2): p. 185-99.

96. Walline, H.M., et al., *Integration of high-risk human papillomavirus into cellular cancer-related genes in head and neck cancer cell lines.* Head Neck, 2017. **39**(5): p. 840-852.

97. Parfenov, M., et al., *Characterization of HPV and host genome interactions in primary head and neck cancers.* Proc Natl Acad Sci U S A, 2014. **111**(43): p. 15544-9.

98. Nulton, T.J., et al., *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**(11): p. 17684-17699.

99. Gao, G., et al., *Whole genome sequencing reveals complexity in both HPV sequences present and HPV integrations in HPV-positive oropharyngeal squamous cell carcinomas.* BMC Cancer, 2019. **19**(1): p. 352.

100. Kalantari, M., et al., *Human papillomavirus-16 and -18 in penile carcinomas: DNA methylation, chromosomal recombination and genomic variation.* Int J Cancer, 2008. **123**(8): p. 1832-40.

101. van de Nieuwenhof, H.P., et al., *The etiologic role of HPV in vulvar squamous cell carcinoma fine tuned.* Cancer Epidemiol Biomarkers Prev, 2009. **18**(7): p. 2061-7.

102. Morel, A., et al., *Mechanistic Signatures of Human Papillomavirus Insertions in Anal Squamous Cell Carcinomas.* Cancers (Basel), 2019. **11**(12).

103. Gao, G., et al., *Common fragile sites (CFS) and extremely large CFS genes are targets for human papillomavirus integrations and chromosome rearrangements in oropharyngeal squamous cell carcinoma.* Genes Chromosomes Cancer, 2017. **56**(1): p. 59-74.

104. Myers, J.E., et al., *Detecting episomal or integrated human papillomavirus 16 DNA using an exonuclease V-qPCR-based assay.* Virology, 2019. **537**: p. 149-156.

105. Senapati, R., N.N. Senapati, and B. Dwibedi, *Molecular mechanisms of HPV mediated neoplastic progression.* Infect Agent Cancer, 2016. **11**: p. 59.

106. Groves, I.J. and N. Coleman, *Human papillomavirus genome integration in squamous carcinogenesis: what have next-generation sequencing studies taught us?* J Pathol, 2018. **245**(1): p. 9-18.

107. Jeon, S. and P.F. Lambert, *Integration of human papillomavirus type 16 DNA into the human genome leads to increased stability of E6 and E7 mRNAs: implications for cervical carcinogenesis.* Proc Natl Acad Sci U S A, 1995. **92**(5): p. 1654-8.

108. McBride, A.A. and A. Warburton, *The role of integration in oncogenic progression of HPV-associated cancers.* PLoS Pathog, 2017. **13**(4): p. e1006211.

109. Amaro-Filho, S.M., et al., *HPV DNA methylation at the early promoter and E1/E2 integrity: A comparison between HPV16, HPV18 and HPV45 in cervical cancer.* Papillomavirus Res, 2018. **5**: p. 172-179.

110. Pokryvkova, B., et al., *Detailed Characteristics of Tonsillar Tumors with Extrachromosomal or Integrated Form of Human Papillomavirus.* Viruses, 2019. **12**(1).

111. Sakakibara, N., R. Mitra, and A.A. McBride, *The papillomavirus E1 helicase activates a cellular DNA damage response in viral replication foci.* J Virol, 2011. **85**(17): p. 8981-95.

112. Liu, L., et al., *Identification of reliable biomarkers of human papillomavirus 16 methylation in cervical lesions based on integration status using high-resolution melting analysis.* Clin Epigenetics, 2018. **10**: p. 10.

113. Brant, A.C., et al., *Characterization of HPV integration, viral gene expression and E6E7 alternative transcripts by RNA-Seq: A descriptive study in invasive cervical cancer.* Genomics, 2019. **111**(6): p. 1853-1861.

114. Ehrig, F., et al., *Differences in Stability of Viral and Viral-Cellular Fusion Transcripts in HPV-Induced Cervical Cancers.* Int J Mol Sci, 2019. **21**(1).

115. Dooley, K.E., A. Warburton, and A.A. McBride, *Tandemly Integrated HPV16 Can Form a Brd4-Dependent Super-Enhancer-Like Element That Drives Transcription of Viral Oncogenes.* MBio, 2016. **7**(5).

116. Olthof, N.C., et al., *Comprehensive analysis of HPV16 integration in OSCC reveals no significant impact of physical status on viral oncogene and virally disrupted human gene expression.* PLoS One, 2014. **9**(2): p. e88718.

117. Safe, S., et al., *Minireview: role of orphan nuclear receptors in cancer and potential as drug targets.* Mol Endocrinol, 2014. **28**(2): p. 157-72.

118. Huebbers, C.U., et al., *Upregulation of AKR1C1 and AKR1C3 expression in OPSCC with integrated HPV16 and HPV-negative tumors is an indicator of poor prognosis.* Int J Cancer, 2019. **144**(10): p. 2465-2477.

119. Koneva, L.A., et al., *HPV Integration in HNSCC Correlates with Survival Outcomes, Immune Response Signatures, and Candidate Drivers.* Mol Cancer Res, 2017.

120. Wald, A.I., et al., *Alteration of microRNA profiles in squamous cell carcinoma of the head and neck cell lines by human papillomavirus.* Head Neck, 2011. **33**(4): p. 504-12.

121. Hui, A.B., et al., *Potentially prognostic miRNAs in HPV-associated oropharyngeal carcinoma.* Clin Cancer Res, 2013. **19**(8): p. 2154-62.

122. Lajer, C.B., et al., *The role of miRNAs in human papilloma virus (HPV)-associated cancers: bridging between HPV-related head and neck cancer and cervical cancer.* Br J Cancer, 2012. **106**(9): p. 1526-34.

123. Shin, H.J., et al., *Physical status of human papillomavirus integration in cervical cancer is associated with treatment outcome of the patients treated with radiotherapy.* PLoS One, 2014. **9**(1): p. e78995.

124. Ibragimova, M., et al., *HPV status and its genomic integration affect survival of patients with cervical cancer.* Neoplasma, 2018. **65**(3): p. 441-448.

125. Kiseleva, V.I., et al., *The Presence of Human Papillomavirus DNA Integration is Associated with Poor Clinical Results in Patients with Third-Stage Cervical Cancer.* Bull Exp Biol Med, 2019. **168**(1): p. 87-91.

126. Lim, M.Y., et al., *Human papillomavirus integration pattern and demographic, clinical, and survival characteristics of patients with oropharyngeal squamous cell carcinoma.* Head Neck, 2016. **38**(8): p. 1139-44.

127. Anayannis, N.V., et al., *Association of an intact E2 gene with higher HPV viral load, higher viral oncogene expression, and improved clinical outcome in HPV16 positive head and neck squamous cell carcinoma.* PLoS One, 2018. **13**(2): p. e0191581.

128. Nulton, T.J., et al., *Patients with integrated HPV16 in head and neck cancer show poor survival.* Oral Oncol, 2018. **80**: p. 52-55.

129. Vojtechova, Z., et al., *Analysis of the integration of human papillomaviruses in head and neck tumours in relation to patients' prognosis.* Int J Cancer, 2016. **138**(2): p. 386-95.

130. Walline, H.M., et al., *Genomic Integration of High-Risk HPV Alters Gene Expression in Oropharyngeal Squamous Cell Carcinoma.* Mol Cancer Res, 2016. **14**(10): p. 941-952.

Chapter 2 HPV Genomic Integration and Survival of HNSCC Patients

Abstract

The molecular drivers of human papillomavirus-related head and neck squamous cell carcinoma (HPV+HNSCC) are not entirely understood. This study evaluated the relationship between HPV integration, expression of E6/E7, and patient outcomes in p16+ HNSCCs. HPV type was determined by HPV PCR-MassArray, and integration was called using Detection of Integrated Papillomavirus Sequences (DIPS) polymerase chain reaction (PCR). We investigated whether fusion transcripts were produced by reverse transcription polymerase chain reaction (RT-PCR). E6/E7 expression was assessed by quantitative reverse transcription polymerase chain reaction (qRT-PCR). We assessed if there was a relationship between integration and E6/E7 expression, clinical variables, or patient outcomes. Most samples demonstrated HPV integration, which sometimes resulted in a fusion transcript. HPV integration was positively correlated with age at diagnosis and E6/E7 expression. There was a significant difference in survival between patients with versus without integration. Contrary to previous reports, HPV integration was associated with improved patient survival. Therefore, HPV integration may act as a molecular marker of good prognosis.

Introduction

Human papillomavirus (HPV)-induced head and neck squamous cell carcinoma (HPV+HNSCC) represents a growing public health concern due to its rapidly increasing incidence worldwide. The incidence rate of HPV+HNSCC in the United States is 4.62 per 100,000 persons

[1]. This cancer type most frequently presents in the oropharynx (HPV+OPSCC) but can also arise in other anatomic subsites of the head and neck region [2]. HPV+HNSCC is clinically distinguished from HPV-negative HNSCC (HPV- HNSCC) by p16 status, which acts as a surrogate immunohistochemical marker for HPV positivity. Currently, HPV+ and HPV- HNSCCs are treated in a similar manner, but HPV+ patients have a significantly better outcome [3, 4]. Despite this improved outcome, still 20-30% of these patients recur or fail to respond to initial therapies [5]. Therefore, it is essential to understand the molecular drivers of this disease to help identify patients at high-risk of recurrence and to develop alternate therapy regimens.

The process of HPV integration into the human genome is of particular interest as a potential driver of HPV+HNSCC. HPV has been reported to be integrated in a large proportion of cervical, head and neck, and other anogenital tumors with estimates ranging from ~50-70% [6-12]. This process has been most heavily investigated in cervical cancers, but there is a growing body of literature implicating integration as a potentially useful biomarker in head and neck cancer. It has been debated whether integration is a stochastic process that occurs randomly throughout the genome or whether it is a targeted process. Some studies have reported that integration occurs into/near genes or other genomic hotspots more frequently than expected by chance and that this can lead to functional alteration of critical genes [6, 12, 13].

In addition to altering cellular gene expression, integration has also been thought to contribute to oncogenesis by increasing HPV oncogene levels within the cell by a variety of mechanisms, including disruption of viral E2 [14]. E2 is frequently, but not always, disrupted as a result of integration, which results in increased E6/E7 due to the role of E2 as a negative transcriptional regulator [15]. Integration of HPV has also been reported to be associated with increased expression of shorter, spliced transcripts of E6 known as E6*I and E6*II [16], which have been shown to be associated with dysregulation of key cancer pathways and worse outcomes for HPV+HNSCC patients [17]. Additionally, integration into cellular genes can lead to the generation of viral-host fusion transcripts, and it has been reported that these transcripts may be more stable than episomally-derived HPV transcripts that then allows for the HPV oncogenes to persist longer [18]. Some have reported that E6/E7 levels are increased in HNSCC cell lines and tumors with integrated HPV [19, 20], but others have reported this is not necessarily true in every case [12, 21]. Therefore, the relationship between HPV integration and E6/E7 levels is not entirely clear.

Due to its impact on both viral and cellular gene expression, it has been of great interest whether integration status can be used clinically as a prognostic marker of poor outcome. A handful of studies have attempted to elucidate the relationship between HPV integration and patient outcomes with conflicting results. Some studies of integration, as measured by loss of E2, revealed that patients with integrated HPV had worse outcomes than those with episomal HPV [22-26], but others reported no significant difference between these two patient groups [27, 28]. Another group recently compared the survival of patients with and without viral-cellular fusion transcripts and found that patients with these transcripts had a significantly worse survival [29]. We recently examined the integration sites in patients who were responsive versus non-responsive to treatment and found that most responsive patients had integration into intergenic regions of the genome, whereas non-responsive patients had integrations into cellular genes [30]. This suggests that integration site may be an important factor in whether integration impacts cellular behavior leading to altered survival.

Due to this conflicting literature, we sought to clarify the relationship between E6/E7 expression and HPV integration, as well the potential impact of integration status and site on patient outcomes. Here we present an analysis of HPV types, HPV integration, and oncogene expression in thirty-six p16+ HNSCCs (**Figure 2.1**). We found that HPV integrated at a similar frequency (60%) in our cohort as previous studies, and sometimes resulted in the generation of a viral-cellular fusion transcript. There was a significant positive correlation between HPV integration status and E6/E7 expression level, and contrary to what others have reported, we found that patients with tumors containing HPV integration had a significantly improved disease-specific survival (DSS).

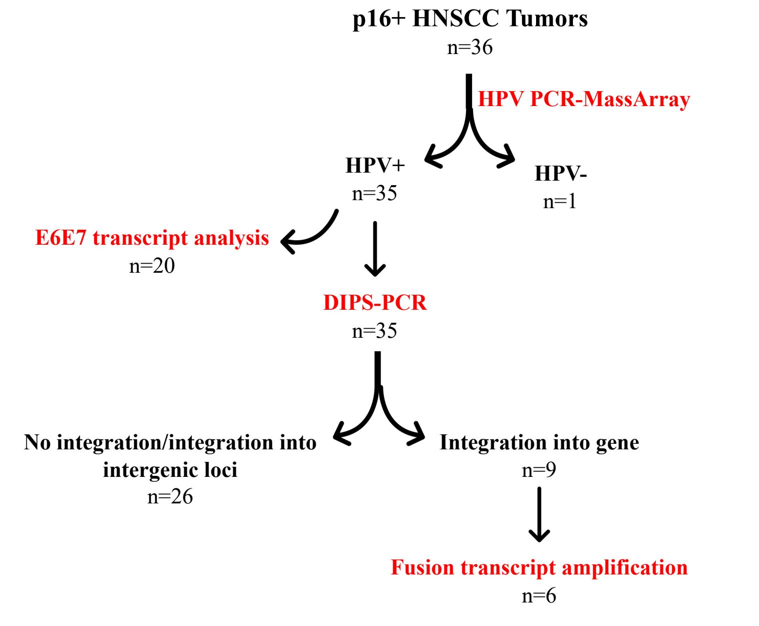


Figure 2.1. Analysis of p16+ HNSCC tumors.

Materials and Methods

Tumor Specimens: Thirty-six p16+ HNSCC tumors were obtained from the Beaumont Hospital BioBank (n=21, fresh frozen) and the Head and Neck Cancer SPORE Biorepository at the University of Michigan (n=15, formalin-fixed, paraffin embedded (FFPE) pre-treatment biopsies/surgical specimens for DNA analysis only. In four of these cases, frozen tissue was available for RNA analysis). Written informed consent to investigate their tissue was obtained from patients under studies approved by the Institutional Review Board at each institution. To reduce selection bias, p16+ HNSCC samples were acquired consecutively.

DNA/RNA Isolation: Tumor tissue was identified by a head and neck pathologist and was subsequently microdissected from 10µm sections of FFPE tissue blocks from the University of Michigan. Following microdissection, DNA was extracted from the tissue using the NucleoSpin DNA FFPE kit (Macherey-Nagel, Duren, Germany) according to the manufacturer's protocol. Briefly, paraffin was dissolved with xylene, and the tissue was lysed with lysis buffer and Proteinase K overnight at 56° C. Following overnight digestion, DNA was de-crosslinked, loaded onto the NucleoSpin DNA columns, washed and then eluted in water. DNA concentration was measured using the QUBIT 2.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA).

RNA was extracted using the RNeasy Mini Kit (Qiagen, Hilden, Germany); RNA isolation was only performed from samples with fresh frozen tissue (n=20). RNA concentration was measured using the QUBIT 2.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA). cDNA was prepared from the resulting RNA using SuperScript III (Thermo Fisher Scientific, Waltham, MA, USA).

Viral Testing: HPV PCR-MassArray was performed as previously described **(Figure 2.2)** [31]. In brief, this method detects and identifies fifteen high-risk HPV subtypes (16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66, 68, and 73), two low-risk subtypes (6 and 11), and HPV90, considered to be a possible high-risk subtype. The test included interrogation of human GAPDH as a control for sample DNA quality and assay validity. Type-specific, multiplex, competitive PCR was performed to amplify the E6 region of HPV, followed by probe-specific single base extension to discriminate between naturally occurring HPV present in the sample and the synthetic competitors included in the reaction. Matrix-assisted laser desorption/ionization time of flight mass spectroscopy was used for separation of products on a matrix-loaded silicon chip array. Samples were run in quadruplicate with appropriate positive and negative controls.

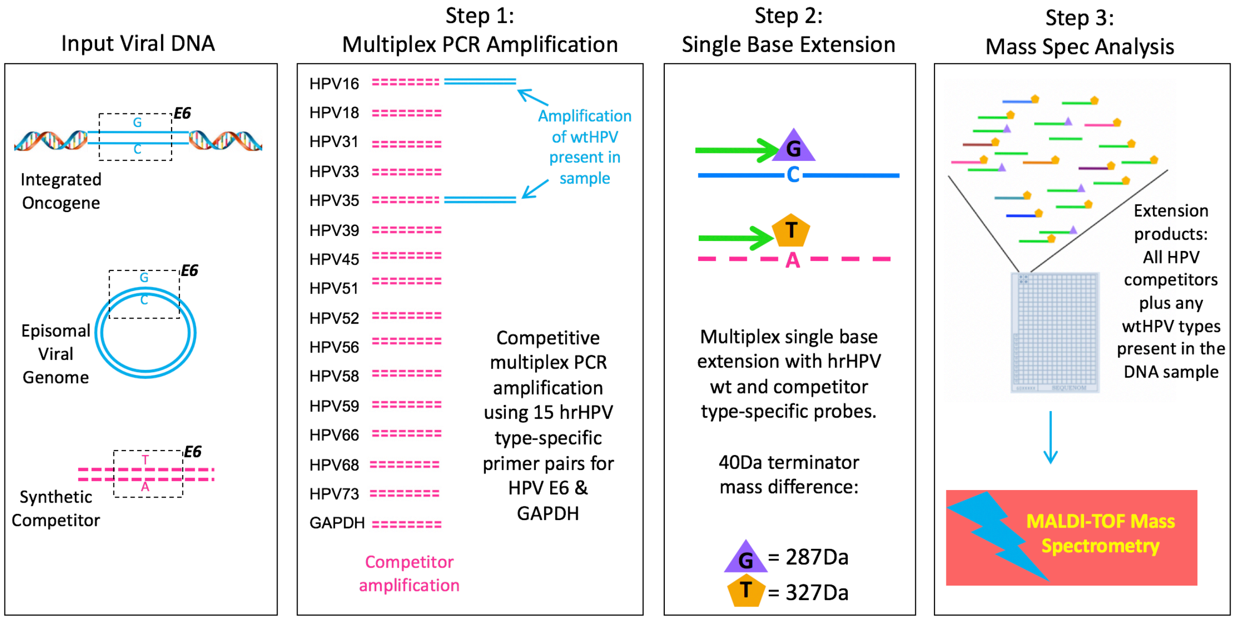


Figure 2.2. HPV PCR-MassArray Method. Type-specific, multiplex, competitive PCR is performed to amplify the E6 region of 15 hrHPV types, followed by probe-specific single base extension to discriminate between naturally occurring HPV and synthetic competitors. These are then separated by matrix-assisted laser desorption/ionization time of flight (MALDI-TOF) mass spectrometry. Figure courtesy of Dr. Heather Walline.

Detection of Integrated Papillomavirus Sequences (DIPS-PCR): DIPS-PCR was performed to identify the sites of HPV integration in the genome of the tumors, as previously described **(Figure 2.3)** [32]. For each tumor, 0.75μg DNA was digested with one of two restriction enzymes, either TaqA1 or Sau3AI (New England Biolabs, Ipswich, MA, USA). Adapters complementary to the unique overhangs created by restriction digestion were annealed to digested DNA. Linear PCR was performed on each sample using multiple viral primers to amplify viral fragments. Following linear PCR, exponential PCR using nested viral primers and an adapter-specific primer was performed. All DIPS-PCR primer sequences are listed in **Table S2.1**. Products of the exponential PCR reactions were separated by gel electrophoresis (3% agarose gel). Bands were excised from the gel and were purified using the Qiaquick Gel Extraction Kit (Qiagen, Hilden, Germany). Sanger sequencing of the isolated products was performed by the University of Michigan Advanced Genomics Core, and the results were mapped to the human and HPV genomes using NCBI-BLAST.

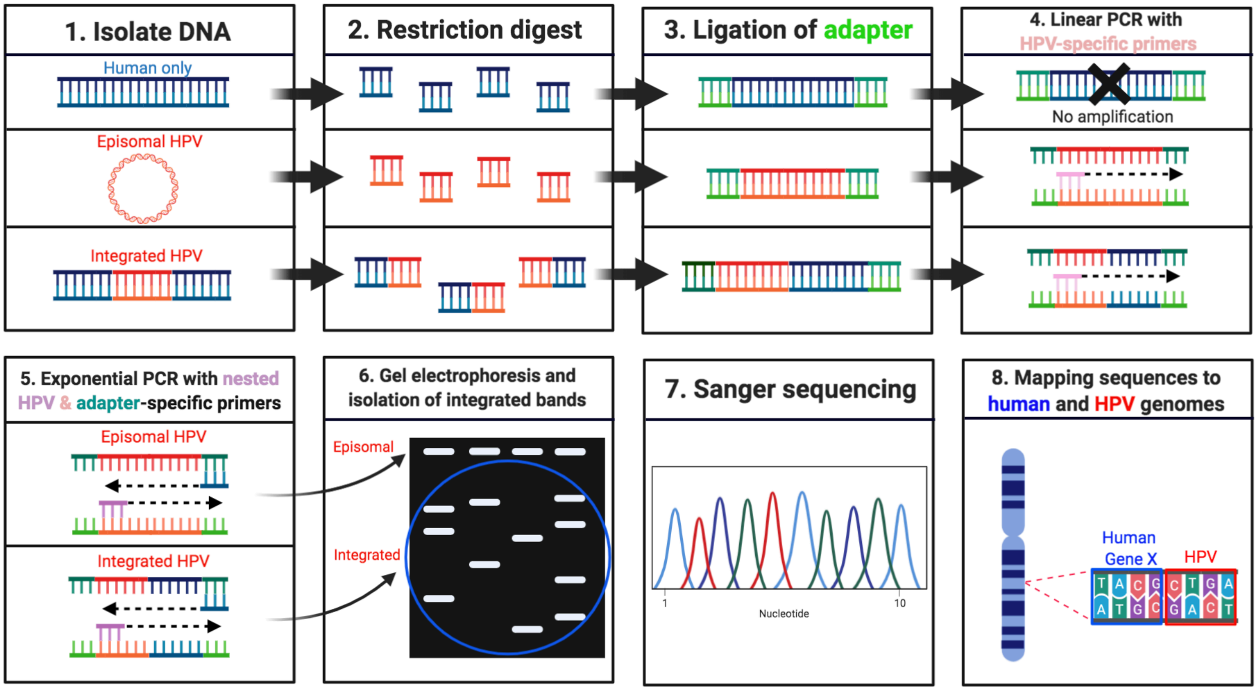


Figure 2.3. DIPS-PCR method. DNA is isolated and digested, potentially generating three types of DNA fragments: human only (top), episomal HPV only (middle) or human+HPV (bottom). An adapter is ligated to the ends of each fragment, followed by two PCR reactions with HPV-specific primers, which eliminates all human only fragments. Products containing junctions between the human and HPV genomes are separated from episomal HPV fragments by size, sequenced and then mapped.

Integration Site Transcript Analysis: RT-PCR assays were designed to amplify predicted viral-cellular transcripts in cases where RNA was available and integration took place within a cellular gene (n=6). The designed primers are listed in **Table S2.2**. All successfully amplified transcripts were sequenced for verification.

Viral Transcript Analysis: Samples with RNA available (n=20) were tested for HPV E6 and E7 transcripts by both quantitative RT-PCR (qRT-PCR) and RT-PCR. qRT-PCR was performed using QuantiTect SYBR Green PCR kit (Qiagen, Hilden, Germany) with GAPDH as an endogenous control. Relative gene expression was calculated using the ΔΔCt method compared to UM-SCC-47 (E6 and E7 expression in UM-SCC-47 were each set to 1). RT-PCR was performed using primers spanning the entire HPV16, HPV18 and/or HPV33 E6E7 region as appropriate; products were separated by gel electrophoresis (1.5% agarose gel). Primer sequences are listed in **Table S2.3**.

Statistical Analysis: Censored Kaplan Meier curves were generated using GraphPad Prism 8; survival curves were compared using log-rank testing (Mantel-Cox). Associations between integration status and clinical variables were analyzed by Spearman's rank correlation testing. P values of 0.05 or lower were considered significant.

Results

Clinical Summary

Two cohorts of p16+ HNSCC patients were analyzed from either Beaumont Hospital (n=21) or Michigan Medicine (n=15). The patients from Beaumont Hospital were collected as part of a retrospective study; patients were diagnosed between 2005-2012. Patients from Michigan Medicine were collected prospectively and were recently diagnosed (2015 onward). Tumor information, patient sex, age, smoking history, year of diagnosis, treatment, and outcomes are summarized in **Tables 2.1** and **S2.4**. We included thirty-four oropharyngeal SCCs, as well as one SCC from the oral cavity and one from the nasopharynx. As expected, there was a higher proportion of males included in this study (79% males, 21% females). Age at diagnosis ranged from 46 to 87 with an average age of 63. The majority of patients were at one time regular smokers (45% former smokers and 15% current smokers) with an average of 22 pack years. The remaining 40% of patients identified as never smokers. Only a small number of patients had history of heavy

Variable	HNSCC Patients n=33*
Av. Age at Diagnosis	62.9 [46-87]
Sex	
Male	26 (79%)
Female	7 (21%)
Smoking Status	
Current	5 (15%)
Former	15 (45%)
Never	13 (40%)
Av. Pack Years	22 [0-100]
Drinking History	
Never	14 (42%)
Social	5 (15%)
Light	8 (24%)
Heavy	6 (18%)
Tumor Site	
Oropharynx	31 (94%)
Oral Cavity	1 (3%)
Nasopharynx	1 (3%)
T Classification†	
T1	5 (15%)
T2	12 (36%)
T3	7 (21%)
T4	7 (21%)
Recurrence	2 (6%)
Treatment	
CRT	22 (67%)
CRT + Immunotherapy	1 (3%)
RT	2 (6%)
Surgery	4 (12%)
Surgery + RT	2 (6%)
Surgery + CRT	2 (6%)
Disease Progression	
No LRF or DM	22 (67%)
LRF and DM	3 (9%)
LRF only	4 (12%)
DM only	3 (9%)
Unknown	1 (3%)
Survival	
Alive, NED	21 (64%)
Died of disease	9 (27%)
Died, unrelated cause	3 (9%)

Table 2.1. Clinical information summary. *Excludes 3 patients (n=1, HPV-negative. n=2, data unavailable). †AJCC 7th edition.

Abbreviations: CRT, chemoradiation. RT, radiation therapy. LRF, locoregional failure. DM, distant metastasis. NED, no evidence of disease.

alcohol use (18%) defined as 8 or more drinks per week for females or 15 or more drinks per week for males; most patients identified as either never, light, or social drinkers.

Patients presented with tumors across the TNM classifications (AJCC 7th edition). The most frequently reported T classification was T2 (36%), but there were patients with T1, T3, and T4 tumors as well. The majority of patients (71%) had some level of nodal involvement (26% N1, 3% N2, 23% N2b, 19% N2c). Only one patient had distant metastasis at diagnosis. The majority of patients were treated with chemoradiation alone or in combination with surgery (73%). A variety of chemotherapy agents were used, including erbitux, cisplatin, carboplatin, taxol, fluorouracil, docetaxel, and gemcitabine. Other treatments included surgery alone (12%), radiation alone (6%), and surgery plus radiation (6%). Patients who developed local recurrences or metastases were treated initially with chemoradiation, followed by different chemotherapy regimens or immunotherapy.

We were able to collect at least two years of follow-up on the majority of this cohort with a median follow-up time of 3.25 years; four patients were lost to follow-up before the two-year mark. Only three patients (9%) developed both locoregional failure (LRF) and distant metastases (DM); four patients developed only LRF (12%), and three patients developed only DM (9%). Nine patients (27%) died of their disease; the average time to death was 1.5 years with a range of 3 months to 3.2 years. The majority of patients who died of disease did so within 2 years of diagnosis. The 3-year disease-specific survival (DSS) of the OPSCC patients was 80% and did not differ significantly from the non-oropharyngeal patients **(Figure 2.4A)**. We compared the survival curves of patients who developed LRF and/or DM versus those who didn't, and as expected, patients whose tumors progressed had a significant worse DSS (**Figure 2.4B**). We also examined the influence of age, smoking and drinking histories, and T and N classification, but none of these variables showed significant differences in survival (**Figure S2.1**).

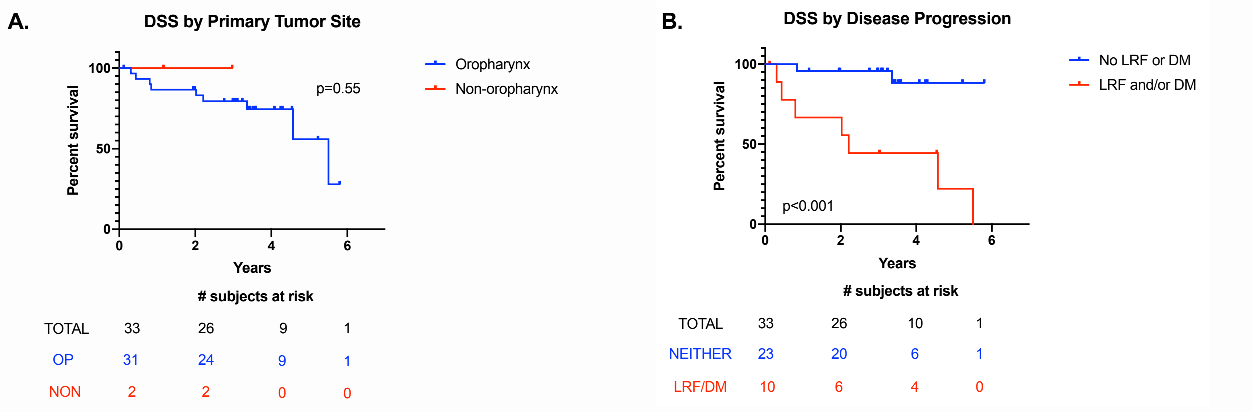


Figure 2.4. Kaplan Meier censored disease-specific survival (DSS) curves. A. Separated by primary tumor site (oropharynx vs non-oropharynx) B. Separated by disease progression (patients with vs without locoregional failure (LRF) and/or distant metastases (DM), includes both oropharynx and non-oropharynx patients.

Viral Genotypes

We tested the HPV genotypes present in thirty-six p16+ HNSCCs by HPV PCR-MassArray (**Table 2.2**). The majority of samples were positive for a single HPV type; thirty samples (83%) were HPV16+ and one sample (3%) was HPV18+. Four additional samples were positive for multiple HPV types; three samples were HPV16+ HPV33+ (8%) and one sample was HPV16+ HPV18+ (3%). Only one sample (3%) was negative for all HPV types and was excluded from further analysis.

HPV Result	**No. of patients (%) by HPV type**
HPV16	30 (83%)
HPV16 + HPV33	3 (8%)
HPV16 + HPV18	1 (3%)
HPV18	1 (3%)
Negative	1 (3%)
TOTAL	**36**

Table 2.2. HPV PCR-MassArray results.

Viral Integration

We tested thirty-five samples for HPV16 and/or HPV18 viral integration as appropriate by DIPS-PCR. We discovered at least one integration site in the majority of samples (60%) but were unable to find any integration sites in fourteen out of thirty-five samples (40%). Interestingly, the sample that was positive for both HPV16 and HPV18 (UM-3898) showed integration of both HPV types into different loci. Of the twenty-one samples with HPV integration, the median number of sites we discovered in each was 1, ranging from 1 to 4.

By Sanger Sequencing, we were able to determine that the vast majority of cellular loci affected by integration were gene-poor intergenic regions of the genome; we discovered a total of thirty-five integration sites and only eight of them involved cellular genes (**Table 2.3**). However, given that the majority of the genome does not consist of coding genes, these findings indicate integration occurs into genes more often than expected by random chance.

Sample ID	HPV Type	HPV Integration Status	HPV/Human Region(s) Involved
BMT-396	Negative	-	-
BMT-8	16	N	-
BMT-56	16	N	-
BMT-280	16+33	N	-
BMT-403	16	N	-
BMT-412	16	N	-
BMT-700	16+33	N	-
BMT-1327	16	N	-
UM-3884	18	N	-
UM-3917	16	N	-
UM-3955	16	N	-
UM-3962	16	N	-
UM-3989	16	N	-
UM-4028	16	N	-
UM-4093	16	N	-
BMT-233	16	I	E1: SCAF
			E1: SCAF
			E2: SCAF
			L1: SCAF
BMT-319	16	I	E1: SCAF
			L1: Chrom 13
BMT-322	16	I	E1: SCAF
BMT-344	16	I	E1: SCAF
BMT-400	16	I	E1: SCAF
BMT-402	16	I	E2: SCAF
			L2: SCAF
			L1: SCAF
BMT-404	16+33	I	E2: Chrom 4
BMT-411	16	I	E1: SCAF
BMT-427	16	I	E1: SCAF
UM-3940	16	I	E2: Chrom 17q21
UM-3948	16	I	L1: Chrom 13q14
UM-4067	16	I	L1: Chrom 13q14
BMT-251	16	I+G	E1: SCAF
			E2: SCAF
			L2: *SGCZ*
BMT-323	16	I+G	E1: Chrom 2q
			L2: *UTP18*
			L1: Chrom 4
BMT-1159	16	I+G	E1: SCAF
			L1: *KIF21B*
UM-3898	16+18	I+G	L1: Chrom 13q14
			(HPV18) E1: *NDST1*
UM-3938	16	I+G	L2: *YIPF1*
			L1: Chrom 6q21
UM-3954	16	I+G	E1: Chrom 3p25
			L1: *DNAI1*-L1: *NPAS3*
BMT-331	16	G	E1: Chrom 1q21: *SCN1B*
UM-4011	16	G	E1: *PTPRN2*
UM-4068	16	G	L1: *RLN1*

Table 2.3. Integration status and site descriptions.

Abbreviations: N, no sites. I, intergenic sites. G, genic sites. SCAF, genomic scaffold region.

Of the samples with HPV integration, the majority had integration into intergenic sites only (n=12) (**Figure 2.5**). Some samples had integration into both intergenic and genic regions (n=6), and a few samples (n=3) had integration into genic regions only.

A large number of integrations occurred in unplaced genomic scaffold regions of the genome (14/35 events) (**Figure 2.6**). The most frequently affected chromosome was chromosome 13 (4/35 events).The cellular genes involved in the integration sites we found included *PTPRN2, SCN1B, YIPF1, SGCZ, DNAI1, NPAS3, UTP18, RLN1*, and *KIF21B*. Integration most frequently involved the HPV genes E1 (n=14) and L1 (n=11). A few integrations also involved E2 (n=5) and L2 (n=4).

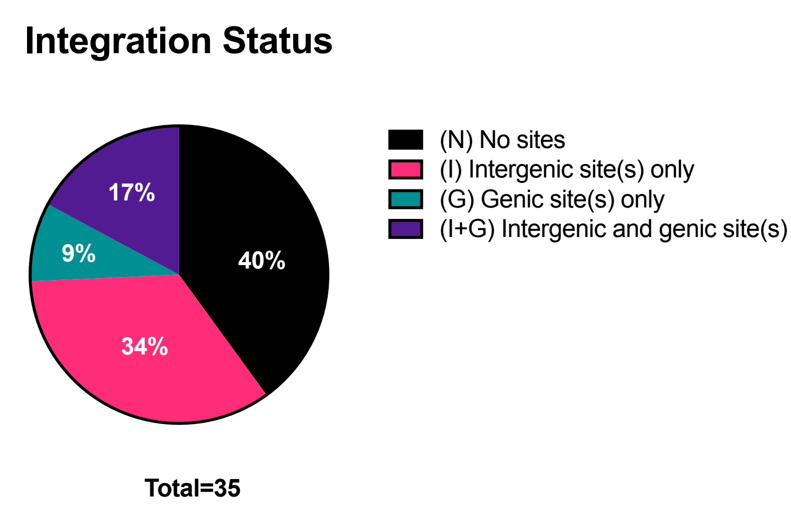


Figure 2.5. Integration status of HPV+ HNSCCs.

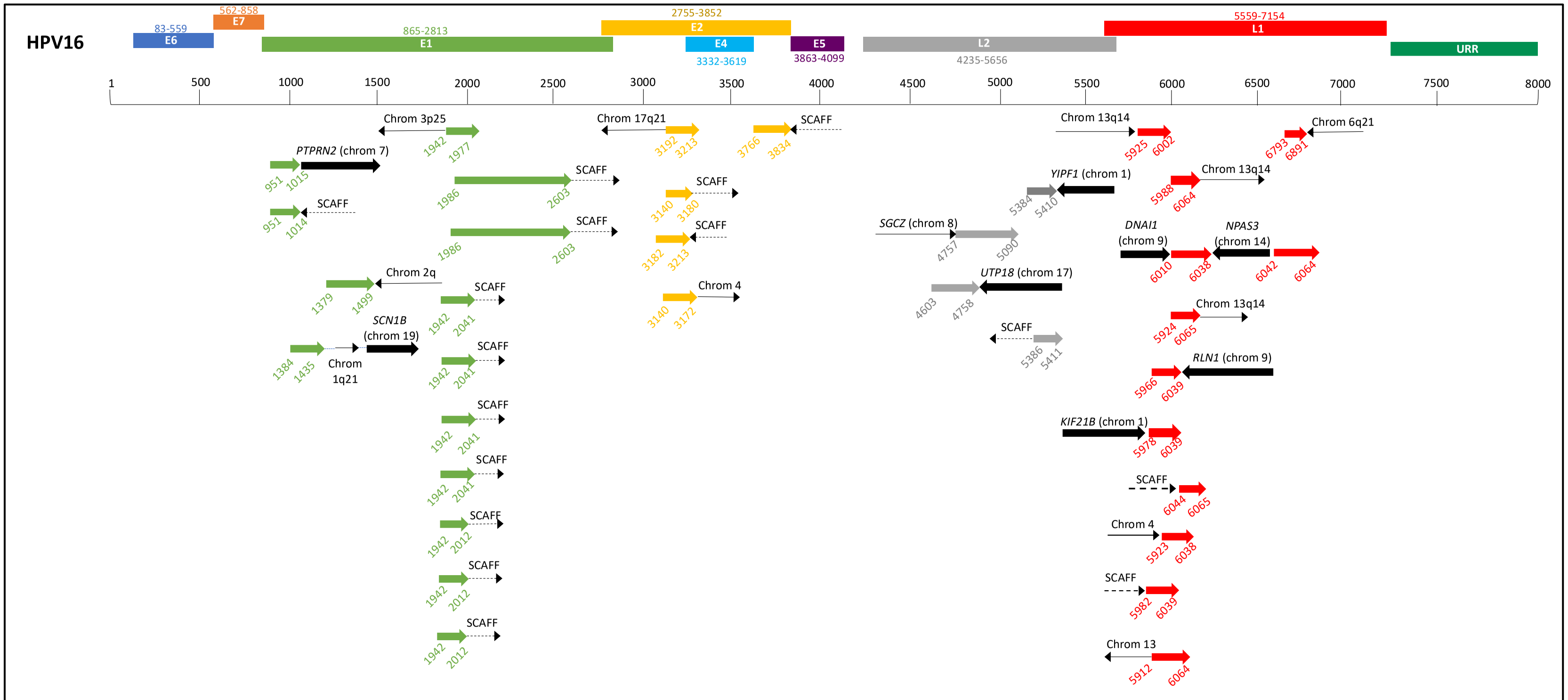


Figure 2.6. HPV integration sites aligned to HPV genome. Corresponding colors represent HPV gene (green=E1 etc). Black arrows indicate human sequence. Wide black arrow, cellular gene. Thin solid black line, intergenic region. Dashed black line labelled SCAFF, genomic scaffold region.

Viral-cellular Fusion Transcript Expression

We were interested whether those integration sites involving cellular genes led to the generation of viral-cellular fusion transcripts that have been reported in many HNSCC samples. Of the nine samples with integration into a gene, RNA was available for fusion transcript analysis for six samples. We attempted to amplify the predicted fusion transcripts with primers designed spanning the junction site discovered by DIPS-PCR (**Figure 2.7**). In BMT-1159, we detected an integration of HPV16 L1 into intron 2 of *KIF21B* by DIPS-PCR and were able to amplify a fusion transcript across this junction as shown in **Fig 2.7A**. This amplicon was sequenced by Sanger sequencing to confirm its identity, and the resulting sequence matched correctly to *KIF21B* and L1. In BMT-251, HPV16 L2 integrated into intron 1 of *SGCZ*; we attempted to amplify junctions up and downstream of L2, but no amplicons were generated. We performed similar amplifications in BMT-323 (*UTP18*:HPV16 L2), UM-3954 (*DNAI1*:HPV16 L1:*NPAS3*:HPV16 L1), UM-3898 (HPV18 E1-*NDST1*) and UM-4068 (HPV16:*RLN1*) with no amplification of any of the predicted fusion transcripts.

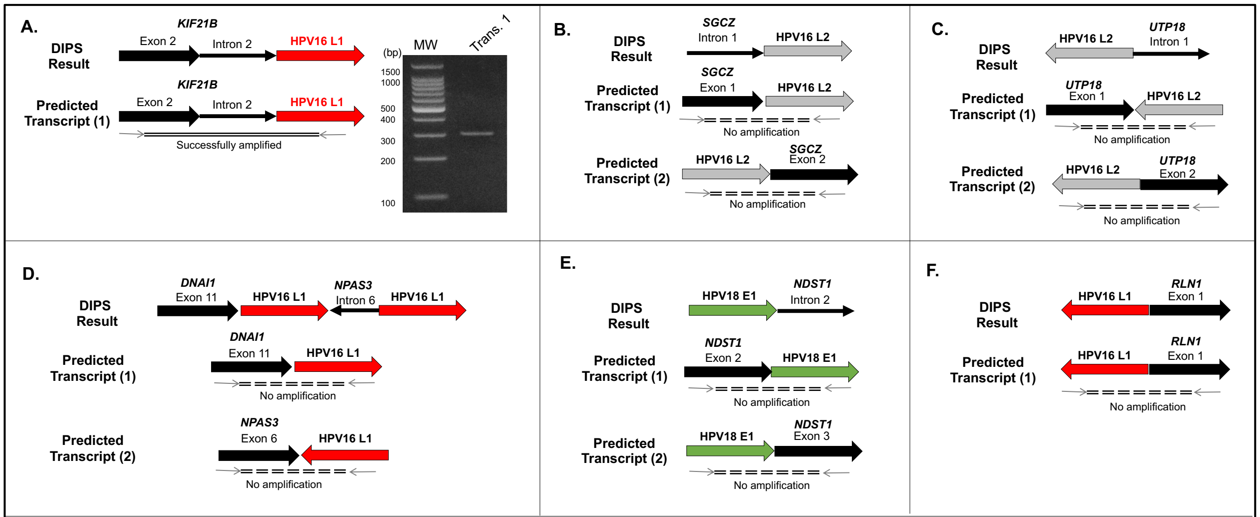
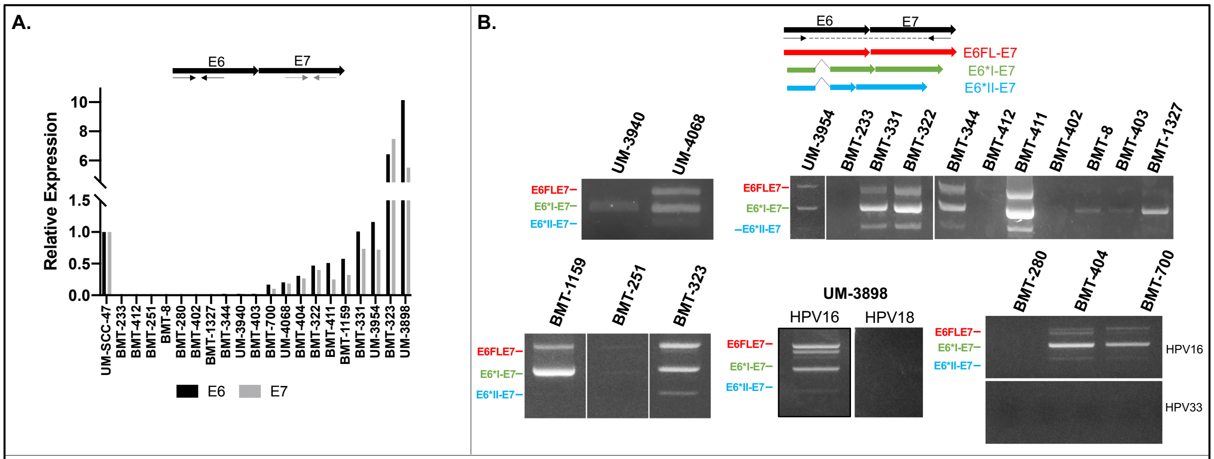


Figure 2.7. Fusion transcript amplification. For each integration event, the DIPS-PCR result (DNA level) is shown and the predicted fusion transcripts (RNA level) are shown below with primer sites indicated by small gray arrows. If the fusion transcript was successfully amplified, the gel is shown. A. BMT-1159. B. BMT-251. C. BMT-323. D. UM-3954. E. UM-3898. F. UM-4068.

Viral E6E7 Transcripts

We assessed expression of HPV E6 and E7 in samples with available RNA by qRT-PCR and RT-PCR (**Figure 2.8**). Of twenty samples tested for HPV16 by qRT-PCR, ten (50%) expressed E6 and E7 transcripts at varying levels relative to expression in UM-SCC-47 which very strongly expresses these transcripts. The remaining ten samples (50%) did not express detectable levels of HPV16 transcripts, despite testing HPV16+ at the DNA level. However, upon assessment of the expression of HPV16 E6-E7 alternate transcripts by RT-PCR, we found that five of these samples showed expression of one or more transcript. We found that the majority of samples expressed both full-length (E6FLE7) and spliced E6* transcripts (n=10), and a small number of samples (n=4) only expressed E6* transcripts. Samples positive for more than one HPV type (HPV16/18+ or HPV16/33+) were tested for transcripts of both HPV types; three samples expressed HPV16 transcripts but not HPV18 (UM-3898) or HPV33 (BMT-700 and BMT-404) transcripts. A fourth sample (BMT-280) did not express HPV16 or HPV33 transcripts. There was no significant difference in survival between patients who expressed any E6/E7 transcripts versus those who didn't, and there was also no significant difference in survival between patients who expressed only E6* transcripts versus both E6FL and E6* transcripts. (**Figure S2.2**).



	BMT-233	BMT-412	BMT-251	BMT-8	BMT-280	BMT-402	BMT-1327	BMT-344	UM-3940	BMT-403	BMT-700	UM-4068	BMT-404	BMT-322	BMT-411	BMT-1159	BMT-331	UM-3954	BMT-323	UM-3898
HPV16 E6/E7 qRT-PCR	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+
HPV16 E6FLE7	-	-	-	-	-	-	-	+	-	-	+	+	+	+	+	+	+	+	+	+
HPV16 E6*-E7	-	-	-	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+
HPV18 E6E7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	-
HPV33 E6E7	NA	NA	NA	NA	-	NA	NA	NA	NA	NA	-	NA	-	NA	NA	NA	NA	NA	NA	NA

Figure 2.8. HPV16 E6 and E7 transcript expression. A) Top: qRT-PCR primer design, Bottom: relative expression of E6 and E7, compared to UM-SCC-47. B) Top: RT-PCR primer design to amplify alternate HPV16/18/33 transcripts, Bottom: expression of alternate transcripts. C) Summary table of results. +, positive result. -, negative result. NA, not applicable.

Association with Clinical Variables

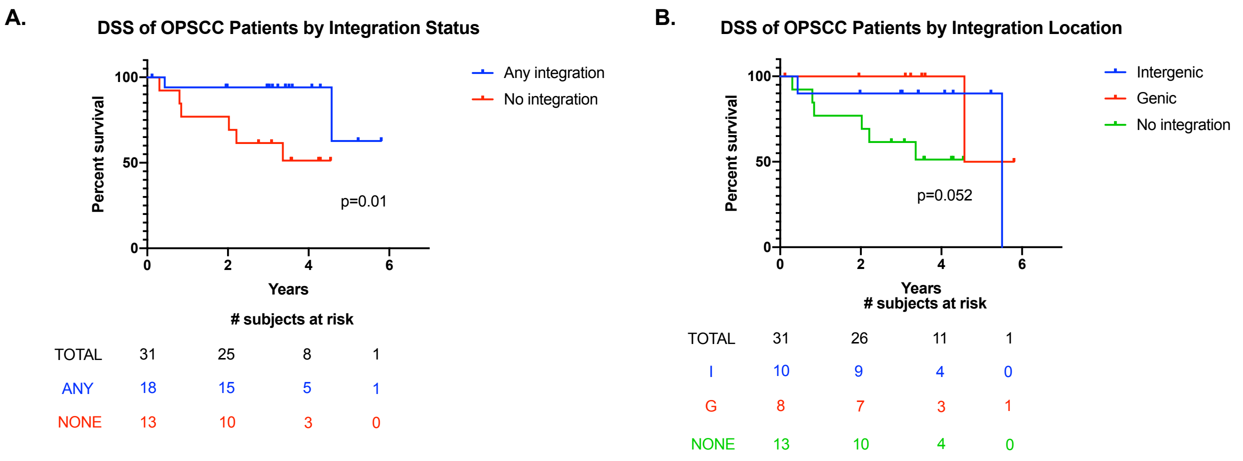
We tested whether there was an association between HPV genomic integration and other variables gathered during this study by Spearman's rank correlation (**Table 2.4**). We tested for a correlation between HPV integration and age, smoking history, drinking history, T classification, nodal involvement, E6/E7 expression level by qRT-PCR, and expression of E6FL or E6*. Of these, only age (r=0.453, p=0.008), E6/E7 expression level by qRT-PCR (r=0.480, p=0.038) and E6FL expression (r=.459, p=0.048) demonstrated a significant positive correlation

HPV Integration vs…	Spearman's r	p value
Age	0.453	0.008*
Smoking	0.112	0.537
Heavy drinking	0.219	0.220
T classification	-0.213	0.251
Nodal involvement	-0.215	0.229
E6/E7 qRT-PCR expression	0.480	0.038*
E6FLE7 expression	0.459	0.048*
E6*-E7 expression	0.186	0.447

Table 2.4. Correlation between HPV integration and other relevant variables. *Significant p-value.

with HPV integration. This indicates that patients with integration were more likely to be older and had higher expression of the HPV oncogenes, specifically the full-length E6 transcript.

We were interested in whether HPV integration influenced patient outcomes. There was no significant association between HPV integration and locoregional failure (p=0.676) or distant metastasis (p=0.659) as assessed by Fisher's exact test, although the number of events in each group was likely too small to power this analysis. The DSS curves of the oropharynx patients separated by integration status and site are shown in **Figure 2.9**. Integration positive OPSCC patients had a significantly improved DSS compared to integration negative patients (p=0.01). When we separated integration positive patients by site of integration (intergenic sites only vs any genic sites), there was no significant difference in the survival curves.



# subjects at risk

TOTAL	31	25	8	1
ANY	18	15	5	1
NONE	13	10	3	0

# subjects at risk

TOTAL	31	26	11	1
I	10	9	4	0
G	8	7	3	1
NONE	13	10	4	0

Figure 2.9. Kaplan Meier curves of oropharynx patients separated by integration status (A) and integration subsite (B), censored.

Discussion

HPV+ HNSCC, particularly HPV+OPSCC, has been increasing in incidence rapidly over the past few decades [33-35]. Despite improved outcomes compared to HPV- HNSCC, still 20-30% of patients fail to respond to initial therapies or recur [5], and the factors that contribute to the progression of this disease are not well understood. Given the high morbidity of HNSCC treatment, there is a push in the field to de-escalate treatment for patients at low risk of disease recurrence [36]. However, the biomarkers for response to treatment are not well developed yet, which makes stratifying patients difficult. Studies of treatment de-escalation are ongoing based on clinical risk factors [37-39], but there is still a need to investigate the molecular drivers of this disease in order to understand what distinguishes high versus low risk patients.

One such process that has been investigated as a potential driver of HPV+ HNSCC is the process of viral integration. Viral integration has been well characterized in cervical cancer as a marker of disease progression [40]. Studies in cervical cancer and HNSCC have shown that integration into the genome can have a variety of effects on both the cellular and viral genomes, including large scale rearrangements, amplifications, deletions, alterations in gene expression and generation of viral-cellular fusion transcripts [6-8, 11-13, 19]. Others have attempted to characterize the relationship between HPV integration and E6/E7 expression as well as between HPV integration and patient outcomes with mixed results [16, 22-31].

Here we have presented an analysis of integration sites, HPV oncogene expression and associations with clinical variables in a cohort of p16+ HNSCCs. Only one patient tested negative for all HPV types by HPV PCR-MassArray and was excluded from further analysis. Of the thirty-five patients tested for HPV16 and/or HPV18 integration by DIPS-PCR, we found at least one integration site in 60% of samples and were unable to find integration in 40%. We considered samples without HPV integration sites to be "integration-negative", although it is theoretically possible sites of integration were missed by DIPS-PCR. However, previous studies of HPV integration using a variety of methods reported similar proportions, ranging from 30-50% integration negative [6-12]. The use of different HPV integration detection methods likely accounts for the variability seen between studies.

The use of DIPS-PCR allows us to identify the number and location of HPV integration sites within each sample. The majority of samples contained only one integration site, although there were samples in which we were able to identify more than one. Of particular interest was UM-3898, which contained integrations for both HPV16 and HPV18; it is unclear how integration of more than one HPV type might affect the progression of tumorigenesis. E1 was the HPV gene most frequently involved in integration (40% of sites), which is in agreement with previous studies [12, 41]. Even though there were a limited number of integration sites detected (n=35), we were able to determine that integration events took place across eleven different chromosomes (chromosomes 1, 2, 3, 4, 6, 7, 8, 9, 13, 14, 17). Of the integration sites detected, only eight (23%) were within cellular genes. Previous studies have proposed that integration is a directed process that occurs preferentially in/near genes or other genomic features, such as miRNAs or lncRNAs [6, 13, 40, 42], but our results show more of a stochastic pattern given the wide range of chromosomes affected and low percentage involving genes. However, the number of events we detected is relatively small, and therefore it is challenging to detect predilections for a specific type of location or chromosomal hotspots. Furthermore, the limiting size of the genomic segments in the SCAF insertions detected by this method prohibits precise identification of the actual locus affected.

We further investigated the integration sites that occurred within cellular genes at the transcript level. Viral-host fusion transcripts have been reported by other groups to increase E6/E7 expression [18-20]. Previous work from our group has shown that viral-cellular fusion transcripts may or may not form depending on the location of the integration site within the gene (within an intron vs exon) [20, 30, 43]. It is possible that some integrations within introns are spliced out and therefore do not produce a fusion transcript, while others may alter splice acceptor/donor sites such that they are retained at the transcript level. We attempted to amplify the predicted fusion transcripts based on the DNA-level information we obtained from DIPS-PCR in six samples but were only successful in amplifying the fusion in one sample (BMT-1159). This fusion involved HPV16 L1 integrating into intron 2 of the cellular gene *KIF21B*, which encodes for a microtubule-dependent motor protein. In this case, we were able to amplify a transcript that included *KIF21B* exon 2–*KIF21B* intron 2–HPV16 L1, indicating this integration resulted in alteration of splice sites such that intron 2 was retained in the transcript. *KIF21B* and other kinesin superfamily proteins have been implicated in the progression of many solid tumors via dysregulation of mitosis [44-46]; therefore, it is of great interest to discover how this fusion may have played a role in the carcinogenesis in this case.

We performed a similar analysis on the other five samples, three of which involved integration into introns and two involved gene exons, but we were unable to amplify any of the predicted fusion transcripts. It is not necessarily surprising that these integration sites did not yield fusion transcripts, but it is possible that the site is more complicated than we expect, resulting in a false negative. Another open question is whether these fusion transcripts are being driven off of a cellular or HPV promoter, which is difficult to address with the relatively short sequences obtained during DIPS-PCR. Gathering more sequence surrounding the site may be helpful in the future to amplify these transcripts.

We also assessed expression of the E6 and E7 oncogenes within tumors with available RNA (n=20) by qRT-PCR, which showed varying levels of expression compared to UM-SCC-47, an HPV+ HNSCC cell line we showed previously has high E6/E7 expression [20]. Interestingly, half of the samples showed no expression of E6 or E7. However, analysis of these samples by RT-PCR using primers designed to amplify alternate E6E7 transcripts revealed that they did in fact express one or more E6E7 transcripts. It is unclear why they lacked expression by qRT-PCR, but it is possible they were below the threshold of detection for this assay. There were still five samples which showed no expression of E6E7, which is curious given that they were p16+ by IHC and HPV16+ at the DNA level. As a whole, the field struggles to agree on the methodology for determining “true” HPV positivity (p16 expression vs. HPV DNA vs. HPV RNA), as there is not always agreement between the methods. P16 is a useful surrogate marker, but there is an estimated discordance rate with HPV expression of 10-20% [47].

E6/E7 are negatively regulated by E2, which is frequently reported to be disrupted by the process of HPV integration; therefore, some have proposed that HPV integration leads to increased E6/E7 levels [15]. In this cohort, we saw a significant positive correlation between HPV integration and E6/E7 expression levels, which supports this idea. However, it is not a perfect correlation; some samples with HPV integration still have no expression of E6/E7. This aligns with those who have published that E2 is not always disrupted during integration, and therefore not all integrated samples will have increased E6/E7 levels [12, 21]. Alternatively, E6/E7 expression could be altered due to methylation of the E2 binding sites in the upstream regulatory region (URR) of HPV16 rather than loss of E2 itself [48, 49].

We assessed the expression of alternate E6* transcripts; these transcripts are thought to contribute to a more aggressive phenotype, resulting in larger tumors and worse patient prognosis [17]. We found that the majority of samples expressed both E6FLE7 and alternate E6* transcripts with a few samples only expressing E6* transcripts. Three out of four samples that contained multiple HPV types only expressed HPV16 transcripts but not from other HPV types. There was a significant positive correlation between HPV integration and E6FL expression, but not between HPV integration and E6* expression. This contrasts with reports that E6* variants are more common in tumors with integrated HPV [16]; however, it is possible our results differed due to our relatively small sample size.

We assessed the association of HPV integration with clinical variables, including age, smoking and drinking histories, and T/N classification, to further examine this process. Of these, only age showed a significant positive correlation with HPV integration, indicating that older patients were more likely to have integrated HPV. It is unclear why this may be; one explanation could be that HPV integration occurs more frequently in older patients because DNA damage accumulates in aging tissue, as it has been previously proposed that HPV integration occurs at sites of unresolved DNA damage [50].

We compared the survival of OPSCC patients with versus without integration and found that integration-positive patients had a significantly improved disease-specific survival over integration-negative patients. This contrasts with what others have previously reported; studies either reported no significant difference between the two groups or that integration-negative patients had a survival advantage over integration-positive patients [23-29]. It has been hypothesized that integration acts as an additional oncogenic driver through its various effects on the human and viral genomes. The reason for the discrepancy between our findings and previous reports is unclear, but it could be due to different methods of detecting HPV integration. These previous studies measured integration indirectly by assessing loss of E2 DNA [22-27] or mRNA [28]. Another study based integration status on the presence of fusion transcripts [29]. However, given that E2 is not always lost due to integration and not every integration results in a fusion transcript, our preferred method to detect integration is DIPS-PCR. We have used DIPS-PCR previously to assess integration sites in a small cohort of responsive vs non-responsive patients and found that non-responsive patients were more likely to have integration into genes rather than intergenic loci [30]. The underlying mechanism behind the improved survival we reported here in integration positive patients is unclear and requires further investigation. One possible hypothesis is that the process of HPV integration generates tumor neoantigens which can then be recognized as non-self by the host immune system and enhance antitumor immune response. HPV+ OPSCC patients with higher levels of infiltrating CD8+ T cells, which are involved in recognizing tumor antigens, have been shown to have improved outcomes [51], but it is currently unknown if integration-positive vs integration negative patients have differential immune infiltration patterns and whether they can present these neoantigens for immune recognition.

There are two major limitations of this study that could be addressed in future research. First, our study population was relatively small, which limited our ability to examine the relationships between HPV integration status/site and LRF or DM given that so few patients experienced these events. Secondly, we used DIPS-PCR as our preferred method of detecting integration sites because it is highly specific, but some of the amplicons we generated were too short to provide enough context for us to be able to place them at a specific locus and therefore had to be denoted as “genomic scaffold”. DIPS-PCR alone is also unable to distinguish between samples with only integrated HPV and samples that contain a mixture of integrated and episomal HPV, although sometimes episomal HPV copies may appear as 6-8 kb bands upon gel electrophoresis. It is unclear how these two samples types may differ in terms of HPV-related genetic or epigenetic changes. In the future, we will focus on pairing DIPS-PCR with long-range sequencing technologies, such as Nanopore sequencing, in order to better define the complex structural rearrangements caused by HPV integration [19] and explain the structural basis of local amplification at integration sites [12]. Comprehensive investigation of HPV integration sites and how they impact the course of HNSCC is necessary to provide insight for the development of alternate therapies for non-responsive tumors. Overall, this study shows that HPV integration influences patient outcomes, which we feel warrants the implementation of viral integration analysis in the clinic.

References

1. Mahal, B.A., et al., *Incidence and Demographic Burden of HPV-Associated Oropharyngeal Head and Neck Cancers in the United States.* Cancer Epidemiol Biomarkers Prev, 2019. **28**(10): p. 1660-1667.
2. Gillison, M.L., et al., *Evidence for a causal association between human papillomavirus and a subset of head and neck cancers.* J Natl Cancer Inst, 2000. **92**(9): p. 709-20.
3. Fakhry, C., et al., *Improved survival of patients with human papillomavirus-positive head and neck squamous cell carcinoma in a prospective clinical trial.* J Natl Cancer Inst, 2008. **100**(4): p. 261-9.
4. Licitra, L., et al., *High-risk human papillomavirus affects prognosis in patients with surgically treated oropharyngeal squamous cell carcinoma.* Journal of Clinical Oncology, 2006. **24**(36): p. 5630-5636.
5. Ang, K.K., et al., *Human papillomavirus and survival of patients with oropharyngeal cancer.* N Engl J Med, 2010. **363**(1): p. 24-35.
6. Bodelon, C., et al., *Genomic characterization of viral integration sites in HPV-related cancers.* Int J Cancer, 2016. **139**(9): p. 2001-11.
7. Gao, G., et al., *Whole genome sequencing reveals complexity in both HPV sequences present and HPV integrations in HPV-positive oropharyngeal squamous cell carcinomas.* BMC Cancer, 2019. **19**(1): p. 352.
8. Holmes, A., et al., *Mechanistic signatures of HPV insertions in cervical carcinomas.* NPJ Genom Med, 2016. **1**: p. 16004.
9. Kalantari, M., et al., *Human papillomavirus-16 and -18 in penile carcinomas: DNA methylation, chromosomal recombination and genomic variation.* Int J Cancer, 2008. **123**(8): p. 1832-40.
10. Morel, A., et al., *Mechanistic Signatures of Human Papillomavirus Insertions in Anal Squamous Cell Carcinomas.* Cancers (Basel), 2019. **11**(12).
11. Nulton, T.J., et al., *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**(11): p. 17684-17699.
12. Parfenov, M., et al., *Characterization of HPV and host genome interactions in primary head and neck cancers.* Proc Natl Acad Sci U S A, 2014. **111**(43): p. 15544-9.
13. Hu, Z., et al., *Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism.* Nat Genet, 2015. **47**(2): p. 158-63.
14. McBride, A.A. and A. Warburton, *The role of integration in oncogenic progression of HPV-associated cancers.* PLoS Pathog, 2017. **13**(4): p. e1006211.
15. McBride, A.A., *The papillomavirus E2 proteins.* Virology, 2013. **445**(1-2): p. 57-79.

16. Zhang, Y., et al., *Subtypes of HPV-Positive Head and Neck Cancers Are Associated with HPV Characteristics, Copy Number Alterations, PIK3CA Mutation, and Pathway Signatures.* Clin Cancer Res, 2016. **22**(18): p. 4735-45.

17. Qin, T., et al., *Significant association between host transcriptome-derived HPV oncogene E6* influence score and carcinogenic pathways, tumor size, and survival in head and neck cancer.* Head Neck, 2020. **42**(9): p. 2375-2389.

18. Jeon, S. and P.F. Lambert, *Integration of human papillomavirus type 16 DNA into the human genome leads to increased stability of E6 and E7 mRNAs: implications for cervical carcinogenesis.* Proc Natl Acad Sci U S A, 1995. **92**(5): p. 1654-8.

19. Akagi, K., et al., *Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability.* Genome Res, 2014. **24**(2): p. 185-99.

20. Walline, H.M., et al., *Integration of high-risk human papillomavirus into cellular cancer-related genes in head and neck cancer cell lines.* Head Neck, 2017. **39**(5): p. 840-852.

21. Olthof, N.C., et al., *Comprehensive analysis of HPV16 integration in OSCC reveals no significant impact of physical status on viral oncogene and virally disrupted human gene expression.* PLoS One, 2014. **9**(2): p. e88718.

22. Ibragimova, M., et al., *HPV status and its genomic integration affect survival of patients with cervical cancer.* Neoplasma, 2018. **65**(3): p. 441-448.

23. Kiseleva, V.I., et al., *The Presence of Human Papillomavirus DNA Integration is Associated with Poor Clinical Results in Patients with Third-Stage Cervical Cancer.* Bull Exp Biol Med, 2019. **168**(1): p. 87-91.

24. Shin, H.J., et al., *Physical status of human papillomavirus integration in cervical cancer is associated with treatment outcome of the patients treated with radiotherapy.* PLoS One, 2014. **9**(1): p. e78995.

25. Anayannis, N.V., et al., *Association of an intact E2 gene with higher HPV viral load, higher viral oncogene expression, and improved clinical outcome in HPV16 positive head and neck squamous cell carcinoma.* PLoS One, 2018. **13**(2): p. e0191581.

26. Nulton, T.J., et al., *Patients with integrated HPV16 in head and neck cancer show poor survival.* Oral Oncol, 2018. **80**: p. 52-55.

27. Lim, M.Y., et al., *Human papillomavirus integration pattern and demographic, clinical, and survival characteristics of patients with oropharyngeal squamous cell carcinoma.* Head Neck, 2016. **38**(8): p. 1139-44.

28. Vojtechova, Z., et al., *Analysis of the integration of human papillomaviruses in head and neck tumours in relation to patients' prognosis.* Int J Cancer, 2016. **138**(2): p. 386-95.

29. Koneva, L.A., et al., *HPV Integration in HNSCC Correlates with Survival Outcomes, Immune Response Signatures, and Candidate Drivers.* Mol Cancer Res, 2017.

30. Walline, H.M., et al., *Genomic Integration of High-Risk HPV Alters Gene Expression in Oropharyngeal Squamous Cell Carcinoma.* Mol Cancer Res, 2016. **14**(10): p. 941-952.

31. Walline, H.M., et al., *High-risk human papillomavirus detection in oropharyngeal, nasopharyngeal, and, oral cavity cancers: Comparison of multiple methods.* JAMA Otolaryngology, 2013.

32. Luft, F., et al., *Detection of integrated papillomavirus sequences by ligation-mediated PCR (DIPS-PCR) and molecular characterization in cervical cancer cells.* Int J Cancer, 2001. **92**(1): p. 9-17.

33. Lundberg, M., et al., *Increased incidence of oropharyngeal cancer and p16 expression.* Acta Otolaryngol, 2011. **131**(9): p. 1008-11.

34. Wittekindt, C., et al., *Increasing Incidence rates of Oropharyngeal Squamous Cell Carcinoma in Germany and Significance of Disease Burden Attributed to Human Papillomavirus.* Cancer Prev Res (Phila), 2019. **12**(6): p. 375-382.

35. Chaturvedi, A.K., et al., *Human papillomavirus and rising oropharyngeal cancer incidence in the United States.* J Clin Oncol, 2011. **29**(32): p. 4294-301.

36. Bigelow, E.O., T.Y. Seiwert, and C. Fakhry, *Deintensification of treatment for human papillomavirus-related oropharyngeal cancer: Current state and future directions.* Oral Oncol, 2020. **105**: p. 104652.

37. Marur, S., et al., *E1308: Phase II Trial of Induction Chemotherapy Followed by Reduced-Dose Radiation and Weekly Cetuximab in Patients With HPV-Associated Resectable Squamous Cell Carcinoma of the Oropharynx- ECOG-ACRIN Cancer Research Group.* J Clin Oncol, 2017. **35**(5): p. 490-497.

38. Misiukiewicz, K., et al., *Standard of care vs reduced-dose chemoradiation after induction chemotherapy in HPV+ oropharyngeal carcinoma patients: The Quarterback trial.* Oral Oncol, 2019. **95**: p. 170-177.

39. Seiwert, T.Y., et al., *OPTIMA: a phase II dose and volume de-escalation trial for human papillomavirus-positive oropharyngeal cancer.* Ann Oncol, 2019. **30**(10): p. 1673.

40. Tian, R., et al., *Risk stratification of cervical lesions using capture sequencing and machine learning method based on HPV and human integrated genomic profiles.* Carcinogenesis, 2019. **40**(10): p. 1220-1228.

41. Liu, L., et al., *Identification of reliable biomarkers of human papillomavirus 16 methylation in cervical lesions based on integration status using high-resolution melting analysis.* Clin Epigenetics, 2018. **10**: p. 10.

42. Schmitz, M., et al., *Non-random integration of the HPV genome in cervical cancer.* PLoS One, 2012. **7**(6): p. e39632.

43. Pinatti, L.M., et al., *Viral Integration Analysis Reveals Likely Common Clonal Origin of Bilateral HPV16-Positive, p16-Positive Tonsil Tumors.* Archives of Clinical and Medical Case Reports, 2020. **4**: p. 680-696.

44. Chen, J., et al., *Kinesin superfamily protein expression and its association with progression and prognosis in hepatocellular carcinoma.* J Cancer Res Ther, 2017. **13**(4): p. 651-659.

45. Sun, Z.G., et al., *Kinesin superfamily protein 21B acts as an oncogene in non-small cell lung cancer.* Cancer Cell Int, 2020. **20**: p. 233.

46. Zhao, H.Q., et al., *Increased KIF21B expression is a potential prognostic biomarker in hepatocellular carcinoma.* World J Gastrointest Oncol, 2020. **12**(3): p. 276-288.

47. Osborn HA, L.D., *P16 and HPV discordance in oropharyngeal squamous cell carcinoma: what are the clinical implications?* Annals of Hematology and Oncology, 2016. **3**(11): p. 1123.

48. Reuschenbach, M., et al., *Methylation status of HPV16 E2-binding sites classifies subtypes of HPV-associated oropharyngeal cancers.* Cancer, 2015. **121**(12): p. 1966-76.

49. von Knebel Doeberitz, M. and E.S. Prigge, *Role of DNA methylation in HPV associated lesions.* Papillomavirus Res, 2019. **7**: p. 180-183.

50. Wallace, N.A., et al., *High-Risk Alphapapillomavirus Oncogenes Impair the Homologous Recombination Pathway.* J Virol, 2017. **91**(20).

51. Oguejiofor, K., et al., *Stromal infiltration of CD8 T cells is associated with improved clinical outcome in HPV-positive oropharyngeal squamous carcinoma.* Br J Cancer, 2015. **113**(6): p. 886-93.

Chapter 3 Clonality of Bilateral Tonsillar Carcinomas

Abstract

With oral HPV infections currently rising at epidemic rates in the western world, high-risk human papillomaviruses (HPV) are responsible for a significant number of oropharyngeal squamous cell carcinomas (OPSCC). Synchronous bilateral HPV+ tumors of both tonsils are a very rare event whose understanding, however, could provide important insights into virus-driven tumor development and progression and whether such integration events are of clonal origin. In this study, we analyzed three cases of bilateral tonsillar p16+ HPV+OPSCC. The viral integration status of the various tumor samples was determined by integration-specific polymerase chain reaction (PCR) methods and sequencing, which identified viral insertion sites and affected host genes. Analysis of the tumors revealed common HPV types and viral integration events in two patients, but unique HPV types and viral integration sites in another patient, providing evidence that multiple mechanisms may exist for the formation of bilateral tonsillar carcinomas.

Introduction

Persistent oral infection of HPV is a risk factor for the development of OPSCC, but the rates of oral infection in the general population are relatively low and clearance of infections is common [1-3]. It is unclear why some HPV infections are cleared while others persist, but smokers, males, and individuals with higher numbers of sexual partners are more likely to have a persistent oral HPV infection [4]. The rate of HPV-related OPSCC is rapidly increasing in the Western world, including the United States and parts of Europe, whereas HPV-negative OPSCC is declining [5, 6]. The majority of newly diagnosed OPSCCs are HPV-related malignancies [7].

HPV genomic integration frequently occurs in OPSCC, with estimated rates of 50-70% depending on the study [8-10]. Expression of HPV oncoproteins E6 and E7 is known to drive carcinogenesis, but recently, the genomic and transcriptomic alterations induced by HPV genomic integration has been investigated as an additional mechanism of carcinogenesis. It has been shown that genomic integration can lead to large-scale genomic rearrangements, deletions and amplifications that can alter expression of critical cellular genes [10, 11]. Other groups have reported that this process can alter gene expression transcriptome-wide, regardless of integration site [12, 13]. HPV integration takes place across the entire human genome with only a few reported hotspots, and as a whole, integration sites vary widely between samples [10, 14].

The oropharynx encompasses the base of tongue, soft palate, posterior and lateral pharyngeal wall, and the tonsils. Cancer can arise in any of these anatomical subsites, but the lingual and palatine tonsils are by far the most common subsite for the development of HPV+OPSCC [15]. It is hypothesized that the thin lymphoepithelium in this region may be more susceptible to infection, and the architecture of the tonsil crypts act as a reservoir for HPV, leading to persistent infection [2]. The non-tonsillar regions of the oropharynx, however, are lined by stratified squamous epithelium that likely acts as a barrier to infection and thereby decreases the likelihood of cancer development [16]. Previous studies have compared HPV+ OPSCCs in tonsillar versus non-tonsillar regions and reported patients with tonsillar tumors had a better disease-specific survival than those with non-tonsillar tumors [17, 18].

Of patients with tonsillar carcinomas, the majority of patients present with unilateral disease. Very few patients present with bilateral HPV+ tumors, and the literature on this phenomenon is somewhat limited [19-23]. A recent retrospective study reported that in a cohort of Danish patients, only 3.3% had synchronous bilateral tonsil cancer [24]. There are multiple hypotheses on the mechanism of this event. It is possible that independent carcinomas spontaneously form in each tonsil as a result of HPV infection at both sites. Others however hypothesize the two tumors are of clonal origin, in which carcinoma develops in one tonsil and a clonal population migrates away from the tumor to the other tonsil **(Figure 3.1)**.

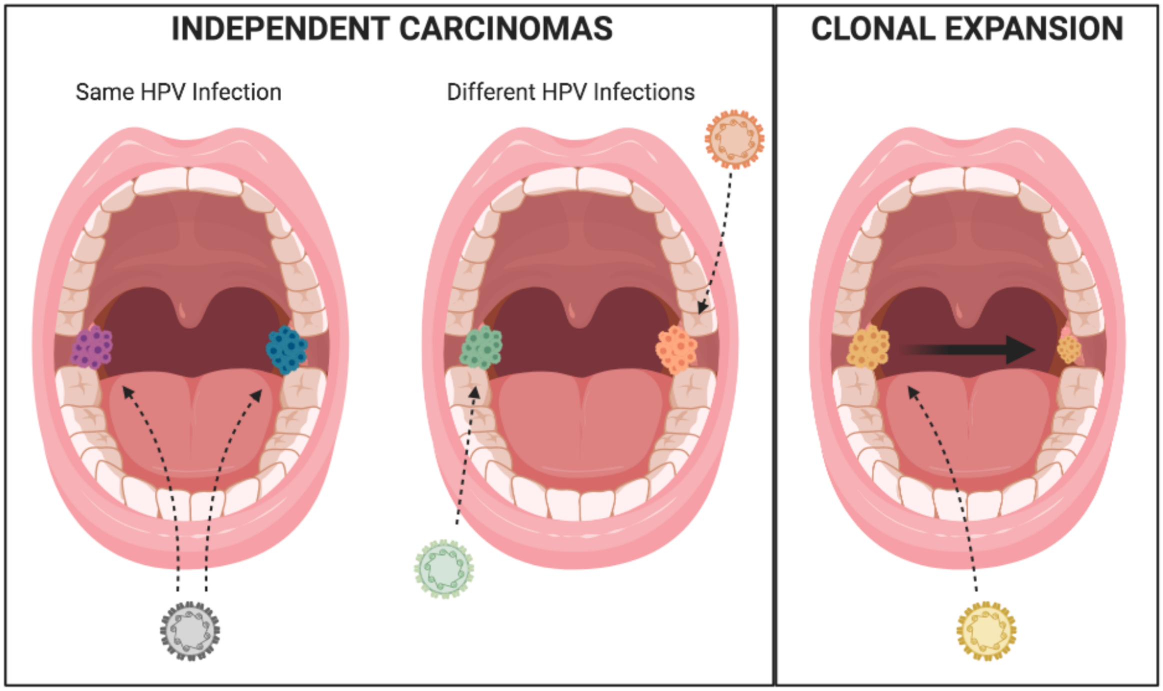


Figure 3.1. Mechanisms of bilateral tonsil tumor formation. Left: Carcinomas form independently in each tonsil as a result of either the same HPV exposure or different HPV exposures. Right: A tumor forms in one tonsil and then establishes a clonal tumor in the contralateral tonsil. Adapted from Joseph 2013 [19]. Copyright permission received on April 27, 2020, license number 4817181010803.

It has long been proposed that cancer is an evolutionary, stepwise process akin to Darwinian natural selection in which cells acquire new molecular alterations, resulting in a heterogeneous mixture of cells with differing profiles [25, 26]. Evolutionary pressure and competition in the tumor microenvironment allow for expansion of the fittest subclones. Subclones with beneficial alterations that permit them to survive the process of metastasis can then establish themselves as a secondary tumor in a new site. In the case of tonsillar carcinoma, it is possible that subclones of the original tumor break away and migrate into the other tonsil to establish a clonal secondary tumor. In cervical cancer, it has been reported that HPV integration may function to inactivate/activate genes that favor clonal expansion [14]. Comparison of the molecular alterations between the two tumors allows us to understand whether the two are related.

Here we present three interesting cases of synchronous bilateral tonsillar p16+ HPV+OPSCC. In order to assess the mechanism by which these tumors arose, we performed HPV genotyping and integration analysis to support or dispute the clonal expansion hypothesis. We would expect shared HPV types and integration sites if tumors were clonally related, as it is unlikely two unrelated tumors would have integration into the same loci due to the stochastic nature of this process. Analysis of two of the tumor sets revealed identical HPV types and both common and unique viral integration events. This suggests a common origin but individual evolution of the tumors, supporting the single-clone hypothesis of bilateral tumor development. However, the other patient did not follow this same pattern, as their tumors contained multiple HPV types unique from one another with no shared HPV integration sites. This suggests either that their tumors were formed spontaneously or that the subclones that grew out were underrepresented in the original tumor and therefore could not be detected. Therefore, we have provided evidence that bilateral carcinomas can sometimes form as a result of clonal expansion from one tonsil to another.

Materials and Methods

Tumor Specimens: Formalin-fixed, paraffin-embedded (FFPE) tissue blocks were received as described in **Table 3.1**. Patient one had two blocks: block 1A from a p16+ squamous cell carcinoma in the left tonsil and block 1B from a p16+ carcinoma in situ suspected to be a squamous cell carcinoma from the right side of the base of tongue. Patient two had three available blocks: block 2A from a p16+ squamous cell carcinoma in the left tonsil, block 2B from a p16+ squamous cell carcinoma in the right tonsil and block 2C from a p16+ squamous cell carcinoma found in the nasopharynx. Three blocks were available from patient three: block 3A from a p16+ squamous cell carcinoma originating from the left tonsil, blocks 3B and 3C originating from different regions of a p16+ squamous cell carcinoma in the right tonsil. Slides were prepared from each block, and the tissues were reviewed by a head and neck pathologist.

Patient #	Sex	Age	Block A	Block B	Block C
1	M	54	p16+ SCC, L tonsil	p16+ CIS, R BOT	-
2	M	52	p16+ SCC, L tonsil	p16+ SCC, R tonsil	p16+ SCC, nasopharynx
3	F	60	p16+ SCC, L tonsil	p16+ SCC, R tonsil	p16+ SCC, R tonsil

Table 3.1. Bilateral patient samples.

Abbreviations: F, female. M, male. SCC, squamous cell carcinoma. CIS, carcinoma in situ. L, left. R, right. BOT, base of tongue.

DNA/RNA Isolation: 10μm sections were taken from FFPE tissue blocks and mounted on a slide. Each section was aligned to the prepared H&E slides to identify the tumor-rich areas, and tissue within the tumor area was microdissected using a scalpel. Following microdissection, DNA and RNA were extracted from the tissue. DNA was isolated using the NucleoSpin DNA FFPE kit (Macherey-Nagel, Duren, Germany) according to the manufacturer's protocol. Briefly, paraffin was dissolved with xylene, and the tissue was lysed with lysis buffer and Proteinase K overnight at 56° C. Following overnight digestion, DNA was de-crosslinked, loaded onto NucleoSpin DNA columns, washed and then eluted in water. DNA concentration was measured using the QUBIT 2.0 Fluorometer.

RNA was extracted from blocks from patient three only using the High Pure RNA Paraffin Kit (Roche, Basel, Switzerland) according to the manufacturer's protocol. Briefly, paraffin was dissolved using heptane and methanol, and the tissue was lysed overnight at 56° C with lysis buffer containing Proteinase K and 10% SDS. RNA was extracted using the supplied High Pure Filters and wash buffers, followed by DNase I treatment. RNA was eluted in Elution Buffer and the concentration was measured using the QUBIT 2.0 Fluorometer.

p16 Staining: Tumor tissue sections were stained for p16 using the Roche/Cintec p16 mouse monoclonal antibody (#805-4713).

HPV Testing: HPV types present in each block were identified using HPV PCR-MassArray as previously described [27]. In brief, this method detects and identifies 15 high-risk HPV subtypes (16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66, 68, and 73), 2 low-risk subtypes (6 and 11), and HPV90, considered to be a possible high-risk subtype. The test included interrogation of human GAPDH as a control for sample DNA quality and assay validity. Type-specific, multiplex, competitive PCR was performed to amplify the E6 region of HPV, followed by probe-specific single base extension to discriminate between naturally occurring HPV present in the sample and the synthetic competitors included in the reaction. Matrix-assisted laser desorption/ionization time of flight mass spectroscopy was used for separation of products on a matrix-loaded silicon chip array. Samples were run in quadruplicate with appropriate positive and negative controls.

Detection of Integrated Papillomavirus Sequences (DIPS-PCR): DIPS-PCR was performed to identify the sites of HPV integration in the genome of the tumors, as previously described [28]. For each tumor, 0.75μg DNA was digested with one of two restriction enzymes, either TaqA1 or Sau3AI (New England Biolabs, Ipswich, MA). Adapters complementary to the unique overhangs created by restriction digestion were annealed to digested DNA. Linear PCR was performed on each sample using 11 different viral primers to amplify viral fragments. Following linear PCR, exponential PCR using 11 nested viral primers and an adapter-specific primer was performed. Products of the exponential PCR reactions were separated by gel electrophoresis (3% agarose gel). Bands were excised from the gel and were purified by Qiaquick Gel Extraction Kit (Qiagen, Hilden, Germany). Sanger sequencing of the isolated products was performed by the University of Michigan Sequencing Core, and the results were mapped using NCBI BLAST.

Direct PCR (Patient 3 only): Additional DNA was used to amplify the viral/cellular regions of integration identified by DIPS-PCR. Only patient three had sufficient DNA for this analysis. PCR was performed with genomic DNA from each tumor sample as well as DNA from the DIPS linear PCR reactions in order to enrich for viral products using the primers shown in **Table S3.1**. Amplicons were separated and visualized with gel electrophoresis and were confirmed by Sanger Sequencing of the excised and purified bands.

Transcript Analysis (Patient 3 only): cDNA was prepared from RNA extracted from the FFPE blocks. cDNA was synthesized from 1μg of RNA using Superscript III and random hexamers (Thermo Fisher Scientific, Waltham, MA). A no-reverse transcriptase control was prepared for each sample to ensure RNA purity. Primers were designed to amplify the native *CD36* and *LAMA3* transcripts proximal to and downstream of each viral integration site using NCBI Primer-BLAST (**Table S3.1**). Primers were also designed to amplify across the predicted fusion transcripts. Reverse transcription PCR (RT-PCR) was performed using Platinum Taq DNA polymerase. Products of RT-PCR were separated and visualized with gel electrophoresis and were confirmed by Sanger Sequencing of the excised and purified bands.

Results

Case Reports

Patient one: A 54-year-old male from Maastricht in the Netherlands presented with p16-positive tonsillar carcinoma in the left tonsil and p16-positive carcinoma in situ (CIS) in the base of tongue.

Patient two: A 52-year-old male from Maastricht in the Netherlands presented with bilateral p16-positive OPSCC in the left and right tonsils with an additional p16-positive, Epstein-Barr virus (EBV)-negative carcinoma in the nasopharynx.

Patient three: A 60-year-old female nun in Cologne, Germany presented with bilateral p16-positive OPSCCs of the tonsils. The left tonsil was diagnosed as pT2N2bcM0, the right as pT2N0M0, both Grade 2. One FFPE block from the left tonsil and two FFPE blocks from biopsies of different regions of the right tonsillar tumor were collected. Combined radio-chemotherapy was recommended, but only radiotherapy was performed (59.5/50.4Gy) because the patient refused chemotherapy. The patient was free of disease at her last visit to the clinic two years post-diagnosis, after which she was lost to follow up but was reported to have died approximately one year later due to pneumonia.

HPV Testing

Staining for p16 was positive for all blocks from patients one, two and three (**Figure 3.2**). HPV PCR-MassArray determined that each block belonging to patient one contained multiple HPV types; block A was HPV16+ and HPV31+, while block B was HPV16+ and HPV33+ **(Figure 3.3)**. HPV16 was the only HPV type present in blocks A, B and C for patient two as well as patient three.

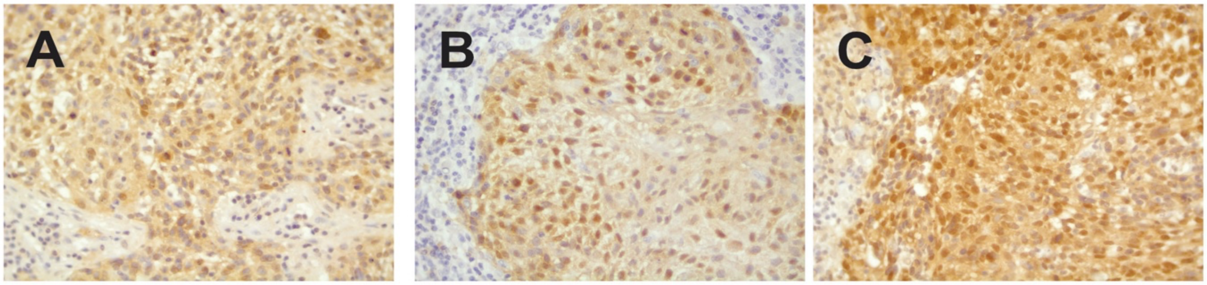


Figure 3.2. Immunohistochemical staining against p16^{INK4a}. Representative images from patient 3 blocks A-C, V=400X.

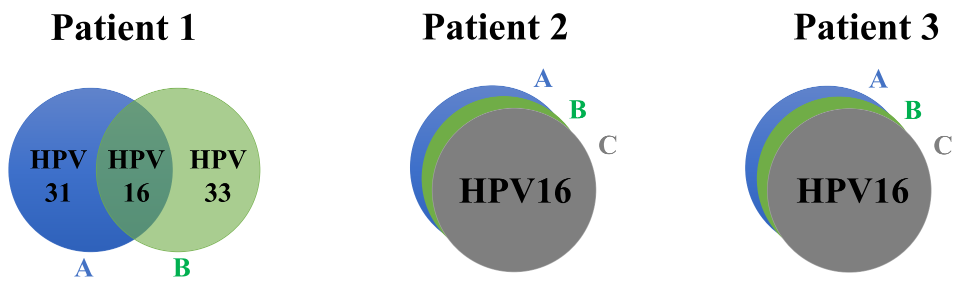


Figure 3.3. HPV types present in bilateral tumors. Each block represented as a circle. Blue, block A. Green, block B. Gray, block C.

Integration Analysis

DIPS-PCR revealed multiple integration sites in the tumor samples with every sample containing at least one site **(Table 3.2)**. A total of fourteen integration sites were identified; eight (57%) viral integrations occurred in intergenic or genomic scaffold regions, and six (43%) viral integrations occurred in cellular genes. Chromosome 4p15 had the most frequently affected loci. Breakpoints in the HPV genome were most frequently detected in E2 (36%), E1 (21%) and E5 (21%). Breakpoints were also detected in E6, L2 and L1.

Patient	Sample	Viral insertion (nt)	Map	Integration locus	Database comparison
1	Left tonsil (1A)	3788 (E2)	7q22	*TRRAP* intron 51	NM_003496.3
	BOT (1B)	5020 (L1)	4p15	Intergenic	NC_000004.12
2	Left tonsil (2A)	50 (E6)	Unplaced	Genomic scaffold	NT_187433.1
		1049 (E1)	1p22	*CLCA4* exon 8	NM_012128.3
		3480 (E2)	4p15	Intergenic	NC_000004.12
	Right tonsil (2B)	2139 (E1)	14q23	*HIF1A* exon 5	NM_001530.3
		3840 (E2)	4p15	Intergenic	NC_000004.12
	Nasopharynx (2C)	3840 (E2)	4p15	Intergenic	NC_000004.12
3	Left tonsil (3A)	2088 (E1)	Unplaced	Genomic scaffold	NT_187433.1
		3886 (E5)	4p15	Intergenic	NC_000004.12
		5617 (L2)	7q21	*CD36* intron 6	NG_008192.1
	Right tonsil (3B)	3867 (E5)	18q11	*LAMA3* intron 1 – intron 68 fusion	NG_007853.2
	Right tonsil (3C)	3213 (E2)	4q28	Intergenic	NC_000004.12
		3854 (E5)	7q21	*CD36* intron 5	NG_008192.1

Table 3.2. HPV integration sites in bilateral tumor sets.

Patient one: Each sample from patient 1 contained a unique HPV16 integration site (**Figure 3.4a)**. Block 1A showed integration of E2 into intron 51 of the cellular gene *TRRAP* located on chromosome 7q22, and block 1B contained an integration of L2 into an intergenic region of chromosome 4q15.

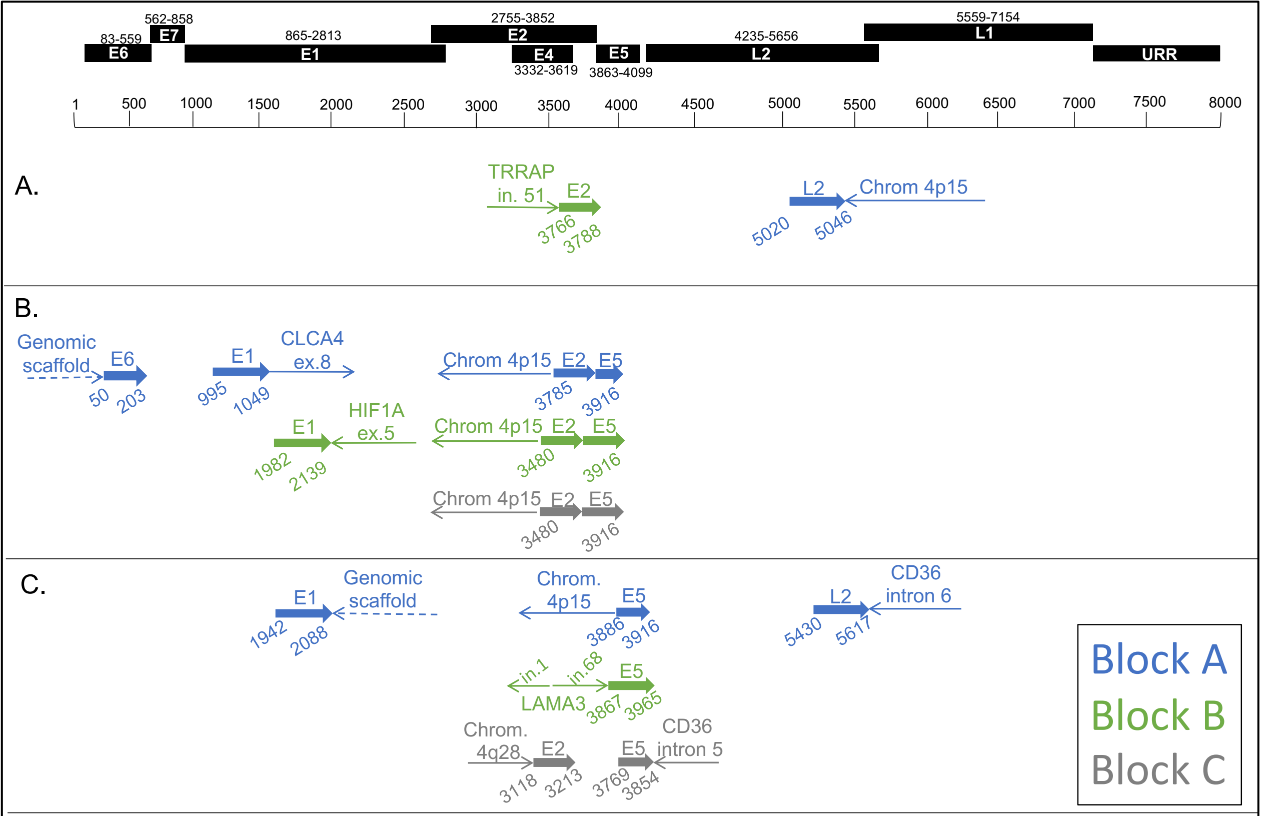


Figure 3.4. Viral integration sites identified in each bilateral tumor. A. Patient one, B. patient 2, C. patient 3. Each color represents block of origin.

Patient two: DIPS-PCR of the tonsils and nasopharynx tumors from patient 2 revealed a common HPV16 E2 integration into chromosome 4p15 (**Figure 3.4b)**. Each tonsil also contained additional integration sites that were not shared; the left tonsil showed integration of E6 into a genomic scaffold region and E1 into the cellular gene *CLCA4*. The right tonsil showed integration of E1 into exon 5 of *HIF1A*.

Patient three: Multiple integration sites were detected in the tumors from patient three (**Figure 3.4c**). The tumor from the block 3A of the left tonsil had integration of HPV16 L2 into intron 6 of the cellular gene *CD36*. Two blocks from the right tonsil (3B and 3C) were analyzed; the tumor cells from the right tonsil had integrations of HPV16 E5 into intron 5 of *CD36*, as well as an additional integration of HPV16 E5 into *LAMA3*, which likely caused a rearrangement of *LAMA3* intron 1 and intron 68.

Due to an abundance of tissue, we were able to perform direct PCR to confirm the rearrangements and to check for gene disruption. Direct PCR and Sanger sequencing confirmed the *LAMA3* rearrangement and integration of HPV16 E5 occurred in the right tonsil (3B), but it was not present in the left tonsil (3A) or the second block (3C) from the right tonsil, suggesting this rearrangement did not persist in the other intratumoral clonal populations (**Figure 3.5a**). Amplification of the native exons of *LAMA3* DNA showed that there is an intact copy of *LAMA3* exon 2 present in all 3 samples (**Figure 3.5b**). However, only tumor 3B, which contained E5 integration into *LAMA3,* showed an intact copy *LAMA3* exon 1.

Direct PCR of the *CD36* integration site found in the right tonsil (E5-CD36 intron 5) in the other blocks yielded no products; similarly, direct PCR of the site found in the left tonsil (L2-CD36 intron 6) in the right tonsil blocks yielded no products. Amplification of the native exons of *CD36* DNA revealed exons 4, 5, and 6 were present in all blocks (**Figure 3.5c**).

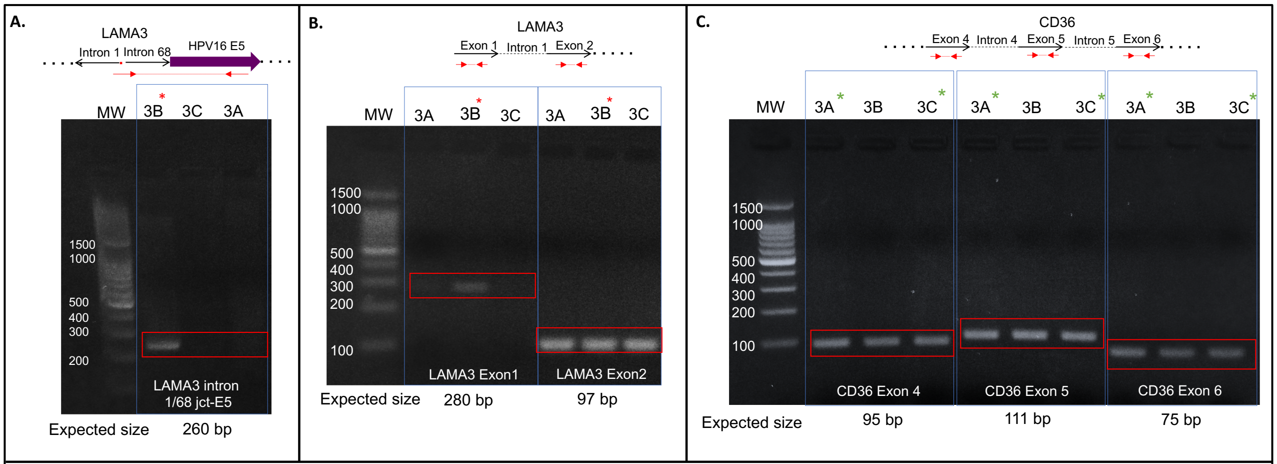


Figure 3.5. *LAMA3* and *CD36* PCRs in bilateral patient 3. Red and green star indicates sample is affected by integration into LAMA3 or CD36, respectively. PCR amplicons A. LAMA3-HPV16 E5 integration, B. Native LAMA3 exons surrounding integration site, C. Native CD36 exons surrounding integration sites.

Transcript Analysis (Patient 3 only)

RNA from the tumors of patient three revealed that HPV16 E6-E7 transcripts were present in each sample (**Figure 3.6a**). Alternate E6-E7 transcripts were visible in each sample, but not full length E6-E7. The left tonsil tumor 3A, and one block from the right tonsil (3B) showed expression of E6*I-E7 and E6*II-E7, but block 3C only expressed E6*II-E7, consistent with evolution of the viral segments with tumor progression.

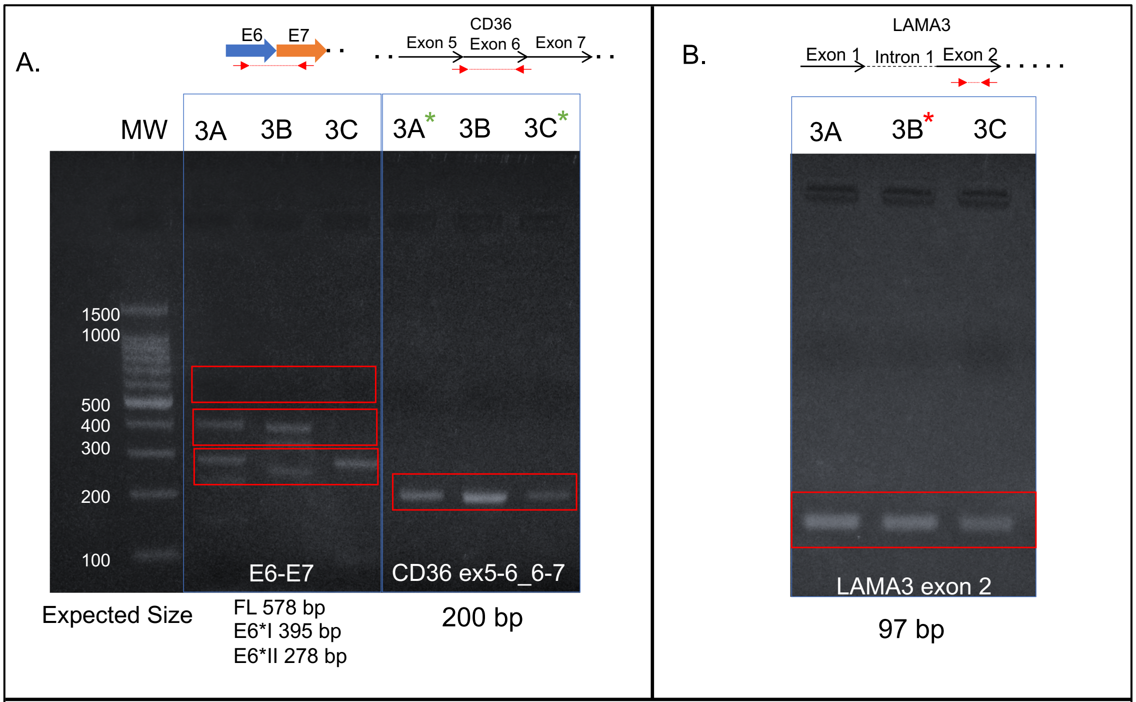


Figure 3.6. Transcript expression in bilateral patient 3. Red and green stars indicate sample is affected by integration into LAMA3 or CD36, respectively. PCR amplicons A. HPV16 E6-E7 and CD36 exon5-6 junction to exon 6-7 junction B. Native LAMA3 exon downstream of integration site.

In order to understand whether viral integration into *LAMA3* and *CD36* disrupted expression of the genes, RT-PCR was performed. RT-PCR revealed that a transcript of the native *LAMA3* exon downstream from the integration site (exon 2) was expressed in all 3 samples (**Figure 3.6a**). Normal *CD36* transcripts spanning the two integration sites (the forward primer was designed to amplify the cDNA junction of exon 5-6 and the reverse primer to amplify from the exon 6-7 junction) were also expressed in all 3 samples (**Figure 3.6b**). These data suggest there is normal expression of at least one copy *CD36* in each tumor or that the HPV L2 and HPV E5 were spliced out of the transcripts along with introns 5 and 6.

We attempted to amplify viral-host fusion transcripts that may have resulted from viral integration into *LAMA3* or *CD36*. However, RT-PCR using primers designed to amplify the predicted fusion transcripts of *LAMA3* and *CD36* failed to yield any products. Every effort was made to limit the size of the amplicons, as FFPE RNA is usually highly fragmented making it difficult to amplify long products. Therefore, it is unclear whether these fusion transcripts are produced but unidentifiable due to fragmentation, or if they lack a viable promoter and are not expressed.

Discussion

The incidence of HPV+OPSCC is rapidly rising in the Western world. Despite an overall survival advantage compared to HPV-negative cancers, there is still a significant proportion of patients who develop local or distant recurrences within 5 years and treatment de-escalation has failed [29, 30]. Therefore, there is a critical need to understand the cellular and molecular characteristics of HPV+OPSCCs with unfavorable outcome. Some studies have shown that viral integration of HPV into the genome is associated with worse prognosis [12, 13, 31] and these tumors have a different mutation signature, particularly of *PIK3CA* [32]. Viral integration into cellular genes may lead to disruption of gene expression and generation of viral-human fusion transcripts.

Synchronous HPV+ bilateral tonsillar carcinomas are relatively rare with about 40 cases reported in the literature; however, studying their characteristics may contribute to our understanding of OPSCC [21, 24]. There is much controversy in the field whether all HPV+ tonsillar carcinoma patients should have their contralateral tonsil removed in order to prevent missing bilateral disease [33]. Others oppose this idea due to increased morbidity of removing both tonsils and lack of sufficient evidence this would benefit patients, especially given that most patients are given adjuvant radiation and monitored closely so disease in the contralateral tonsil would be treated and detected [34]. The mechanism behind synchronous bilateral carcinoma development is debated; there is evidence for both clonal expansion of a single primary tumor as well as simultaneous development of independent carcinomas due to similar HPV exposure. Understanding how these tumors develop would help inform whether or not all patients with unilateral tonsillar carcinoma are at risk for development of bilateral disease.

Here we have presented three cases of synchronous bilateral HPV+ tonsillar OPSCC. HPV genotyping confirmed shared HPV types in two out of three patients, whereas the remaining patient had discordant types. Viral integration analysis of each set of tumors highlighted a number of HPV16 integration sites into both intergenic and genic regions of the cellular genome. There were viral integrations unique to each tonsil as well as some shared sites.

HPV analysis of the tumors from patient one revealed multiple HPV types; HPV16 was shared among both tumors, but other discordant types were present in each. Integration analysis of HPV16 in these samples revealed no shared integration sites. This suggests that these tumors could have formed independently, given that they contain unique HPV types and unique integration loci. However, it is also a possibility that these tumors came from one original tumor; a minor highly metastatic subclone that metastasized and dominated in the other site but could have been represented by too few cells and therefore too diluted in the original population to be detected by our assays, such that it would appear the two tumors do not share common HPV types. Further investigation of the genetic makeup of these two tumors would help clarify if that is the case.

Patients two and three, however, demonstrated clear evidence of clonally related tumors. All paired samples shared an HPV16 infection and demonstrated HPV16 integration into the same loci (chromosome 4p15 in patient two and the gene CD36 on chromosome 7q21 in patient three). Patient two's shared sites were essentially identical; the same regions of HPV and chromosome 4 were involved. In patient three however, the same cellular gene was involved, but different but adjacent regions (E5-L2) of HPV were involved, and the integration sites were located in two different but sequential *CD36* introns (intron 5 and intron 6). Although the locations of integration are slightly different, we expect it would be unlikely to discover viral integration into the same gene in unrelated samples; the majority of samples have unique viral integration sites due to the stochastic nature of viral integration [10, 35]. There is some evidence that viral integration occurs preferentially into genes [14], but it is unlikely that these integrations into *CD36* are the result of separate viral integration events in two independent tumors. Furthermore, *CD36* is not reported to be a hotspot of HPV integration, in contrast to other genes that have been reported in several cases [10, 36-38]. We believe that these two tumors arose from one primary tumor with viral integration into *CD36* that underwent clonal expansion and was subsequently established in the other tonsil. However, the altered site of integration also suggests the DNA was edited with tumor evolution over time. It is unknown whether viral integration events are stable over time or if they are subject to changes due to either mobile element characteristics or genomic instability. It is also possible that the initial *CD36* integration site was established via a "looping" integration mechanism and some of the integrated DNA was subsequently excised during clonal expansion, as has been described by others [39, 40].

In patient three, the *LAMA3* gene containing the viral integration was rearranged and inverted as a result of HPV16 E5 insertion as shown in **Figure 3.4c**. The implications of the viral integration on *LAMA3* function are unclear because no intact *LAMA3* gene was present in blocks 3A and 3C. Limited analysis of the *LAMA3* gene revealed that there is loss of *LAMA3* exon 1 in the blocks without HPV E5-*LAMA3* integration, but exon 2 is present in all three tumor areas. One possibility to explain the observation that block 3A and 3C lost exon 1 DNA but it is retained in block 3B is that the initial *LAMA3* integration site was established via a "looping" integration mechanism which reversed the orientation of the *LAMA3* gene and the integrated DNA, including *LAMA3* exon 1, was subsequently excised during clonal expansion. This suggests that the *LAMA3* rearrangement included exon 1 and that this was an early event which was later excised as the tumor progressed in the right tonsil (3C) and the left tonsil (3A), but the only copy of *LAMA3* carried in those tumor cells was the copy missing exon 1.

The cellular genes located at the HPV integration sites were of particular interest because of their involvement in head and neck cancer. *TRRAP* encodes for a protein that complexes with histone acetyltransferases to mediate diverse cellular processes by acetylation of histones [41]. Mutations in histones and histone modifiers are frequent in HNSCC, and *TRRAP* is amplified or mutated in 11% of HNSCC patients [42]. *CLCA4*, known as chloride channel accessory 4, has been shown to be an inhibitor of cellular proliferation frequently downregulated during tumor progression [43]. *HIF1A* is a well characterized marker of hypoxia that mediates cellular responses to hypoxic stress; HNSCC patients who overexpress HIFs have an increased risk of mortality [44]. *LAMA3* encodes for the laminin subunit alpha-3, which is one of three members of the complex glycoprotein laminin 5. Laminins are components of the cellular basement membrane, and laminin 5 is reported to be involved in cell adhesion, migration, and the differentiation of keratinocytes [45]. Laminin 5 has been shown to be overexpressed in invasive oral squamous cell carcinomas but not in premalignant lesions [46, 47]. *CD36*, or cluster of differentiation 36, encodes an integral membrane protein involved in fatty acid import and binds many ligands including collagen, lipoproteins, phospholipids and long-chain fatty acids. Studies done suggest *CD36* may promote cell migration and proliferation in oral cancers and other solid tumors [48-50]. Further investigation of protein expression in these samples is warranted to understand whether these proteins are disrupted as a result of HPV integration and whether that may have played a role in tumor spread.

Our assessment of tumor clonality was limited to HPV genotype and HPV integration site, which other groups have also used to demonstrate tumor clonality [51, 52], but future work will include mutational profiling of cellular genes to strengthen our ability to assess the clonal nature of these samples. This work was somewhat limited by having access to only a small amount of FFPE tissue, resulting in a low amount of DNA and lack of quality RNA for integrated viral transcript analysis. This limited our ability to detect gene expression changes of regions involved in integration, as well as detection of the predicted viral-host fusion transcripts. It also would have been valuable to evaluate the metastatic lymph node associated with the left pT2N2bcM0 tonsil of patient 3, but DNA was not available for analysis from that metastatic lesion. Given the highly metastatic nature of HPV+ tonsillar carcinomas [53, 54], it is not be surprising to us that metastasis of HPV transformed cancer cells from one lymphoid bed within the oropharynx to another across the midline occurs. Overall, our study supports clonal spread from one tonsil to another, and future work will be focused on validating these results in other bilateral tonsil pairs.

References

1. D'Souza, G., T.S. McNeel, and C. Fakhry, *Understanding personal risk of oropharyngeal cancer: risk-groups for oncogenic oral HPV infection and oropharyngeal cancer.* Ann Oncol, 2017. **28**(12): p. 3065-3069.
2. Mena, M., et al., *Might Oral Human Papillomavirus (HPV) Infection in Healthy Individuals Explain Differences in HPV-Attributable Fractions in Oropharyngeal Cancer? A Systematic Review and Meta-analysis.* J Infect Dis, 2019. **219**(10): p. 1574-1585.

3. Ortiz, A.P., et al., *Periodontitis and oral human papillomavirus infection among Hispanic adults.* Papillomavirus Res, 2018. **5**: p. 128-133.
4. Pierce Campbell, C.M., et al., *Long-term persistence of oral human papillomavirus type 16: the HPV Infection in Men (HIM) study.* Cancer Prev Res (Phila), 2015. **8**(3): p. 190-6.
5. Brouwer, A.F., M.C. Eisenberg, and R. Meza, *Age Effects and Temporal Trends in HPV-Related and HPV-Unrelated Oral Cancer in the United States: A Multistage Carcinogenesis Modeling Analysis.* Plos One, 2016. **11**(3).
6. Chaturvedi, A.K., et al., *Human papillomavirus and rising oropharyngeal cancer incidence in the United States.* J Clin Oncol, 2011. **29**(32): p. 4294-301.
7. Mehanna, H., et al., *Prevalence of human papillomavirus in oropharyngeal and nonoropharyngeal head and neck cancer--systematic review and meta-analysis of trends by time and region.* Head Neck, 2013. **35**(5): p. 747-55.
8. Gao, G., et al., *Whole genome sequencing reveals complexity in both HPV sequences present and HPV integrations in HPV-positive oropharyngeal squamous cell carcinomas.* BMC Cancer, 2019. **19**(1): p. 352.
9. Nulton, T.J., et al., *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**(11): p. 17684-17699.
10. Parfenov, M., et al., *Characterization of HPV and host genome interactions in primary head and neck cancers.* Proc Natl Acad Sci U S A, 2014. **111**(43): p. 15544-9.
11. Akagi, K., et al., *Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability.* Genome Res, 2014. **24**(2): p. 185-99.
12. Huebbers, C.U., et al., *Upregulation of AKR1C1 and AKR1C3 expression in OPSCC with integrated HPV16 and HPV-negative tumors is an indicator of poor prognosis.* Int J Cancer, 2019. **144**(10): p. 2465-2477.
13. Koneva, L.A., et al., *HPV Integration in HNSCC Correlates with Survival Outcomes, Immune Response Signatures, and Candidate Drivers.* Mol Cancer Res, 2017.
14. Bodelon, C., et al., *Genomic characterization of viral integration sites in HPV-related cancers.* Int J Cancer, 2016. **139**(9): p. 2001-11.
15. Haeggblom, L., et al., *Time to change perspectives on HPV in oropharyngeal cancer. A systematic review of HPV prevalence per oropharyngeal sub-site the last 3 years.* Papillomavirus Res, 2017. **4**: p. 1-11.
16. Fossum, C.C., et al., *Characterization of the oropharynx: anatomy, histology, immunology, squamous cell carcinoma and surgical resection.* Histopathology, 2017. **70**(7): p. 1021-1029.
17. Tham, T., et al., *Anatomical subsite modifies survival in oropharyngeal squamous cell carcinoma: National Cancer Database study.* Head Neck, 2020. **42**(3): p. 434-445.
18. Tham, T., et al., *The prognostic effect of anatomic subsite in HPV-positive oropharyngeal squamous cell carcinoma.* Am J Otolaryngol, 2019. **40**(4): p. 567-572.
19. Joseph, A.W., et al., *Molecular etiology of second primary tumors in contralateral tonsils of human papillomavirus-associated index tonsillar carcinomas.* Oral Oncol, 2013. **49**(3): p. 244-8.
20. Roeser, M.M., et al., *Synchronous bilateral tonsil squamous cell carcinoma.* Laryngoscope, 2010. **120 Suppl 4**: p. S181.
21. Theodoraki, M.N., et al., *Synchronous bilateral tonsil carcinoma: case presentation and review of the literature.* Infect Agent Cancer, 2017. **12**: p. 38.

22. Shimizu, F., et al., *Synchronous HPV-Related Cancer of Bilateral Tonsils Detected Using Transoral Endoscopic Examination with Narrow-Band Imaging.* Case Rep Otolaryngol, 2017. **2017**: p. 9647010.

23. Rasband-Lindquist, A., Y. Shnayder, and M. O'Neil, *Synchronous bilateral tonsillar squamous cell carcinoma related to human papillomavirus: Two case reports and a brief review of the literature.* Ear Nose Throat J, 2016. **95**(4-5): p. E30-4.

24. Rokkjaer, M.S. and T.E. Klug, *Prevalence of synchronous bilateral tonsil squamous cell carcinoma: A retrospective study.* Clin Otolaryngol, 2018. **43**(1): p. 1-6.

25. Greaves, M. and C.C. Maley, *Clonal evolution in cancer.* Nature, 2012. **481**(7381): p. 306-13.

26. Nowell, P.C., *The clonal evolution of tumor cell populations.* Science, 1976. **194**(4260): p. 23-8.

27. Walline, H.M., et al., *High-risk human papillomavirus detection in oropharyngeal, nasopharyngeal, and oral cavity cancers: comparison of multiple methods.* JAMA Otolaryngol Head Neck Surg, 2013. **139**(12): p. 1320-7.

28. Luft, F., et al., *Detection of integrated papillomavirus sequences by ligation-mediated PCR (DIPS-PCR) and molecular characterization in cervical cancer cells.* Int J Cancer, 2001. **92**(1): p. 9-17.

29. Mehanna, H., et al., *Radiotherapy plus cisplatin or cetuximab in low-risk human papillomavirus-positive oropharyngeal cancer (De-ESCALaTE HPV): an open-label randomised controlled phase 3 trial.* Lancet, 2019. **393**(10166): p. 51-60.

30. Gillison, M.L., et al., *Radiotherapy plus cetuximab or cisplatin in human papillomavirus-positive oropharyngeal cancer (NRG Oncology RTOG 1016): a randomised, multicentre, non-inferiority trial.* Lancet, 2019. **393**(10166): p. 40-50.

31. Walline, H.M., et al., *Genomic Integration of High-Risk HPV Alters Gene Expression in Oropharyngeal Squamous Cell Carcinoma.* Mol Cancer Res, 2016. **14**(10): p. 941-952.

32. Leemans, C.R., P.J.F. Snijders, and R.H. Brakenhoff, *The molecular landscape of head and neck cancer.* Nat Rev Cancer, 2018. **18**(5): p. 269-282.

33. Kim, C.M. and M.A. St John, *Should the Contralateral Tonsil Be Removed in Cases of HPV-Positive Squamous Cell Carcinoma of the Tonsil?* Laryngoscope, 2019. **129**(6): p. 1257-1258.

34. Cognetti, D., et al., *In Reference to Should the Contralateral Tonsil Be Removed in Cases of HPV-Positive Squamous Cell Carcinoma of the Tonsil?* Laryngoscope, 2019. **129**(6): p. E194.

35. Lace, M.J., et al., *Human papillomavirus type 16 (HPV-16) genomes integrated in head and neck cancers and in HPV-16-immortalized human keratinocyte clones express chimeric virus-cell mRNAs similar to those found in cervical cancers.* J Virol, 2011. **85**(4): p. 1645-54.

36. Hu, Z., et al., *Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism.* Nat Genet, 2015. **47**(2): p. 158-63.

37. Cancer Genome Atlas Research, N., et al., *Integrated genomic and molecular characterization of cervical cancer.* Nature, 2017. **543**(7645): p. 378-384.

38. Li, W., et al., *Characteristic of HPV Integration in the Genome and Transcriptome of Cervical Cancer Tissues.* Biomed Res Int, 2018. **2018**: p. 6242173.

39. Groves, I.J. and N. Coleman, *Human papillomavirus genome integration in squamous carcinogenesis: what have next-generation sequencing studies taught us?* J Pathol, 2018. **245**(1): p. 9-18.
40. Holmes, A., et al., *Mechanistic signatures of HPV insertions in cervical carcinomas.* NPJ Genom Med, 2016. **1**: p. 16004.
41. Murr, R., et al., *Orchestration of chromatin-based processes: mind the TRRAP.* Oncogene, 2007. **26**(37): p. 5358-72.
42. Cancer Genome Atlas, N., *Comprehensive genomic characterization of head and neck squamous cell carcinomas.* Nature, 2015. **517**(7536): p. 576-82.
43. Yu, Y., V. Walia, and R.C. Elble, *Loss of CLCA4 promotes epithelial-to-mesenchymal transition in breast cancer cells.* PLoS One, 2013. **8**(12): p. e83943.
44. Gong, L., et al., *Prognostic value of HIFs expression in head and neck cancer: a systematic review.* PLoS One, 2013. **8**(9): p. e75094.
45. Hamill, K.J., A.S. Paller, and J.C. Jones, *Adhesion and migration, the diverse functions of the laminin alpha3 subunit.* Dermatol Clin, 2010. **28**(1): p. 79-87.
46. Rani, V., M. McCullough, and A. Chandu, *Assessment of laminin-5 in oral dysplasia and squamous cell carcinoma.* J Oral Maxillofac Surg, 2013. **71**(11): p. 1873-9.
47. Rahman, F., et al., *The expression of laminin-5 in severe dysplasia/carcinoma in situ and early invasive squamous cell carcinoma: an immunohistochemical study.* Minerva Stomatol, 2013. **62**(5): p. 139-46.
48. Nath, A., et al., *Elevated free fatty acid uptake via CD36 promotes epithelial-mesenchymal transition in hepatocellular carcinoma.* Sci Rep, 2015. **5**: p. 14752.
49. Hale, J.S., et al., *Cancer stem cell-specific scavenger receptor CD36 drives glioblastoma progression.* Stem Cells, 2014. **32**(7): p. 1746-58.
50. Sakurai, K., et al., *CD36 expression on oral squamous cell carcinoma cells correlates with enhanced proliferation and migratory activity.* Oral Dis, 2019.
51. Arfi, A., et al., *HPV DNA integration site as proof of the origin of ovarian metastasis from endocervical adenocarcinoma: three case reports.* BMC Cancer, 2019. **19**(1): p. 375.
52. Lillsunde Larsson, G., et al., *HPV16 viral characteristics in primary, recurrent and metastatic vulvar carcinoma.* Papillomavirus Res, 2018. **6**: p. 63-69.
53. Goldenberg, D., et al., *Cystic lymph node metastasis in patients with head and neck cancer: An HPV-associated phenomenon.* Head Neck, 2008. **30**(7): p. 898-903.
54. McHugh, J.B., *Association of cystic neck metastases and human papillomavirus-positive oropharyngeal squamous cell carcinoma.* Arch Pathol Lab Med, 2009. **133**(11): p. 1798-803.

Chapter 4 SearcHPV: Novel viral integration detection methodology

Abstract

Human papillomavirus (HPV) is a well-established driver of malignant transformation resulting in squamous cell carcinomas of the head and neck, uterine cervix, vulva, anus-rectum and penis; however, the impact of HPV integration into the host human genome on this process remains largely unresolved. This is due to the technical challenge of identifying HPV integration sites, which includes limitations of existing informatics approaches to discover viral-host breakpoints from short read sequencing data. To overcome this limitation, we optimized a new sequencing and analysis pipeline called SearcHPV. Through analysis of HPV+ models, we show that SearcHPV detects HPV-host integration sites with a higher confirmation rate than existing callers. We then performed an integrated analysis of SearcHPV-defined breakpoints with genome-wide linked read sequencing. These methods demonstrated that HPV integration sites were found not only adjacent to known cancer-related genes such as *TP63* and *MYC*, but also near regions of large structural variation in the tumor genome. Further, analysis of SearcHPV-assembled junction contigs demonstrated that the tool can be used to accurately identify viral-host junction sequences and showed that viral integration occurs through a variety of DNA repair mechanisms including non-homologous end joining, alternative end joining and microhomology mediated repair. In summary, we show that SearcHPV is a new optimized tool for the detection of HPV-human integration sites from short read DNA sequencing data.

Introduction

Human papillomavirus (HPV) is a well-established driver of malignant transformation in a number of cancers, including head and neck squamous cell carcinoma (HNSCC). Although HPV genomic integration is not a normal event in the lifecycle of HPV, it is frequently reported in HPV+HNSCC and studies have shown it may be a contributor to oncogenesis [1-4]. In cervical cancer, HPV integration increases in incidence during progression from stages of cervical intraepithelial neoplasia (CIN) I/II, CIN III and invasive cancer development [5]. This process has a variety of impacts on both the HPV and cellular genomes, including disruption of HPV E2, which acts as a transcriptional repressor of HPV E6/E7, leading to dysregulation of E6/E7 expression and an increase in genetic instability [6]. HPV integration occurs within or near known cellular genes more often than expected by chance [7] and has been reported to be associated with structural variations [8] and increases in DNA copy number [3]. Recent studies in HNSCCs have also suggested that additional oncogenic mechanisms of HPV integration may exist through direct effects on cancer-related gene expression and generation of hybrid viral-host fusion transcripts [9].

A wide array of methods has been previously used for the detection of HPV integration. Polymerase chain reaction (PCR)-based methods, such as Detection of Integrated Papillomavirus Sequences PCR (DIPS-PCR) [10] and Amplification of Papillomavirus Oncogene Transcripts (APOT) [11], are direct ways to interrogate sites of HPV integration, but they are low sensitivity assays and are therefore still limited in their ability to detect the broad spectrum of genomic changes resulting from this process. The use of next-generation sequencing (NGS) technologies provides the opportunity for in-depth characterization of these events. Previous groups have assessed HPV integration within HNSCC tumors in The Cancer Genome Atlas (TCGA) and cell lines by a combination of whole-genome sequencing (WGS) and RNAseq [2, 3, 8]. To process WGS data, viral-human fusion callers, such as VirusFinder2 [12, 13] and VirusSeq [14], have been developed. However, these strategies are designed for a broad range of virus types and require whole genomes to be sequenced at uniform coverage, which can result in a lower sensitivity of detection for specific types of rare viral integration events.

To overcome this issue, others have begun to use HPV targeted capture sequencing, mainly focused on anogenital samples [5, 15-18]. This strategy allows for enrichment for integration sites and better coverage of regions of interest than an untargeted approach like WGS. However, assessing HPV integration sites from this type of data requires sensitive and accurate viral-human fusion detection bioinformatic tools, of which the field has been lacking. In our lab, we have found the previously available viral integration callers VirusFinder2 and VirusSeq, which were designed for WGS instead of a targeted capture approach, to have a relatively low validation rate and limitations on the structural information surrounding the fusion sites, which impairs the ability to investigate the mechanisms of integration from capture based sequencing data. Therefore, we set out to generate a novel pipeline specifically for targeted capture sequencing data to serve as a new gold standard in the field of viral integration calling.

Materials and Methods

Cell Line Model: UM-SCC-47 was previously derived in our lab from a surgical resection of a previously untreated p16+ T3N1M0 carcinoma of the lateral tongue in a 53-year-old male smoker [19]. The patient died within a year of diagnosis. Subsequent HPV testing demonstrated the cell line to be p16+ and HPV16+ [20, 21].

Patient Derived Xenograft (PDX) Model: Flash frozen tissue from an HPV16+ OPSCC PDX model (PDX-932174-294-R, subsequently abbreviated PDX-294R) was obtained from the National Cancer Institute Patient-Derived Models Repository (NCI-PDMR), NCI-Frederick, Frederick National Laboratory for Cancer Research (Frederick, MD) – https://pdmr.cancer.gov/.

The PDX was derived from a base of tongue squamous cell carcinoma from a 62-year-old, treatment naive male patient.

DNA Isolation: High molecular weight DNA was isolated by treating the samples overnight at 37° with lysis buffer (10 mM Tris-HCl, 400 mM NaCl, 2 mM EDTA), 10% SDS, RNase A and a proteinase K solution (1 mg/mL Proteinase K, 1% SDS, 2 mM EDTA). DNA was then salted out of the solution with 5 M NaCl for 1 hour at 4° and precipitated with ice cold ethanol for 5 hours at -20°C. High molecular weight DNA was eluted in TE buffer; the quality and integrity of the DNA was assessed using the Tapestation Genomic DNA ScreenTape kit (Agilent, Santa Clara, CA).

Targeted Capture Sequencing: DNA from UM-SCC-47 and PDX-294R were submitted to the University of Michigan Advanced Genomics Core for targeted capture sequencing. Targeted capture was performed using a custom designed probe panel with high density coverage of the HPV16 genome, the HPV18/33/35 L2/L1 regions, as well as over 200 HNSCC-related genes; the list of genes and approach for library preparation and targeted capture are detailed in Heft Neal et. al 2020 [22]. Following library preparation and capture, the samples were sequenced on an Illumina NovaSEQ6000 or HiSEQ4000, respectively, with 300nt paired end run. Data was de-multiplexed and FastQ files were generated.

Novel Integration Caller (SearcHPV): The pipeline of SearcHPV was illustrated in **Figure 4.1**, which overall has four main steps: (1) Alignment; (2) Genome fusion point calling; (3) Assembly; (4) HPV fusion point calling. These steps are elaborated in detail below.

Alignment

The customized reference genome used for alignment was constructed by catenating the HPV16 genome (from Papillomavirus Episteme (PAVE) database [23, 24]) and the human genome reference (1000 Genomes Reference Genome Sequence, hs37d5). We aligned paired-end reads from targeted capture sequencing against the customized reference genome using BWA mem aligner [25]. Then we performed an indel realignment by Picard Tools [26] and GATK [27]. Duplications were marked by Picard MarkDuplicates Tool [26] for the filtering in downstream steps.

Genome Fusion Points Calling

To identify the fusion points on the human genome, we extracted reads that have regions matched to HPV16 and filtered those reads to meet these criteria: (1) not secondary alignment; (2) mapping quality greater than or equal than to 50; (3) not duplicated. Genome fusion points were called by split reads (reads spanning the human genome and HPV genome) and the paired-end reads (reads that have one end matched to HPV and the other end matched the human genome) at the surrounding region (+/-300 base pairs (bp)) were used as supportive evidence to identify the fusion points (**Figure 4.1A**). The cut-off criteria for identifying the fusion points were based on empirical practice. We then clustered the integration sites within 100bp to avoid duplicated counting of integration events due to the stochastic feature of read mapping and structural variations.

Assembly

To construct longer sequence contigs from individual reals, we extracted supporting split reads and paired-end reads used for genome fusion points calling for local assembly from each integration event. Due to the library preparation methods we implemented for the targeted capture approach, some reads exhibited an insertion size less than 2 x read length, resulting in overlapping read segments. For such events, we first merged these reads using PEAR [28] and then combined them with other individual reads to perform a local assembly by CAP3 [29].

HPV Fusion Point Calling

For each integration event, the assembly algorithm was able to report multiple contigs. We developed a procedure to evaluate and select contigs for each integration event to call HPV fusion point more precisely. First, we aligned the contigs against the human genome and HPV genome separately by BWA mem. If the contig met the following criteria, we marked it as high confidence:

(1) Has at least 10 supportive reads

(2) $10\% < \frac{matched\ length\ of\ the\ contig\ to\ HPV}{length\ of\ contig} < 95\%$

Then we separated the contigs we assembled into two classes: from left side (Contig A in **Fig 4.1B**) and from right side (Contig B in **Fig 4.1B**). For each class, if there were high confidence contigs in the class, we selected the contig with maximum length among them. If the class has only low confidence contigs, we selected the contig with most supportive reads. For each insertion event, we reported one contig if it only had contigs from one side and we reported two contigs if it had contigs from both sides (**Figure 4.1C**). Finally, we identified the fusion points within HPV based on the alignment results of the selected contigs against the HPV genome. The bam/sam file processing in this pipeline was done by Samtools[25] and the analysis was performed with R 3.6.1 [30] and Python [31].

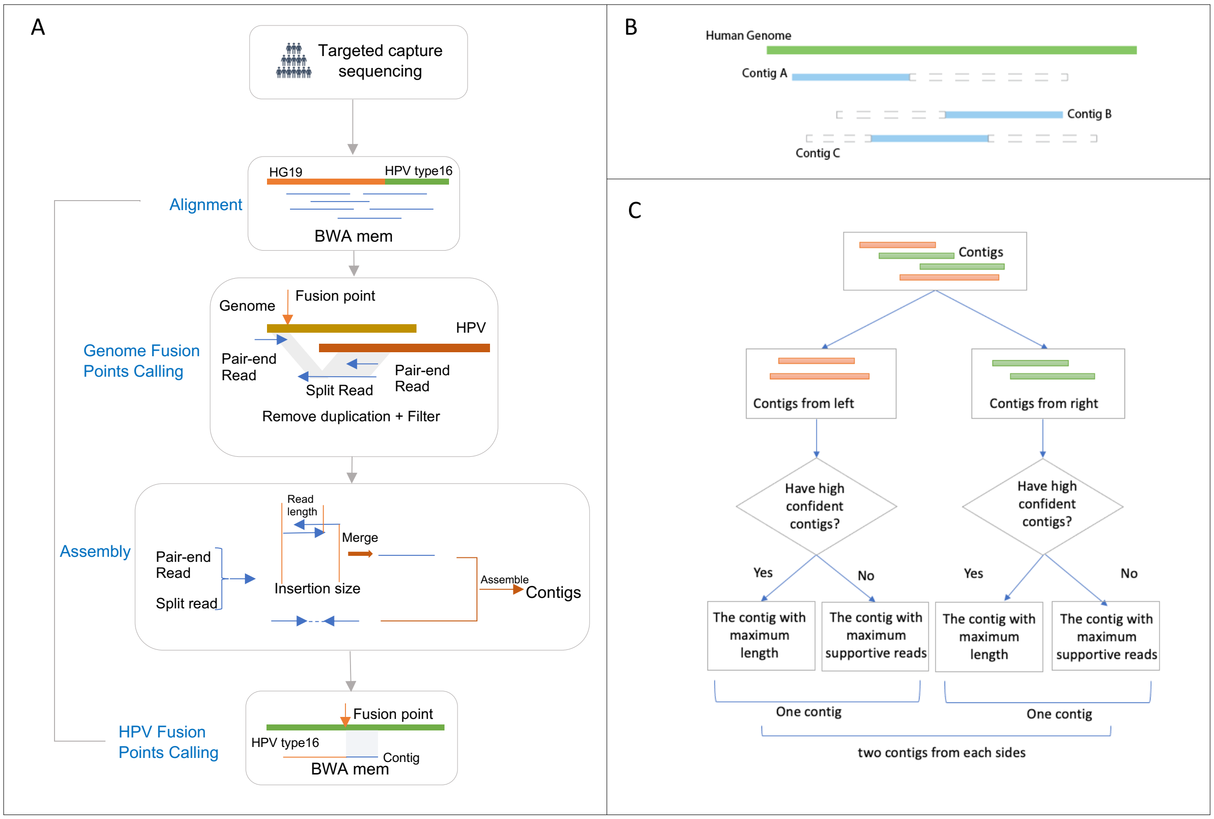

A
Targeted capture sequencing
Alignment
HG19
HPV type16
BWA mem
Genome Fusion Points Calling
Genome
Fusion point
HPV
Pair-end Read
Pair-end Read
Split Read
Remove duplication + Filter
Assembly
Read length
Merge
Pair-end Read
Split read
Insertion size
Assemble
Contigs
HPV Fusion Points Calling
Fusion point
HPV type16
Contig
BWA mem
B
Human Genome
Contig A
Contig B
Contig C
C
Contigs
Contigs from left
Contigs from right
Have high confident contigs?
Have high confident contigs?
Yes
No
Yes
No
The contig with maximum length
The contig with maximum supportive reads
The contig with maximum length
The contig with maximum supportive reads
One contig
One contig
two contigs from each sides

Figure 4.1: Workflow of SearcHPV. (A) Paired-end reads from targeted capture sequencing were aligned to a catenated human-HPV reference genome. After filtering, fusion points were identified by split reads and paired-end reads. Informative reads were extracted for local assembly. Read pairs that have overlaps were merged first before assembly. Assembled contigs were aligned to HPV genome to identify the breakpoints on HPV. (B) Contigs were divided to two classes. Contig A would be assigned to left group and Contig B would be assigned to right group. Contig C would be randomly assigned to left or right group. (C) Workflow for the contig selection procedures for fusion point with multiple candidates contigs. For each fusion point, we reported at least one contig and at most two contigs representing two directions.

Other Integration Callers: We installed and ran two other integration callers VirusSeq and VirusFinder2 as described below.

VirusSeq

We installed and ran VirusSeq following the user guide using the default parameter settings. We first installed the MOSAIK aligner [32]. As VirusSeq did not have a feature for users to customize the reference genome, we indexed and used the provided built-in reference database (GIB-V [33], hg19 and hybrid reference genome concatenated by hg19 and 17 viral genomes). VirusSeq aligned the paired end reads to the catenated reference genome by MOSAIK and extracted and clustered split reads using a Perl script (Spanner_cross_converter.pl). Integration sites were detected by another Perl script (VirusSeq_Integration.pl).

VirusFinder2

To install VirusFinder2, we first installed all the third-party tools and Perl modules required by VirusFinder2. As required by VirusFinder2, we indexed the human reference genome (hs37d5) by Bowtie2 [34] and Blast+ [35], as well as the virus database [36] suggested by VirusFinder2 using Blast+. VirusFinder2 used Bowtie2 to align raw reads against the human reference genome. The informative reads were extracted and assembled to contigs using Trinity [37]. By mapping contigs to the virus reference database, VirusFinder2 detected the virus and identified the virus type. It then applied the VERSE algorithm to customize the reference genome.

Sanger Sequencing: Primer sets (n=46) were designed to amplify across the predicted HPV-human junctions from the contigs generated by the integration callers (**Table S4.1**). Primers were designed using NCBI Primer-BLAST. PCR was performed using each of the 46 primer sets multiplexed with *GAPDH* control primers using 50 ng DNA and Platinum Taq (Thermo Fisher Scientific, Waltham, MA) following the manufacturer’s instructions. The PCR products were run by gel electrophoresis on a 1.5% agarose gel, followed by isolation of DNA from the bands at the predicted molecular weights using the Qiaquick Gel Extraction Kit (Qiagen, Hilden, Germany). These products were sent for Sanger sequencing at Eurofins Genomics (Louisville, KY) and mapped back to the predicted sequence to confirm sequence identity.

Characterization of Integration Calls: Circos plots detailing integration sites were generated using the Circlize package [38] in R 3.6.1 [30]. Distance of each integration site from genes was calculated based on NCBI RefSeq genes (Release 105.20190906). Microsatellite repeats (2-6 bp in length, minimum of 3 repeats) were detected using the Tandem Repeats Finder (http://insilico.ehu.es/mini_tools/microsatellites/?info) [39].

10X Linked Reads Sequencing: High-molecular weight DNA from UM-SCC-47 and PDX-294R was submitted to the University of Michigan Advanced Genomics Core for 10x-based linked read library generation and sequencing on an Illumina NovaSeq6000 with 300nt paired end run. Samples were de-multiplexed and FastQ files with matched index files were generated using Long Ranger Version 2.2.2. Data was visualized using the Loupe software package, Version2.1.1 (2.4). Structural variation calls were considered high confidence if they occurred in unambiguous regions of the reference genome and there were 3 or more supporting sequencing barcodes detected at the site. The raw data was deposited to the sequencing read archives under identification number: PRJNA668771.

Results

SearcHPV pipeline

Viral integration has traditionally been detected using whole genome sequencing data, but these events are relatively rare in the genome, so a targeted approach is helpful to enrich for these events to improve coverage of these regions. HPV targeted capture sequencing allows for deeper investigation of these events, but the current bioinformatics pipelines available are not designed for this type of data. Given the limitations with previous sequencing approaches and their associated viral integration callers, we set out to design a new targeted sequencing-based pipeline to improve HPV integration calling in HNSCC samples. A schematic of our pipeline which we termed “SearcHPV” is shown in **Figure 4.1**. Two HPV16+ HNSCC models, UM-SCC-47 and PDX-294R, were subjected to targeted-capture based Illumina sequencing using a custom panel of probes spanning the entire HPV16 genome, L1 and L2 of HPV18/33/35, and over two hundred human genes known to be frequently altered in HNSCC. The paired end reads for each sample then went through the four steps of analysis of SearcHPV: alignment to custom reference genome, genome fusion points calling, local assembly and precise fusion point calling. Analysis of the integration sites in the cell line and PDX models using our pipeline SearcHPV showed a high frequency of HPV16 integration with a total of six events in UM-SCC-47 and sixty-nine unique events in PDX-294R (**Figure 4.2, Table S4.2-S4.3**).

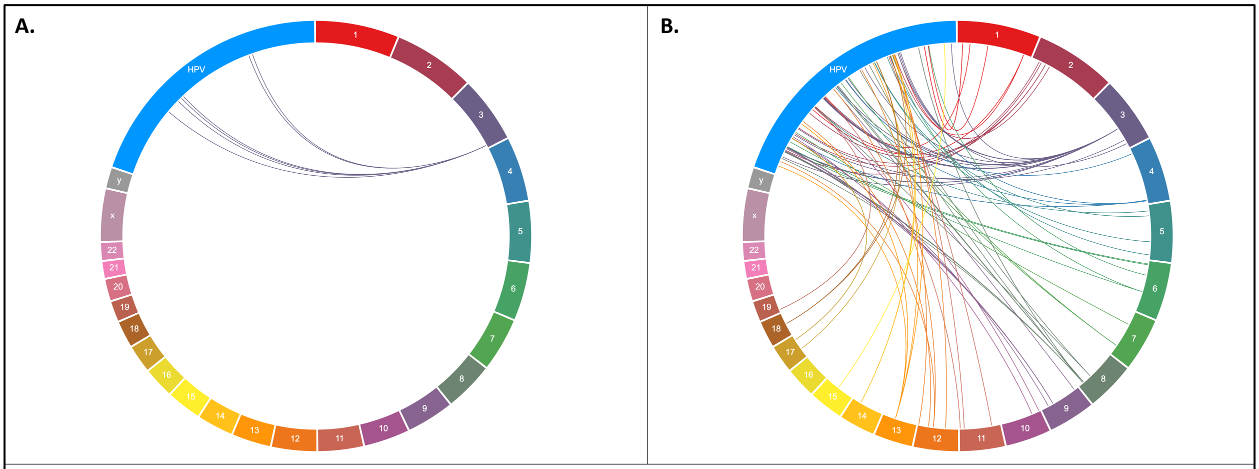


Figure 4.2: Distribution of breakpoints in the human and HPV genomes called by SearcHPV. Links of breakpoints in the human and HPV16 genomes for (A) UM-SCC-47 and (B) PDX-294R.

Comparison to other integration callers and confirmation of integration sites

In addition to using SearcHPV, we assessed UM-SCC-47 and PDX-294R for HPV integration events using two previously established integration callers, VirusFinder2 and VirusSeq (**Figure 4.3**). We found that SearcHPV called HPV integration at a much higher rate than either previous caller. There were a large number of sites that were only identified by SearcHPV (n=49), although there were also sites that were identified by two or more callers (n=26). VirusFinder2 and VirusSeq also had a number of sites (n=20 and n=8, respectively) that were not detected by our pipeline. In order to assess the accuracy of each caller, we performed PCR on source genomic DNA followed by Sanger sequencing with primers spanning the HPV-human junction sites predicted by either SearcHPV, VirusFinder2 and/or VirusSeq (**Figure S4.1**). We were able to test a total of forty-six integration sites using this method, twenty-five of which were unique to SearcHPV and eight which were unique to VirusSeq. VirusFinder2 does not allow for local assembly of the integration junctions which rendered us unable to test sites that were unique to this program. SearcHPV had an overall confirmation rate of 27/38 (71%), with the confirmation rate for sites unique to SearchHPV showing slightly higher (18/25 (72%)). In contrast, the overall confirmation rate for VirusSeq was 7/14 (50%), with only a 2/8 (25%) success rate for sites unique to VirusSeq. The sites that were identified by all three integration callers had the highest confirmation rate (4/4 (100%)). The confirmation rate of high confidence SearcHPV sites was higher than that for low confidence sites (23/31 (74%) versus 4/7 (57%)).

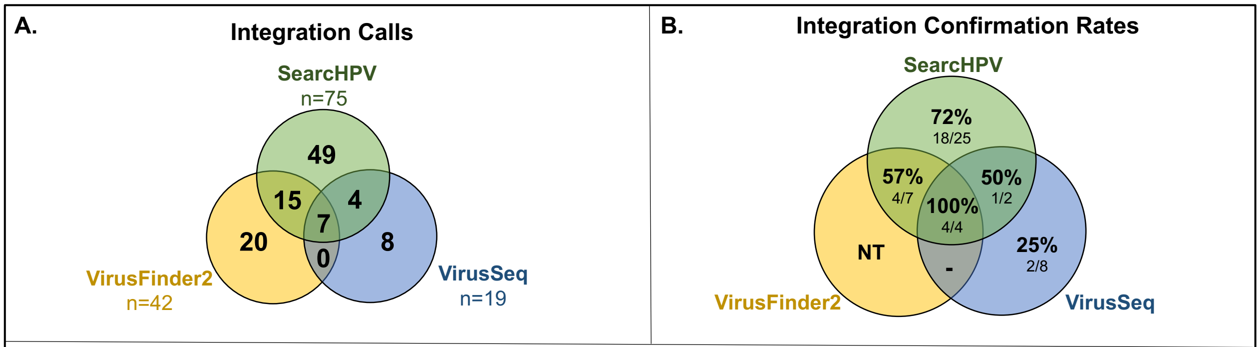


Figure 4.3: Comparison of integration sites called by SearcHPV, VirusSeq and VirusFinder2. (A) Number of integration sites called by each program. (B) PCR confirmation rate of sites called by each program.

Localization of integration sites

We next examined the integration sites detected in the HNSCC models by SearcHPV. We identified a large number of integration events, with six called in UM-SCC-47 and sixty-nine unique HPV integration sites called in PDX-294R **(Figure 4.4)**.

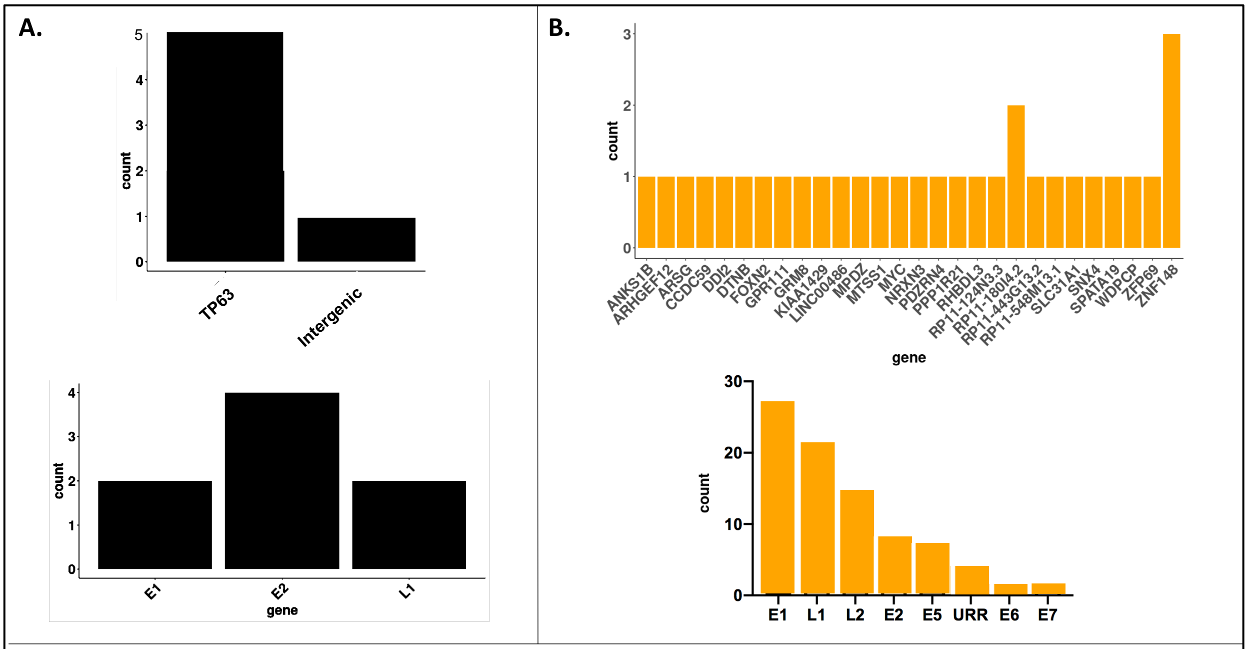


Figure 4.4: Quantification of breakpoint calls in human and HPV16 genes for (A) UM-SCC-47 and (B) PDX-294R.

The six integration sites discovered in UM-SCC-47 were clustered on chromosome 3q28 within/near the cellular gene *TP63* and either involved the HPV16 genes E1, E2 or L1. Three integration sites were called within intron 10 of *TP63*, and there was one integration each in intron 12 and exon 14. One integration site was 8.6 kilobases (kb) downstream of the *TP63* coding region.

Within PDX-294R, HPV16 integration sites were identified across 18 different chromosomes (chromosomes 1-15, 17-19), occurring most frequently on chromosome 3. The most frequently involved HPV genes were E1 (21/ 69 (31%)) and L1 (18/69 (26%)). Most of the integration sites discovered in this sample mapped to within/near (<50 kb) a known cellular gene (45/69 (66%)). Of the sites that fell within a gene, the majority of integrations took place within an intronic region (35/41 (85%)) with only a small number of events occurring within a gene promoter or exon. Although the integration sites were scattered throughout the human genome, we saw some examples of loci that contained multiple integration sites closely clustered around cancer-relevant genes, including *ZNF148* and *SNX4* on chromosome 3q21.2, *MYC* on chromosome 8q24.21 and *FOXN2* on chromosome 2p16.3.

Association of integration sites and large-scale duplications

We predicted that the complex integration sites we discovered in UM-SCC-47 and PDX-294R would be associated with large-scale structural alterations of the genome, such as rearrangements, deletions and duplications. To identify these alterations, we subjected UM-SCC-47 and PDX-294R to 10X linked-read sequencing. We generated over 1 billion reads for each sample (**Table S4.4**), with phase blocks (contiguous stretches of DNA from the same allele) of up to 28.9M and 3.8M bases in length for UM-SCC-47 and PDX-294R, respectively (**Figure S4.2**). This led to the identification of 444 high confidence large structural events in UM-SCC-47 and 126 events in the PDX-294R model. Upon performing integrated analysis with our SearcHPV results, a 130 kb duplication surrounding the integration events in *TP63* in UM-SCC-47 was discovered (**Figure 4.5A**). Similarly, in PDX-294R, 32/69 (46%) integration sites were within a region that contained a large-scale duplication, while the other 37 integration events fell outside regions of large structural variation. This suggested that in this PDX model, 32/126 (25%) large structural events were potentially induced during HPV integration. For example, the clusters of integration events surrounding *ZNF148* and *SNX4*, *MYC*, as well as *FOXN2* were all associated with large genomic duplications in PDX-294R (**Figure 4.5B-D**).

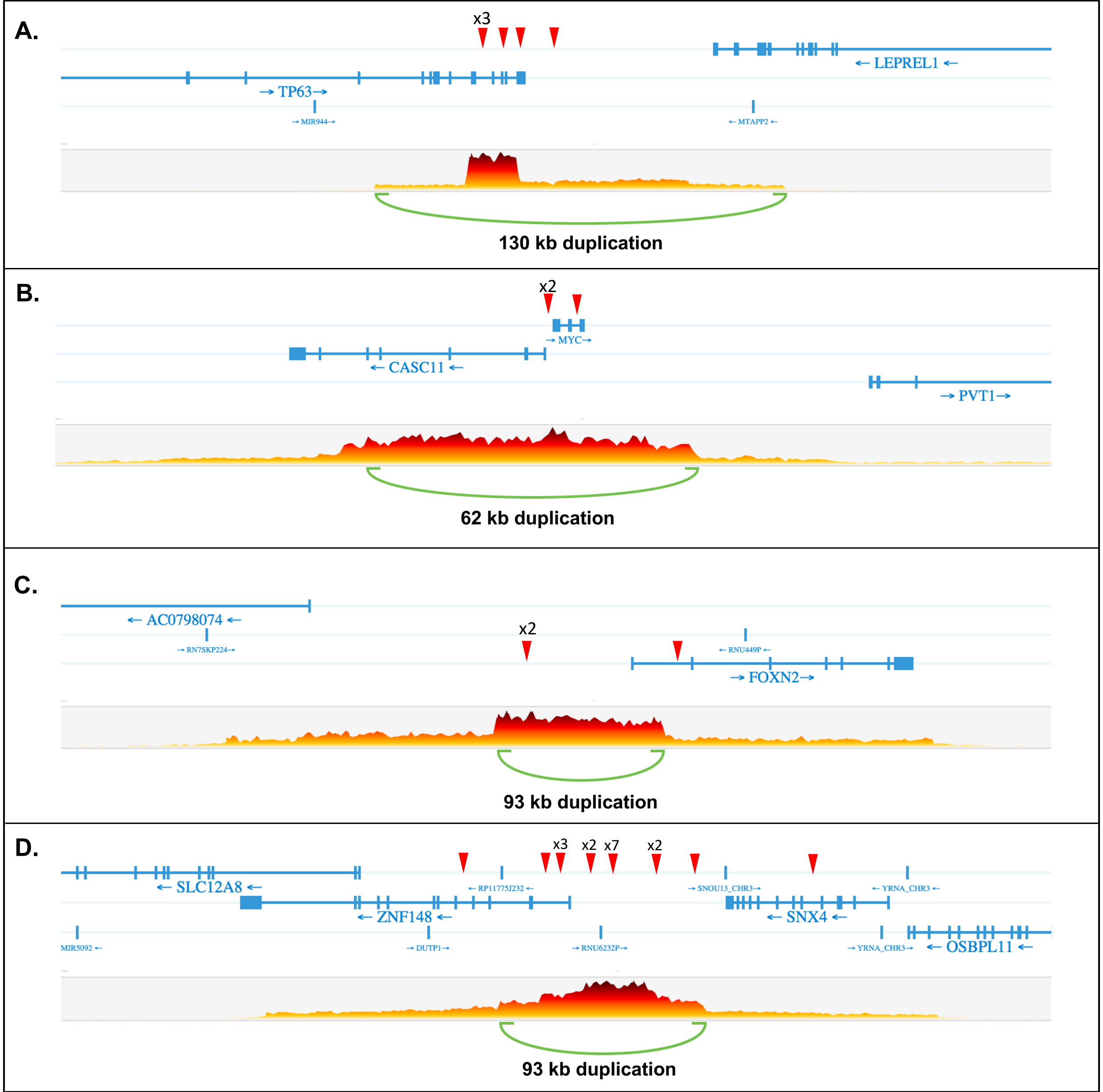


Figure 4.5: Genomic duplications associated with HPV integration in UM-SCC-47 (A) and PDX-294R (B-D). Red arrows indicate integration site. Each plot shows the number of overlapping barcodes observed in sequencing reads of that region.

Descriptive analysis at junction sites

Finally, we were interested to see whether the junction sites called in UM-SCC-47 and PDX-294R by SearcHPV followed any patterns in terms of 1) direction of the HPV genome in relation to the host genome, 2) the presence of microsatellite repeats and 3) microhomology between the genomes. The direction HPV16 integrated relative to the host genome appeared to be stochastic; HPV read in the same and opposite direction as the human genome at approximately equal rates (49% and 51%, respectively). HPV integration into microsatellite repeats within the human genome was a relatively rare event and only occurred in a few cases (6/69 (8%)).

To evaluate microhomology, we examined the degree of sequence overlap at the junction site. We saw three types of junction points: those with a gap of unmapped sequence between the human and HPV genomes, those that had a clean breakpoint between the genomes, and those with sequence that could be mapped to both the human and HPV16 genomes (**Figure 4.6A**). The majority of sites in both samples had at least some degree of microhomology (56%) (**Figure 4.6B-C**). Integration sites with clean breaks (0 bp overlap) and 3 bp of overlap were the most frequently seen junctions in PDX-294R, but there was a wide range of levels seen, going up to 17 bp of overlap. There was also a large number of junctions with gaps between the human and HPV genomes ranging from 1 - 48 bp long.

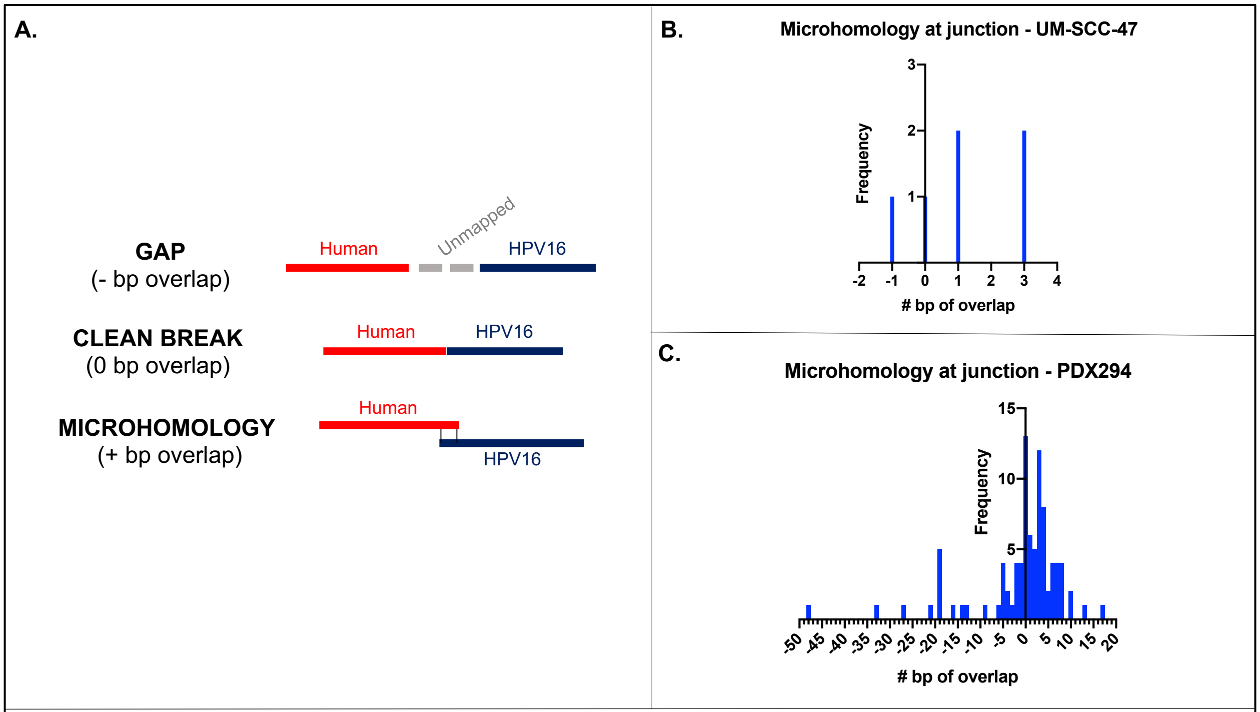


Figure 4.6: Microhomology at junction points. (A) The three types of junction points. (B) Level of microhomology (in bp) in UM-SCC-47. (C) Level of microhomology (in bp) in PDX-294R. Junctions with a gap are shown as negative numbers.

Discussion

To interrogate HPV integration sites through targeted capture sequencing data, we developed a novel bioinformatics pipeline that we termed "SearcHPV" and preliminarily show that it operated in a more accurate and efficient manner than existing pipelines. The software also has the advantage of performing local contig assembly around the junction sites, which simplifies downstream confirmation experiments. We used our new caller to interrogate the integration sites found in two HNSCC models in order to compare the accuracy of our caller to the existing pipelines. We then evaluated the genomic effects of these integrations on a larger scale by 10X linked-reads sequencing and performed an integrated analysis of the capture-based and whole genome data sets to identify the role of HPV integration in driving structural variation in the tumor genome.

Using SearcHPV, we were able to investigate the HPV-human integration events present in UM-SCC-47 and PDX-294R. Importantly, UM-SCC-47 has been previously assessed for HPV integration by our group and others using a variety of methods [8, 20, 21, 40, 41], which we leveraged as ground truth knowledge to validate our integration caller. All previous studies were in agreement that HPV16 is integrated specifically within the cellular gene *TP63*, although the exact number of sites and locations within the gene varied by study. In this study, SearcHPV also called HPV integration sites within *TP63*. We found integrations of E1, E2 and L1 within *TP63* intron 10, L1 within intron 12 and E2 within *TP63* exon 14. These integration sites were also detected using DIPS-PCR [21] and/or WGS [8] with the exception of E1 into intron 10, which was unique to our caller and confirmed by direct PCR followed by Sanger sequencing. It is possible that the integration sites detected in this sample represent multiple fragments of one larger integration site. There were additional sites called by other WGS studies that we did not detect (intron 9 [8] and exon 7 [20]), although it is possible that alternate clonal populations grew out due to different selective pressures in different laboratories. Nonetheless, the analysis clearly demonstrated that SearcHPV was able to detect a well-established HPV insertion site.

In contrast to UM-SCC-47, to our knowledge, PDX-294R has not been previously analyzed for viral-host integration sites and therefore represented a true discovery case in our study. We identified widespread HPV integration sites throughout the host human genome and also observed that 66% of integration sites were found within or near genes (<50 kb). This aligns with previous reports that integrations are detected in host genes more frequently than expected by chance [2, 3, 7, 42]. Further, we identified several integration events at or near previously established cancer-related genes, including *MYC*. Importantly, *MYC* has also been identified as a potential hotspot for HPV integration [7, 43] and the junctions we detected in/near this gene had 2-4 bp of homology, potentially driving this observation. Accordingly, an HPV-integration related promoter duplication event, which may be expected to drive expression, would be consistent with a novel genetic mechanism to drive expression of this oncogene.

TP63 has also been reported to be a hotspot for HPV integration, as it has been recorded in multiple samples besides UM-SCC-47 [3, 7, 44, 45]. There is a high degree of microhomology between HPV16 and this gene [44], and the junctions we found within TP63 mostly had 1-3 bp of microhomology, again serving as a possible mechanism for frequent integration here. Given the high frequency of molecular alterations in the epidermal differentiation pathway (e.g. *NOTCH1/2*, *TP63* and *ZNF750*) in HPV+ HNSCCs, this data supports HPV integration as a pivotal mechanism of viral-driven oncogenesis in this model [46].

HPV integration sites have been associated with structural variations in the human genome and have been found at regions of amplification or deletion [3, 8, 46], which may support the selective advantage of integration into/adjacent to host cancer-related genes. This structural variation event is thought to be due to the rolling circle amplification that takes place at the integration breakpoint, leading to the formation of amplified segments of genomic sequence flanked by HPV segments [8, 47]. These structural alterations are frequently associated with changes in gene expression [3]. Our data are consistent with these previous reports in that approximately half of the integration events we discovered were associated with a large-scale amplification. Accordingly, in UM-SCC-47, the integration sites within *TP63* were also associated with a large-scale amplification. It is unclear why some integration sites were associated with structural variants and others were not, but it is possible that at some points in the genome, HPV integrated by an alternative mechanism to rolling circle amplification as has been previously described [47]. It is also unclear how these large amplifications may affect gene expression in these samples, as we did not evaluate this in the current study.

Importantly, this observation that HPV integration events tended to be enriched in cellular genes could be due to multiple different mechanisms. Integration could occur preferentially in regions of open chromatin during cell replication and keratinocyte differentiation. Other potential mechanisms are: 1) that HPV integration is directed to specific host genes by homology or 2) that HPV integration is random, but events that are advantageous for oncogenesis are clonally selected and expanded, and we would postulate that the later mechanism may be enriched for non-homology based DNA repair mechanisms. Therefore, to help resolve differences in the mechanism of integration, we assessed microhomology at the HPV-human junction points. Early in the HPV literature, it was described that HPV integration may be targeted to chromosomal fragile sites where DNA double strand breaks are unrepaired [42, 48, 49], but it is still unclear at this point how DNA damage repair pathways play a role in resolving these breakpoints. We saw that the majority of breakpoints had at least some level of microhomology, ranging from 1-17 bp of overlap. The most frequent levels of overlap were 0 and 3 bp, which potentially implicates non-homologous end joining (NHEJ) in repair at these sites, since this pathway most frequently results in 0-5 bp of overlap [50]. There were also a number of junction sites that demonstrated a gap of inserted sequence between the HPV and human genomes. It has been described that during polymerase theta-mediated end joining (TMEJ), stretches of 3-30 bp are frequently inserted at the site of repair, possibly accounting for the sites we saw with a gap between genomes [51]. However, given the relatively small number of events we examined, we expect that future analysis with our pipeline will be able to help resolve the specific role of each DNA repair pathway in HPV-human fusion breakpoints.

Overall, our new HPV detection pipeline SearcHPV overcomes a gap in the field of viral-host integration analysis. We recognize that the performance of SearcHPV has only been examined on two HPV+ HNSCC models, and we were only able to compare its sensitivity and accuracy to other programs based on a relatively small number of overall events. Therefore, we are unable to determine statistical significance differences in the accuracy and sensitivity of our caller. However, based on our preliminary findings, we demonstrated a trend that our caller was more accurate. In the future, cohort-based studies of HPV+ HNSCC samples with similarly rigorous validation will further our understanding of the sensitivity of the software. Most importantly, we expect that the application of this pipeline in large HPV+ cancer tissue cohorts will also help advance our understanding of the potential oncogenic mechanisms associated with viral integration-based oncogenesis. Indeed, with the emerging set of tools such as SearcHPV that are rapidly becoming available for different types of next generation sequencing data, we believe the field is now primed to make major advances in the understanding of HPV-driven pathogenesis, some of which may lead to the development of novel biomarkers and/or treatment paradigms.

References

1. Gao, G., et al., *Whole genome sequencing reveals complexity in both HPV sequences present and HPV integrations in HPV-positive oropharyngeal squamous cell carcinomas.* BMC Cancer, 2019. **19**(1): p. 352.
2. Nulton, T.J., et al., *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**(11): p. 17684-17699.
3. Parfenov, M., et al., *Characterization of HPV and host genome interactions in primary head and neck cancers.* Proc Natl Acad Sci U S A, 2014. **111**(43): p. 15544-9.
4. Pinatti, L.M., et al., *Association of human papillomavirus integration with better patient outcomes in oropharyngeal squamous cell carcinoma.* Head Neck, 2020.
5. Tian, R., et al., *Risk stratification of cervical lesions using capture sequencing and machine learning method based on HPV and human integrated genomic profiles.* Carcinogenesis, 2019. **40**(10): p. 1220-1228.
6. McBride, A.A. and A. Warburton, *The role of integration in oncogenic progression of HPV-associated cancers.* PLoS Pathog, 2017. **13**(4): p. e1006211.
7. Bodelon, C., et al., *Genomic characterization of viral integration sites in HPV-related cancers.* Int J Cancer, 2016. **139**(9): p. 2001-11.
8. Akagi, K., et al., *Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability.* Genome Res, 2014. **24**(2): p. 185-99.
9. Pinatti, L.M., H.M. Walline, and T.E. Carey, *Human Papillomavirus Genome Integration and Head and Neck Cancer.* J Dent Res, 2018. **97**(6): p. 691-700.
10. Luft, F., et al., *Detection of integrated papillomavirus sequences by ligation-mediated PCR (DIPS-PCR) and molecular characterization in cervical cancer cells.* Int J Cancer, 2001. **92**(1): p. 9-17.
11. Klaes, R., et al., *Detection of high-risk cervical intraepithelial neoplasia and cervical cancer by amplification of transcripts derived from integrated papillomavirus oncogenes.* Cancer Res, 1999. **59**(24): p. 6132-6.
12. Wang, Q., P. Jia, and Z. Zhao, *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**(5): p. e64465.
13. Wang, Q., P. Jia, and Z. Zhao, *VERSE: a novel approach to detect virus integration in host genomes through reference genome customization.* Genome Med, 2015. **7**(1): p. 2.
14. Chen, Y., et al., *VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue.* Bioinformatics, 2013. **29**(2): p. 266-7.
15. Holmes, A., et al., *Mechanistic signatures of HPV insertions in cervical carcinomas.* NPJ Genom Med, 2016. **1**: p. 16004.
16. Montgomery, N.D., et al., *Identification of Human Papillomavirus Infection in Cancer Tissue by Targeted Next-generation Sequencing.* Appl Immunohistochem Mol Morphol, 2016. **24**(7): p. 490-5.
17. Morel, A., et al., *Mechanistic Signatures of Human Papillomavirus Insertions in Anal Squamous Cell Carcinomas.* Cancers (Basel), 2019. **11**(12).
18. Nkili-Meyong, A.A., et al., *Genome-wide profiling of human papillomavirus DNA integration in liquid-based cytology specimens from a Gabonese female population using HPV capture technology.* Sci Rep, 2019. **9**(1): p. 1504.

19. Brenner, J.C., et al., *Genotyping of 73 UM-SCC head and neck squamous cell carcinoma cell lines.* Head Neck, 2010. **32**(4): p. 417-26.
20. Olthof, N.C., et al., *Viral load, gene expression and mapping of viral integration sites in HPV16-associated HNSCC cell lines.* Int J Cancer, 2015. **136**(5): p. E207-18.
21. Walline, H.M., et al., *Integration of high-risk human papillomavirus into cellular cancer-related genes in head and neck cancer cell lines.* Head Neck, 2017. **39**(5): p. 840-852.
22. Heft Neal, M.E., et al., *Prognostic Significance of Oxidation Pathway Mutations in Recurrent Laryngeal Squamous Cell Carcinoma.* Cancers (Basel), 2020. **12**(11).
23. NIAID. *Papillomavirus Episteme.* [cited 2020; Available from: https://pave.niaid.nih.gov/.
24. Van Doorslaer, K., et al., *The Papillomavirus Episteme: a major update to the papillomavirus sequence database.* Nucleic Acids Res, 2017. **45**(D1): p. D499-D506.
25. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
26. Institute, B., *Picard toolkit.* Broad Institute GitHub Repository, 2019.
27. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Res, 2010. **20**(9): p. 1297-303.
28. Zhang, J., et al., *PEAR: a fast and accurate Illumina Paired-End reAd mergeR.* Bioinformatics, 2014. **30**(5): p. 614-20.
29. Huang, X. and A. Madan, *CAP3: A DNA sequence assembly program.* Genome Res, 1999. **9**(9): p. 868-77.
30. Team, R.C., *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, 2019.
31. Van Rossum, G., Drake F.L., *Python 3 Reference Manual: Python Documentation Manual Part 2.* CreateSpace Independent Publishing Platform., 2009.
32. Lee, W.P., et al., *MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping.* PLoS One, 2014. **9**(3): p. e90581.
33. Hirahata, M., et al., *Genome Information Broker for Viruses (GIB-V): database for comparative analysis of virus genomes.* Nucleic Acids Res, 2007. **35**(Database issue): p. D339-42.
34. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.
35. Camacho, C., et al., *BLAST+: architecture and applications.* BMC Bioinformatics, 2009. **10**: p. 421.
36. Bhaduri, A., et al., *Rapid identification of non-human sequences in high-throughput sequencing datasets.* Bioinformatics, 2012. **28**(8): p. 1174-5.
37. Grabherr, M.G., et al., *Full-length transcriptome assembly from RNA-Seq data without a reference genome.* Nat Biotechnol, 2011. **29**(7): p. 644-52.
38. Gu, Z., et al., *circlize Implements and enhances circular visualization in R.* Bioinformatics, 2014. **30**(19): p. 2811-2.
39. Benson, G., *Tandem repeats finder: a program to analyze DNA sequences.* Nucleic Acids Res, 1999. **27**(2): p. 573-80.
40. Khanal, S., et al., *Viral DNA integration and methylation of human papillomavirus type 16 in high-grade oral epithelial dysplasia and head and neck squamous cell carcinoma.* Oncotarget, 2018. **9**(54): p. 30419-30433.

41. Myers, J.E., et al., *Detecting episomal or integrated human papillomavirus 16 DNA using an exonuclease V-qPCR-based assay.* Virology, 2019. **537**: p. 149-156.

42. Hu, Z., et al., *Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism.* Nat Genet, 2015. **47**(2): p. 158-63.

43. Ferber, M.J., et al., *Preferential integration of human papillomavirus type 18 near the c-myc locus in cervical carcinoma.* Oncogene, 2003. **22**(46): p. 7233-42.

44. Schmitz, M., et al., *Non-random integration of the HPV genome in cervical cancer.* PLoS One, 2012. **7**(6): p. e39632.

45. Walline, H.M., et al., *Genomic Integration of High-Risk HPV Alters Gene Expression in Oropharyngeal Squamous Cell Carcinoma.* Mol Cancer Res, 2016. **14**(10): p. 941-952.

46. Cancer Genome Atlas, N., *Comprehensive genomic characterization of head and neck squamous cell carcinomas.* Nature, 2015. **517**(7536): p. 576-82.

47. Groves, I.J. and N. Coleman, *Human papillomavirus genome integration in squamous carcinogenesis: what have next-generation sequencing studies taught us?* J Pathol, 2018. **245**(1): p. 9-18.

48. Dall, K.L., et al., *Characterization of naturally occurring HPV16 integration sites isolated from cervical keratinocytes under noncompetitive conditions.* Cancer Res, 2008. **68**(20): p. 8249-59.

49. Thorland, E.C., et al., *Common fragile sites are preferential targets for HPV16 integrations in cervical tumors.* Oncogene, 2003. **22**(8): p. 1225-37.

50. Pannunzio, N.R., et al., *Non-homologous end joining often uses microhomology: implications for alternative end joining.* DNA Repair (Amst), 2014. **17**: p. 74-80.

51. Carvajal-Garcia, J., et al., *Mechanistic basis for microhomology identification and genome scarring by polymerase theta.* Proc Natl Acad Sci U S A, 2020. **117**(15): p. 8476-8485.

Chapter 5 Discussion

Summary

Human papillomaviruses (HPV) are implicated in the development of a number of cancers, including head and neck squamous cell carcinomas (HNSCC), particularly of the oropharynx (OPSCC). One particular process that has been recently investigated as a potential driver of HPV-related cancers is viral integration into the human host genome, as this is frequently observed in tumors but is not required for the lifecycle of this DNA virus [1-6]. In **Chapter 1**, we explored the reported mechanisms of this process and how it affects the HPV/human genomes and transcriptomes. The leading model of HPV integration to explain the genomic instability associated with this process is described as looping amplification that results the formation of concatemers consisting of amplified segments of human sequence flanked by HPV segments [7]. This rolling circle amplification results in disruption of viral regulatory regions, such as E2 and E1. This potentially results in overexpression of the viral oncogenes E6 and E7, specifically alternate transcripts of E6 known as E6* that are associated with a more aggressive phenotype [8-11]. It can also result in disruption of adjacent host genes and lead to large scale structural variants, including deletions, amplifications and rearrangements [5, 7]. These genome-level changes can then impact the tumor transcriptome; gene expression differences have been reported for genes affected directly by integration, but transcriptome-wide changes have also been reported [5, 7, 11, 12]. Despite all that has been reported about HPV integration, how exactly this process drives oncogenesis of HPV+ OPSCCs is still unclear.

Our group has previously assessed HPV integration sites in patient specimens but only in a limited cohort of ten patients [13]. Therefore, in **Chapter 2**, we aimed to examine these events in a larger set of patients (n=36) in order to investigate the relationship between HPV integration and 1) HPV oncogene expression and 2) clinical outcome. Previous work investigating the relationship between HPV integration and HPV oncogene expression has shown conflicting results. It has long been thought that the process of HPV integration leads to an increase in HPV E6/E7 expression due to the loss of the transcriptional repressor E2 [8]. This unregulated expression of E6/E7 can lead to increased proliferation and genomic instability due to their interaction with cell cycle control proteins TP53 and RB, respectively [14, 15]. However, others have reported that not all tumors with HPV integration showed increased expression of E6 and E7 [5, 16].

In our cohort of thirty-six patients, we assessed HPV integration status by Detection of Integrated Papillomavirus Sequences polymerase chain reaction (DIPS-PCR) and expression of the HPV oncogenes E6 and E7 by both quantitative and non-quantitative reverse transcription polymerase chain reaction (qRT-PCR and RT-PCR) to show expression of E6* variants. Our results demonstrated a high frequency of HPV integration similar to that of other reports (60% of samples). We saw a significant positive correlation between HPV integration and E6E7 expression level as measured by qRT-PCR. When we looked at the association between integration and the expression of either E6FL-E7 or E6*-E7 transcripts, integration was only significantly associated with the expression of E6FLE7 transcripts. This indicated to us that in most cases HPV integration does lead to increased E6/E7 expression, which likely contributes to increased genomic instability. However, there was no difference in survival between patients with high versus low expression of E6E7 in our cohort.

In the study of these thirty-six patients, we also aimed to investigate the relationship between HPV integration and clinical outcome, specifically disease-specific survival (DSS). A handful of previous studies have tried to assess whether patients with HPV integration have a worse outcome, mainly measuring integration status by E2 loss, although one group used the presence or absence of viral-cellular fusion transcripts to distinguish between integration positive versus negative [17-20]. One group reported no significant difference in survival between the two groups [20], but others reported that integration-positive patients as assessed by viral-host fusion transcripts had a significantly worse outcome, suggesting that this process may act as an additional mechanism of carcinogenesis.

When we examined integration-positive versus integration-negative patients in our cohort, our results actually showed the opposite; integration positive patients had a significantly improved DSS. We also tried to separate integration-positive patients into two subcategories, those with integration into gene-poor regions of the genome and those with integration into genes, given that integration into genes may impact the tumor transcriptome. In our cohort, there was no significant difference in survival between these groups, likely due to small sample sizes. The reason for the discrepancy in results between studies is not entirely clear, but the method for integration detection likely contributes to this issue. One possible hypothesis to explain our results is that the high levels of genomic instability due to HPV integration somehow leads to a better response to therapy in patients, potentially through an improved immune response.

In **Chapter 3**, we leveraged the stochastic nature of HPV integration to investigate the clonality of bilateral tonsillar carcinomas. These rare tumors only occur in about 3% of OPSCC patients, but the mechanism behind their formation is of great clinical interest [21, 22]. When a patient presents with HPV+ tumors in both tonsils, it is unclear whether those tumors formed independently due to exposure to one or more HPV types or if there was one original tumor that migrated across the midline into the other tonsil. There have been reports that support both mechanisms, but in general, this is an understudied process [23-25]. It would not be entirely surprising for a clonal population to establish a secondary tumor given the highly metastatic nature of HPV+ OPSCC which frequently is observed in the surrounding lymph nodes [26, 27]. However, there is not currently a good understanding of how that may occur. Investigation of these tumors also provides the opportunity to think more broadly about how and why HPV integrations may contribute to the fitness of cancer cell clones.

In order to investigate the clonality of bilateral tonsillar carcinomas, we assessed the HPV genotypes and HPV integration sites found within three pairs of tumors. We would expect clonal tumors to share HPV genotypes and HPV integration sites. Given that HPV integration rarely occurs in the same locus from sample to sample, we would expect that unrelated tumors would not share integration sites. In two out of three patients (Patients 2 and 3), HPV16 was the only HPV type present in each tumor. In the remaining patient (Patient 1), both tumors shared an HPV16 infection, but they also had an additional discordant HPV type (HPV31 and HPV33). When the HPV16 integration events were assessed by DIPS-PCR in all samples, nearly identical HPV integration sites were detected between the tumor pairs in Patients 2 and 3, but Patient 1 had only discordant HPV integration sites. The tumor pairs from Patient 2 showed HPV16 E2 integration into the exact same intergenic locus on chromosome 4p15. The tumor pairs from Patient 3 showed integration into the same cellular gene, *CD36*, but the HPV gene and exact site within *CD36* differed slightly.

These data suggest that Patient 1's tumors were likely independently formed, given the discordant HPV types and integration sites. However, it is also possible that these tumors were in fact clonally derived, but the establishing clone was so diluted that it could not be detected. Further investigation of the genetic alterations in these tumors would be necessary to differentiate between these two hypotheses. Based on the shared HPV types and HPV integration sites seen within the tumor pairs of Patients 2 and 3, we concluded these tumors were likely of clonal origin. The integration site in Patient 3 was not identical, but *CD36* is not a known integration hotspot, so it is unlikely to see integration here in unrelated samples. However, the altered site of integration also suggests the HPV-containing DNA was edited with tumor evolution over time. It is unknown whether viral integration events are stable over time or if they are subject to changes due to either mobile element characteristics or genomic instability. The observation that clonal populations of cells in different anatomic sites with the same integration sites are preserved also implies that these are beneficial fusions that somehow give the cells a survival advantage, but how exactly these fusions contribute to cancer cell fitness is still unclear.

In **Chapter 4**, we explored a new methodology to detect HPV integration and leveraged this novel sequencing pipeline to explore in depth the viral-host fusions in two HPV+HNSCC models. We have mainly used DIPS-PCR to detect HPV integration [10, 13, 28, 29]; however, there are limitations to this method. Due to its time-consuming and labor-intensive protocol, it cannot be easily scaled up to assess a high volume of samples. Next-generation sequencing (NGS) overcomes this issue and has the capability to rapidly process a large number of samples. Whole genome sequencing has been used on HPV+ cell lines and a relatively large panel of HPV+ HNSCCs in The Cancer Genome Atlas (TCGA), as has RNAseq [4, 5, 7]. However, the rare nature of HPV integration sites makes it challenging to detect these events because of sequencing depth issues. Therefore, we feel the use of HPV capture technologies prior to sequencing is optimal because these sites will be enriched for and covered at a much higher depth. The issue with NGS data is that analysis relies on high-quality fusion detection bioinformatics tools. Our group has attempted to use two previously available viral-host fusion callers, VirusSeq [30] and VirusFinder2 [31, 32], with little success. These callers are not optimized for targeted capture sequencing data, and they provide too little information for downstream analysis of the sites. Therefore, we generated a novel bioinformatics pipeline "SearcHPV" specifically for HPV targeted capture data, which we showed to be accurate by Sanger sequencing validation.

We performed HPV targeted capture sequencing on two HPV+HNSCC models and ran them through the SearcHPV pipeline. The first model we included was an HPV16+ cell line derived in our lab, UM-SCC-47, which has been assessed for HPV integration by multiple methods [7, 10, 33-35]. Given that the integration sites are well-described in this sample, we felt it was useful to determine the accuracy of our caller. SearcHPV called a total of six integration sites in UM-SCC-47 clustered within/near *TP63*, aligning with previous reports by our group and others [7, 10, 33-35]. Depending on the method used, the sites found within this gene differed slightly, but the majority of the sites we detected here have also been reported by one or more groups [7, 10]. We detected one site that had not been previously reported within intron 10 of *TP63,* which we were able to validate independently by Sanger sequencing on the source DNA. There also two sites that were reported by others that we did not detect by SearcHPV [7, 35]. Despite these minor differences which could easily represent different clones that expanded in vitro in different laboratories, we feel that SearcHPV demonstrated its ability to detect a well-characterized integration site.

The second model was an HPV16+ patient-derived xenograft (PDX) model (PDX-294R), which to our knowledge has not previously been investigated for HPV integration. SearcHPV called a high number of integration sites in this model (n=69), a large percentage of which fell into known cellular genes, some with important roles in cancer. Although the majority of fusions detected were spread across the genome, we found a number of loci where integration events were clustered together surrounding cancer-related genes, including *MYC*, *ZNF148* and *FOXN2*.

Other groups have previously reported that integration is associated with large-scale structural variations, including deletions, duplications and rearrangements [5, 7]. In order to investigate this in our chosen models, we also subjected them to 10X linked reads sequencing. We found that approximately half of the integration sites we detected in PDX-294R were associated with large genomic duplications ranging from 50 to 100 kb, as were the integration sites within *TP63* in UM-SCC-47. It is likely that these large genomic duplications were formed during the process of viral integration because they are consistent with previous reports that rolling circle amplification generates large HPV-human concatemers, resulting in large-scale duplications [7]. Based on the proximity of the integration sites to each other surrounding *TP63* in UM-SCC-47, it is likely that these sites actually represent multiple fragments of one larger integration site that was established by rolling circle amplification. Similar events likely resulted in the multiple integration events detected within *ZNF148*, *MYC* and *FOXN2* in PDX-294R, explaining why there were large-scale duplications seen here as well. Other integration sites that were not associated with genomic duplications could have possibly integrated by a simpler method, such as direct integration [36].

Future Directions

Many questions about the impact of HPV integration on the progression of HNSCC still remain, but the rapidly developing set of tools available will allow us to address these unknowns. We demonstrated the utility of our HPV targeted capture-based pipeline SearcHPV with two HPV+ HNSCC models, and in the future, we plan to use this tool to assess the HPV integration sites within a larger cohort of samples, including a larger panel of HPV+ cell lines, additional

HPV+ PDXs from the National Cancer Institute, and approximately 300 HNSCC tumors we have sequenced thus far which we will continue to expand upon. We expect that the application of this pipeline in these large cohorts will help advance our understanding of the potential oncogenic mechanisms associated with this process by allowing for analysis of the patterns seen across many samples. This especially would allow for deeper investigation of the potential DNA repair mechanisms at play at these sites described in Chapter 4, as we have only been able to assess microhomology levels in a limited number of junctions.

Pairing this targeted capture data with long-range DNA sequencing technologies, such as Nanopore sequencing, will allow us to better define the complex structural rearrangements caused by HPV integration and explain the structural basis of local amplification at integration sites. It would be particularly useful to run this parallel sequencing approach on the HPV+ cell lines, as these samples would have an abundance of tissue for downstream analyses and can be manipulated in knockout or knock-in experiments, although any tumor or PDX with fresh frozen tissue would also be great candidates for this approach. We have initiated this paired sequencing approach by performing Nanopore long-range whole genome sequencing of UM-SCC-47, for which analysis is ongoing. We expect long-range sequencing results from this cell line will help resolve the structure of the complex integration sites seen in *TP63* which we can then use to understand the mechanism by which they were formed.

Additional analysis of these samples at the RNA-level by traditional RNAseq or a more advanced approach on the Nanopore platform would also help clarify how integration affects tumor progression. This would allow for investigation of gene expression alterations near sites of integration but also genome wide. Importantly, long-range RNA sequencing would allow for resolution of the structure of viral-human fusion transcripts, which have been associated with worse patient outcomes [18], but whose functions within the cell are entirely unknown. We have previously been limited in our ability to understand these functions due to a lack of a full sequence; however, obtaining the full sequence would allow us to tell if these fusion transcripts code for a fusion protein which could have novel oncogenic functions. Indeed, we have seen a truncated TP63 protein in UM-SCC-47, which may in fact be a fusion protein [10]. Cloning of these full-length fusion transcripts into normal cells would also clarify what function they have and if they contribute to tumorigenesis or tumor progression.

Use of these technologies will also allow us to address other interesting questions that remain unanswered. Through our assessment of bilateral tonsillar tumors, it has become clear that integration events may be clonal within tumors and that cell populations with advantageous HPV integrations can expand. However, what makes one HPV integration advantageous over another is unclear. Evaluation of additional bilateral tumor pairs, as well as paired primary tumors and lymph nodes or distant metastases, by targeted capture and Nanopore long-range sequencing would allow us to explore the clonality of integration events. Additionally, single cell sequencing of fresh HPV+ tumors obtained from our surgical collaborators would allow us to explore the heterogeneity of integration sites at the single cell level.

Lastly, given the translational nature of our work, we ultimately care how this molecular process affects the patients we treat. In chapter 2, we observed that HPV+ OPSCC patients with HPV integration had improved disease-specific survival over those without HPV integration, which contradicts what has been previously reported. The underlying mechanism for this is unclear and requires further investigation. One possibility we plan to investigate in the future is differences in the antitumor immune response. If HPV integration generates tumor neoantigens which can then be recognized as non-self by the host immune system, this could enhance antitumor immune response. We will investigate if integration-positive vs integration negative patients have differential immune infiltration patterns and whether they can present these neoantigens for immune recognition. Our overarching goal has been, and continues to be, to make advances in the understanding of HPV-driven pathogenesis to lead to the development of novel biomarkers and/or treatment paradigms.

References

1. Holmes, A., et al., *Mechanistic signatures of HPV insertions in cervical carcinomas.* NPJ Genom Med, 2016. **1**: p. 16004.
2. Kalantari, M., et al., *Human papillomavirus-16 and -18 in penile carcinomas: DNA methylation, chromosomal recombination and genomic variation.* Int J Cancer, 2008. **123**(8): p. 1832-40.
3. Morel, A., et al., *Mechanistic Signatures of Human Papillomavirus Insertions in Anal Squamous Cell Carcinomas.* Cancers (Basel), 2019. **11**(12).
4. Nulton, T.J., et al., *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**(11): p. 17684-17699.
5. Parfenov, M., et al., *Characterization of HPV and host genome interactions in primary head and neck cancers.* Proc Natl Acad Sci U S A, 2014. **111**(43): p. 15544-9.
6. van de Nieuwenhof, H.P., et al., *The etiologic role of HPV in vulvar squamous cell carcinoma fine tuned.* Cancer Epidemiol Biomarkers Prev, 2009. **18**(7): p. 2061-7.
7. Akagi, K., et al., *Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability.* Genome Res, 2014. **24**(2): p. 185-99.
8. McBride, A.A. and A. Warburton, *The role of integration in oncogenic progression of HPV-associated cancers.* PLoS Pathog, 2017. **13**(4): p. e1006211.
9. Qin, T., et al., *Significant association between host transcriptome-derived HPV oncogene E6* influence score and carcinogenic pathways, tumor size, and survival in head and neck cancer.* Head Neck, 2020. **42**(9): p. 2375-2389.
10. Walline, H.M., et al., *Integration of high-risk human papillomavirus into cellular cancer-related genes in head and neck cancer cell lines.* Head Neck, 2017. **39**(5): p. 840-852.
11. Zhang, Y., et al., *Subtypes of HPV-Positive Head and Neck Cancers Are Associated with HPV Characteristics, Copy Number Alterations, PIK3CA Mutation, and Pathway Signatures.* Clin Cancer Res, 2016. **22**(18): p. 4735-45.
12. Huebbers, C.U., et al., *Upregulation of AKR1C1 and AKR1C3 expression in OPSCC with integrated HPV16 and HPV-negative tumors is an indicator of poor prognosis.* Int J Cancer, 2019. **144**(10): p. 2465-2477.
13. Walline, H.M., et al., *Genomic Integration of High-Risk HPV Alters Gene Expression in Oropharyngeal Squamous Cell Carcinoma.* Mol Cancer Res, 2016. **14**(10): p. 941-952.
14. Jeon, S. and P.F. Lambert, *Integration of human papillomavirus type 16 DNA into the human genome leads to increased stability of E6 and E7 mRNAs: implications for cervical carcinogenesis.* Proc Natl Acad Sci U S A, 1995. **92**(5): p. 1654-8.

15. Wiest, T., et al., *Involvement of intact HPV16 E6/E7 gene expression in head and neck cancers with unaltered p53 status and perturbed pRb cell cycle control.* Oncogene, 2002. **21**(10): p. 1510-1517.
16. Olthof, N.C., et al., *Comprehensive analysis of HPV16 integration in OSCC reveals no significant impact of physical status on viral oncogene and virally disrupted human gene expression.* PLoS One, 2014. **9**(2): p. e88718.
17. Anayannis, N.V., et al., *Association of an intact E2 gene with higher HPV viral load, higher viral oncogene expression, and improved clinical outcome in HPV16 positive head and neck squamous cell carcinoma.* PLoS One, 2018. **13**(2): p. e0191581.
18. Koneva, L.A., et al., *HPV Integration in HNSCC Correlates with Survival Outcomes, Immune Response Signatures, and Candidate Drivers.* Mol Cancer Res, 2017.
19. Nulton, T.J., et al., *Patients with integrated HPV16 in head and neck cancer show poor survival.* Oral Oncol, 2018. **80**: p. 52-55.
20. Vojtechova, Z., et al., *Analysis of the integration of human papillomaviruses in head and neck tumours in relation to patients' prognosis.* Int J Cancer, 2016. **138**(2): p. 386-95.
21. Joseph, A.W., et al., *Molecular etiology of second primary tumors in contralateral tonsils of human papillomavirus-associated index tonsillar carcinomas.* Oral Oncol, 2013. **49**(3): p. 244-8.
22. Rokkjaer, M.S. and T.E. Klug, *Prevalence of synchronous bilateral tonsil squamous cell carcinoma: A retrospective study.* Clin Otolaryngol, 2018. **43**(1): p. 1-6.
23. Rasband-Lindquist, A., Y. Shnayder, and M. O'Neil, *Synchronous bilateral tonsillar squamous cell carcinoma related to human papillomavirus: Two case reports and a brief review of the literature.* Ear Nose Throat J, 2016. **95**(4-5): p. E30-4.
24. Roeser, M.M., et al., *Synchronous bilateral tonsil squamous cell carcinoma.* Laryngoscope, 2010. **120 Suppl 4**: p. S181.
25. Theodoraki, M.N., et al., *Synchronous bilateral tonsil carcinoma: case presentation and review of the literature.* Infect Agent Cancer, 2017. **12**: p. 38.
26. Goldenberg, D., et al., *Cystic lymph node metastasis in patients with head and neck cancer: An HPV-associated phenomenon.* Head Neck, 2008. **30**(7): p. 898-903.
27. McHugh, J.B., *Association of cystic neck metastases and human papillomavirus-positive oropharyngeal squamous cell carcinoma.* Arch Pathol Lab Med, 2009. **133**(11): p. 1798-803.
28. Pinatti, L.M., et al., *Association of human papillomavirus integration with better patient outcomes in oropharyngeal squamous cell carcinoma.* Head Neck, 2020.
29. Pinatti, L.M., et al., *Viral Integration Analysis Reveals Likely Common Clonal Origin of Bilateral HPV16-Positive, p16-Positive Tonsil Tumors.* Archives of Clinical and Medical Case Reports, 2020. **4**: p. 680-696.
30. Chen, Y., et al., *VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue.* Bioinformatics, 2013. **29**(2): p. 266-7.
31. Wang, Q., P. Jia, and Z. Zhao, *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**(5): p. e64465.
32. Wang, Q., P. Jia, and Z. Zhao, *VERSE: a novel approach to detect virus integration in host genomes through reference genome customization.* Genome Med, 2015. **7**(1): p. 2.

33. Myers, J.E., et al., *Detecting episomal or integrated human papillomavirus 16 DNA using an exonuclease V-qPCR-based assay.* Virology, 2019. **537**: p. 149-156.

34. Khanal, S., et al., *Viral DNA integration and methylation of human papillomavirus type 16 in high-grade oral epithelial dysplasia and head and neck squamous cell carcinoma.* Oncotarget, 2018. **9**(54): p. 30419-30433.

35. Olthof, N.C., et al., *Viral load, gene expression and mapping of viral integration sites in HPV16-associated HNSCC cell lines.* Int J Cancer, 2015. **136**(5): p. E207-18.

36. Groves, I.J. and N. Coleman, *Human papillomavirus genome integration in squamous carcinogenesis: what have next-generation sequencing studies taught us?* J Pathol, 2018. **245**(1): p. 9-18.

Appendices

Appendix 1: Supplemental Tables

Primer ID	DIPS Primer Sequence	Nested DIPS Primer Sequence
HPV16-E6	5'-GTATTGCTGTTCTAATGTTGTTCC-3'	5'-GCAAAGTCATATACCTCACGTCG-3'
HPV16-E1a	5'-ACGGGATGTAATGGATGGTTTTATG-3'	5'-AGGGGATGCTATATCAGATGACGAG-3'
HPV16-E1b	5'-ATGTTACAGGTAGAAGGGCG-3'	5'-AGTCAGTATAGTGGTGGAAGTG-3'
HPV16-E1c	5'-ACGCCAGAATGGATACAAAGACAAAC-3'	5'-ATGGTACAATGGGCCTACGATAATG-3'
HPV16-E2a	5'-ACCCGCATGAACTTCCCATAC-3'	5'-TCAACTTGACCCTCTACCAC-3'
HPV16-E2b	5'-GTGGACATTACAAGACGTTAGCCTTG-3'	5'-CATGGATATACAGTGGAAGTGCAG-3'
HPV16-E2c	5'-CGTCTACATGGCATTGGACAGG-3'	5'-GATAGTGAATGGCAACGTGACC-3'
HPV16-E5	5'-AGAGGCTGCTGTTATCCACAATAG-3'	5'-ATGTAGACACAGACAAAAGCAGC-3'
HPV16-L2a	5'-GTACGCCTAGAGGTTAATGCTGG-3'	5'-CCAAAAAGTCAGGATCTGGAGC-3'
HPV16-L2b	5'-CCACTTTACATGCAGCCTCACC-3'	5'-CTGTACCCTCTACATCTTTATCAGG-3'
HPV16-L1	5'-ATCCACACCTGCATTTGCTGC-3'	5'-GCACTAGCATTTTCTGTGTCATCC-3'
HPV18-E7	5'-CCAGAAGGTACAGACGGGGAG-3'	5'-CGGGTTGTAACGGCTGGTTTTATG-3'
HPV18-E1a	5'-ATAGACAACGGGGGCACAGAG-3'	5'-GGGGCACAGAGGGCAACAAC-3'
HPV18-E1b	5'-CCACCAAAATTGCGAAGTAGTG-3'	5'-TAATGGGAGACACACCTGAGTGGATAC-3'
HPV18-E1c	5'-GAGGAAGAGGAAGATGCAGACAC-3'	5'-AAGATGCAGACACCGAAGGAAACC-3'
HPV18-E2	5'-ACCTACAGGCAACAACAAAAGAC-3'	5'-CAGGCAACAACAAAAGACGGAAAC-3'
HPV18-E5	5'-GGGGACGTTATTACCACAATATACACA-3'	5'-ACAGATGGCAAAAGCGGG-3'
HPV18-L2a	5'-GAAATAGACACAGAGGTAGACGAAGGT-3'	5'-TCAAACCCAGACGTGCCAGTAAAC-3'
HPV18-L2b	5'-ATGTTAATGTAGTGTCCACAGGCTCA-3'	5'-GCCGGGTTGTCATATGTAATTAAAGA-3'
HPV18-L1	5'-CAGTATCTACCATATCACCATCTTCCAA-3'	5'-AACTGTGTTTTTAAGT-3'

Table S2.1. HPV16 and HPV18 DIPS-PCR primer sequences. Nested PCR reverse adapter primer: 5’-GATGCTGACGACTGATACCGG-3’.

Sample	Target	F primer sequence	R primer sequence
BMT-251	*SGCZ* exon 1- HPV16 L2	5’- CCACTTCGTTT AGTTGCGCT-3’	5’- AGGGGGTCTT ACAGGAGCAA-3’
	HPV16 L2 - *SGCZ* exon 2	5’- CAGATGTCTCTT TGGCTGCCT-3’	5’- TAACCAACA GCAGAAGGACAA-3’
BMT-323	*UTP18* exon 1 – HPV16 L2	5’- GTTCCACGTG AGCGCCT-3’	5’- TAACCAACAG CAGAAGGACAA-3’
	HPV16 L2 – *UTP18* exon 2	5’- TCCTATAGTTCC AGGGTCTCCAC-3’	5’- CCACTTCTG AGTCACCCGAG-3’
BMT-1159	*KIF21B* exon 2 – HPV16 L1	5’- GGACAAGGCC TTCACCTATGA-3’	5'-ATCCACACCT GCATTTGCTGC-3'
UM-3898	*NDST1* exon 2 - HPV18 E1	5’- ACTTCTGCTCT GCACAGGACC-3’	5’- CCGAAAGGG TTTCCTTCGGT-3’
	HPV18 E1 - *NDST1* exon 3	5’-GGCTGGAGG TGGATACAGAGT-3’	5’- CTTCAGGCCC AGGTTTGAGT -3’
UM-3954	*DNAI1* exon 11 – HPV16 L1	5’- TGCTGATGAAT ACCGGGACC-3’	5’-GCACTAGCAT TTTCTGTGTCATCC-3’
	NPAS3 exon 6 – HPV16 L1	5’- TTCCGAAACAG TCTCCATCTACC-3’	5’- ATCCACACCTG CATTTGCTGC-3’
UM-4068	HPV16 L1 – *RLN1* exon 1	5’- GCACTAGCATTT TCTGTGTCATCC-3’	5’- TGCCACTGGT CTAGGTGTCT-3’

Table S2.2. Primers to amplify predicted fusion transcripts.

Application	Target	F primer sequence	R primer sequence	Amplicon size (bp)
qRT-PCR	HPV16 E6	5’-TGCAATGTTTC AGGACCCAC-3’	5’-ATAGTTGTTT GCAGCTCTGTGC-3’	72
qRT-PCR	HPV16 E7	5’-AGAACCGGACAG AGCCCATTACAA-3’	5’-TGTGCTTTGTAC GCACAACCGAAG-3’	82
qRT-PCR	*GAPDH*	5’-CAAGAAGGT GGTGAAGCAG-3’	5’- TGAGCTTGAC AAAGTGGTCG-3’	63
RT-PCR	HPV16 E6-E7	5’-GAACTGCAAT GTTTCAGGACCCAC-3’	5’-ATTTCATCCTC CTCCTCTGAGCTG-3’	578 (E6FL-E7) 397 (E6*I-E7) 278 (E6*II-E7)
RT-PCR	HPV18 E6-E7	5’-TGTGCACGG AACTGAACACT-3’	5’-TGGAATGCT CGAAGGTCGTC-3’	695 (E6FL-E7)
RT-PCR	HPV33 E6-E7	5’-GCCAAGCATT GGAGACAACT-3’	5’-TGGTTCGTAGG TCACTTGCT-3’	652 (E6FL-E7)

Table S2.3. Primers to amplify HPV E6 and E7.

Sample ID	Sex	Age at Dx	Smoking Status	Pack yrs	Alcohol Use	Year of Diagnosis	TNM (AJCC 7th)	Treatment	Chemo	LRF?	DM?	Survival	Integration Status
BMT-8	M	64	N	NA	N	2007	Recurrence	CRT	ERB, CIS, CAR, TAX, 5FU	✓	✗	NED	-
BMT-56	M	56	F	100	H	2006	Recurrence	CRT	ERB, CIS, CAR, 5FU	✓	✗	DOD	-
BMT-280	F	64	N	NA	S	2008	T2N2bM0	CRT	CIS	✗	✗	NED	-
BMT-403	Data unavailable												-
BMT-412	M	50	C	31	L	2007	T2N0M0	RT	NA	✗	✗	NED	-
BMT-700	M	67	F	30	N	2009	T3N0M0	CRT	?	✗	✗	DOD	-
BMT-1327	M	57	C	60	S	2012	T2N2bM0	CRT	ERB	✗	✗	DOD	-
UM-3884	M	55	N	NA	N	2015	T2N2bM0	CRT	ERB	✗	✓	DOD	-
UM-3917	M	54	N	NA	L	2016	T4N2cM0	CRT	CAR, TAX	✗	✗	NED	-
UM-3955	M	50	N	NA	N	2016	T4N2bM0	CRT	CIS	✗	✗	NED	-
UM-3962	M	56	F	42	N	2016	T3N2cM0	surgery + CRT	CIS	✗	✗	NED	-
UM-3989	M	62	N	NA	L	2016	T4aN2cM1	palliative RT	NA	✗	✓	DOD	-
UM-4028	M	66	F	11	L	2016	T4aN2cMx	CRT, then immunotherapy	CIS	✗	✓	DOD	-
UM-4093	M	55	F	18.5	N	2017	T1N1M0	RT	NA	✗	✗	NED	-
BMT-233	M	79	C	60	H	2009	T2N0M0	CRT	DOC	✓	✓	NED	+
BMT-319	Data unavailable												+
BMT-322	F	69	N	NA	N	2005	T2N0M0	CRT	CIS	✓	✓	DOD	+
BMT-344	M	72	F	10	L	2007	T4N2bM0	CRT	ERB, CIS, GEM, DOC	✓	✗	DOD	+
BMT-400	F	65	N	NA	N	2007	T2N1M0	CRT	?	✗	✗	NED	+
BMT-402	M	60	F	40	H	2007	T3N0M0	CRT	?	?	?	DUC	+
BMT-404	M	81	F	40	S	2007	T2N1M0	CRT	ERB	✗	✗	NED	+
BMT-411	M	46	N	NA	S	2008	T2N2M0	CRT	CIS	✗	✗	NED	+
BMT-427	F	85	F	?	S	2009	T3N0M0	CRT	ERB	✗	✗	NED	+
UM-3940	F	66	F	15	N	2016	T1N1M0	surgery + RT	NA	✗	✗	NED	+
UM-3948	M	59	C	60	L	2016	T3N2bM0	surgery	NA	✗	✗	DUC	+
UM-4067*	F	67	F	12.5	H	2017	T3N1M0	surgery + CRT	CIS	✗	✗	NED	+
BMT-251	F	87	F	10	N	2009	T2N0M0	CRT	ERB	✓	✗	NED	+
BMT-323	M	61	C	100	H	2005	T3N2cM0	CRT	TPF	✓	✓	DOD	+
BMT-331	M	62	N	NA	N	2006	T1N2cM0	CRT	ERB	✗	✗	NED	+
BMT-1159	M	51	F	10	L	2010	T2N1M0	CRT	?	✗	✗	NED	+
UM-3898†	M	53	F	25	L	2016	T4N0M0	surgery	NA	✗	✗	DUC	+
UM-3938	M	51	F	22	H	2016	T4aN0Mx	CRT	CAR	✗	✗	NED	+
UM-3954	M	74	N	NA	N	2016	T1N1M0	surgery	NA	✗	✗	NED	+
UM-4011	M	63	N	NA	N	2016	T1N2bM0	surgery + RT	NA	✗	✗	NED	+
UM-4068	M	70	N	NA	N	2017	T2N1M0	surgery	NA	✗	✗	NED	+

Table S2.4. Patient clinical information. Oropharyngeal SCC unless otherwise indicated by *(nasopharyngeal SCC) or †(oral cavity SCC).

Abbreviations: LRF, locoregional failure. DM, distant metastasis. M, male. F, female. N, never. F, former. C, current. NA, not applicable. H, heavy. L, light. S, social. CRT, chemoradiation. RT, radiation. ERB, erbitux. CIS, cisplatin. CAR, carboplatin. TAX, taxol. 5FU, fluorouracil. DOC, docetaxel. TPF, docetaxel+cisplatin+fluorouracil. GEM, gemcitabine. ✓, yes. ✗, no. DOD, died of disease. DUC, died of unrelated cause. NED, alive with no evidence of disease. ?, unknown. -, negative. +, positive.

Target	F primer sequence	R primer sequence	PCR Product size (bp)	RT-PCR Product size (bp)
LAMA3 Exon 1	5’-CATATCCCCGGCTGCGCTA-3’	5’-GCCAGGTTGAAGTAAGTCGGG-3’	280	-
LAMA3 Exon 2	5’-CATCCTGTCACCAATGCCATC-3’	5’-CCAAGGTGAGGTTGACTCTGTT-3’	97	97
LAMA3 Intron1:68 junction – HPV16 E5	5’-TGTATTTTTAGATAAAGATGCCGC-3’	5’-ATGTAGACACAGACAAAAGCAGC-3’	260	260
CD36 Exon 4	5’-TGGGTTAAAACAGGCACAGAA-3’	5’-ACTTGAATGTTGCTGCTGTTCA-3’	95	-
CD36 Exon 5	5’-CGCTGAGGACAACACAGTCT-3’	5’-GCCACAGCCAGATTGAGAAC-3’	111	-
CD36 Exon 6	5’-TGTTCCAAGTCAGAACTTTGAGAG-3’	5’-CAGGGTACGGAACCAAACTCA-3’	75	-
CD36 Exon5-6_6-7	5’-GGCAGCTGCATCCCATATCT-3’	5’-CCATCTGCAGTATTGTTGTAAGGA-3’	-	204
HPV16 L2 – *CD36* intron 6	5’-TGCTGATGCAGGTGACTTTTAT-3’	5’-GCACCTTTCACAATTTTTAAGGCCA-3’	-	212
HPV16 E2 - *CD36* intron 5	5’-TACAGTGTCTACTGGATTTATGTCT-3’	5’-AAACTTACCTCCGTACCAGTA-3’	-	160
HPV16 E6-E7	5’-GAACTGCAATGTTTCAGGACCCAC-3’	5’-ATTTCATCCTCCTCCTCTGAGCTG-3’	-	FL 578 E6*I 395 E6*II 278

Table S3.1. Primer sequences for bilateral patient three.

Primer Set #	Integration Loci	Caller	F primer seq	R primer seq	Size (bp)
1	*MYC*	SRCH/VS	5’-TGTAACCTTGCTAAAGGAGTGA-3’	5’-TGCACCACCAAAAGGAATTGT-3’	79
2	*RHBDL3*	SRCH	5’-ATTATCCACACCTGCATTTGCT-3’	5’-GTGCTGACATTGGGTATGGGG-3’	145
3	*VIRMA*	SRCH/VF2	5’-AGGGATGTCCAACTGCAAGTA-3’	5’-CAGGAACAGCTCAGAAGCAT-3’	93
4	*CCDC59*	SRCH/VF2	5’-TGCCACAATACAAGCTTTAGTCA-3’	5’-GGGTGGTTGCAGTCAGTACA-3’	74
5	*ANKS1B*	SRCH	5’-CACATGCGCCTAGAATGTGC-3’	5’-TGAGGTGGTGTGTTAAGTAACCT-3’	99
6	*ZFP69*	SRCH	5’-GGCGTGCTTTTTGCTTTGC-3’	5’-TACTTTAAGCCCATCCAATGAATTT-3’	73
7	*MTSS1*	SRCH/VS/VF2	5’-TTCAGGTCGAGACCCTTGTC-3’	5’-CACTTGCTCCTGTAAGACCCC-3’	97
8	*DDI2*	SRCH/VF2	5’-AGGCCAACTAAACACCACGG-3’	5’-ACAATTGAAAAAGCCACTATCGG-3’	76
9	*ADGRF2*	SRCH/VF2	5’-TCCTCACTCCCACCTAACCG-3’	5’-TTCGGTTACGCCCTTAGTTTT-3’	103
10	*NRXN3*	SRCH	5’-CCAGGTAGACATAGATCCTTGACC-3’	5’-ACTGCAAATTTAGCCAGTTCAAA-3’	82
11	*ZNF148*	SRCH/VS/VF2	5’-TGCATCCACAACATTACTGGC-3’	5’-AGCATGATTCTGAAGGAGGGA-3’	134
12	*ZNF148*	SRCH	5’-ACACAGACAAAAGCAGCGGA-3’	5’-CCATTGACGTGTCAAGGCTC-3’	97
13	*SPATA19*	SRCH	5’-GCAGCCTCTGCGTTTAGGT-3’	5’-CAGGTAGGGTGGGGGTGACT-3’	87
14	*ARGHGEF12*	SRCH	5’-TGGTTACCTCTGATGCCCAAA-3’	5’-TTGCTGTCTAGATTCCCGCC-3’	92
15	Intergenic (6)	SRCH/VF2	5’-AGTTGGTTACCCCAACAAATGC-3’	5’-AGAGGAGAATAAAATAGCCAGAGCA-3’	70
16	Intergenic (3)	SRCH	5’-ATGCAAAGGCAGCAATGTTAGC-3’	5’-TGGTTATACAGAGCCAGCCC-3’	130
17	Intergenic (6)	SRCH	5’-GATGCTGGACGCTGCAAAAG-3’	5’-TGGCGTGTCTCCATACACTT-3’	91
18	Intergenic (19)	SRCH	5’-CACAGACGACTATCCAGCGA-3’	5’-TGGCTCACGCCTATTATCACT-3’	90
19	Intergenic (5)	SRCH	5’-TGGCCACTAATGCCCACAC-3’	5’-AATCTCAGCAACAGAAAGGGGG-3’	115
20	Intergenic (3)	SRCH	5’-TCAAATAGCTTCCACCTTGGCT-3’	5’-TCTTCTTTAGGTGCTGGAGGTG-3’	126
21	Intergenic (5)	SRCH	5’-GAACAATTGTGTTATTTACTGGGGA-3’	5’-AACTTAGTGGTGTGGCAGGG-3’	123
22	Intergenic (3)	SRCH	5’-GGCACTGGTCAAGGCATTTG-3’	5’-TGGGGGAGGTTGTAGACCAA-3’	112
23	Intergenic (3)	SRCH	5’-AGACCAAAATTCCAGTCCTCCA-3’	5’-GGCACTGGTCAAGGCATTTG-3’	99
24	Intergenic (5)	SRCH/VF2	5’-ACATTTTCACCAACAGCACCAG-3’	5’-AAACCTGCTATTGAGACCTACTGC-3’	83
25	Intergenic (8)	SRCH	5’-GGGAGAGGGTGTTAGTGAAACT-3’	5’-TCCCCACAACAGTACTAAAACGTA-3’	96
26	Intergenic (8)	SRCH/VS	5’-GCATGTTCATGGGGAATGGTT-3’	5’-ACTGAGTCCCCCAATTTGCT-3’	104
27	Intergenic (13)	SRCH/VS/VF2	5’-GCTTCAGCATTCCACGATGC-3’	5’-TCCTCCCCATGTCGTAGGTA-3’	109
28	Intergenic (1)	SRCH	5’-GACCAAAATTCCAGTCCTCCA-3’	5’-GCTCCAATTGGGCATTTTTCAG-3’	88
29	Intergenic (3)	SRCH	5’-ATCTTCTAGTGTGCCTCCTGG-3’	5’-CTCCAGCAGAGATGTTCCAGA-3’	106
30	Intergenic (6)	SRCH/VF2	5’-ACACATTGTTGCACAATCCTTTACA-3’	5’-TCTGTCTGAGCATTCACAACT-3’	101
31	Intergenic (3)	SRCH	5’-TCGGAATGACTCGCAGGTG-3’	5’-GAAGGGCCCACAGGATCTAC-3’	139
32	Intergenic (4)	SRCH	5’-CAGTGGCACGCCTAGGATTA-3’	5’-AGCTCTTAACCAGTTACTAATGGAA-3’	96

Continued on next page

33	Intergenic (18)	SRCH/VS/VF2	5’-CAATAGAGTGAGTGCTCCATAACT-3’	5’-GCTTATGCAGCAAATGCAGGT-3’	147
34	Intergenic (13)	VS	5’-GTGGACCGGTCGATGTATGT-3’	5’-CCTCTTGCTTACCCACCCCT-3’	80
35	Intergenic (13)	VS	5’-GTGTGACTCTACGCTTCGGT-3’	5’-ACCTTGGGGTGTTACAAGGC-3’	199
36	Intergenic (2)	VS	5’-CGACCCATACCAAAGCCGT-3’	5’-GGGCCACACTTGGAGTAGTAA-3’	174
37	Intergenic (2)	VS	5’-CCCTGCCACACCACTAAGTT-3’	5’-CTTGGGCCACACTTGGAGTA-3’	167
38	*VIRMA*	VS	5’-AACCTCCCATCACTGACCCA-3’	5’-AACCAGCCGCTGTGTATCTG-3’	250
39	*MTSS1*	VS	5’-AGGAAGCGTTCCCTGCAAAA-3’	5’-TGAGGTGGTGGGTGTAGCTT-3’	242
40	Intergenic (18)	VS	5’-AGGATAACAACTTTTGCAGCGT-3’	5’-TTCCTCCCCATGTCGTAGGT-3’	179
41	*MTSS1*	VS	5’-TGCCTGATCGCATTCCAAGT-3’	5’-GCATGACACAATAGTTACACAAGC-3’	193
42	*TP63*	SRCH	5’-AGGACTGAGCCTGATTCTGC-3’	5’-TCTGCATCATCTTTAAACTGCACA-3’	103
43	*TP63*	SRCH	5’-AGGAGGGTAGGTCAGAAACCA-3’	5’-TTCCTCCCCATGTCGTAGGT-3’	110
44	*TP63*	SRCH	5’-TGGCTCCCTTCCAACACAAG-3’	5’-TTACTGGCGTGCTTTTTGCT-3’	122
45	Intergenic (3)	SRCH	5’-CAGCAAGGCAAAGAAGAACCAG-3’	5’-CAGAGGCTGCTGTTATCCACAA-3’	99
46	*TP63*	SRCH	5’-AACCAGCATGGAAACAAGGGAA-3’	5’-TACAACGAGCACAGGGCCAC-3’	123

Table S4.1. HPV-human junction validation primers. Intergenic sites of integration are listed as “Intergenic (chromosome #).” Each primer set was run multiplexed with GAPDH control primers: F seq 5’-GAGTCAACGGATTTGGTCGT-3’ R seq 5’-GGAGGCATTGCTGATGATCT-3’.

Abbreviations: CL, cell line. SRCH, SearcHPV. VS, VirusSeq. VF2, VirusFinder2.

HumPos	Site	HPVPos	SP	PE	Conf
chr3:189596814	*TP63*	2379;2379	694	926	high
chr3:189597479	*TP63*	2915;3000	142	167	high
chr3:189601562	*TP63*	5807	172	190	high
chr3:189607491	*TP63*	5678	135	168	high
chr3:189612850	*TP63*	2855	7784	10321	high
chr3:189620989	Int	3091	106	130	high

Table S4.2: Detailed integration sites in UM-SCC-47 called by SearcHPV.

Abbreviations: HumPos, position of integration site in the human genome. Int, intergenic. HPVPos, position of integration site in the HPV genome. SP, number of split reads. PE, number of paired-end reads. Conf, confidence of integration site.

HumPos	Site	HPVPos	SP	PE	Conf
chr1:101055413	Int	6972	5	13	high
chr1:15946207	*DDI2*	6841;6841	5	11	high
chr1:224074048	Int	5935;2482	6	16	high
chr1:40952566	*ZFP69*	3055	2	3	low
chr2:25619085	*DTNB*	1710	3	7	high
chr2:33141307	*LINC00486*	847	1	9	low
chr2:48532381	Int	2618;1485	403	1330	high
chr2:48549010	*FOXN2*	2322	6	6	high
chr2:48677465	*PPP1R21*	3050	2	9	low
chr2:63781782	*WDPCP*	5819	4	10	high
chr3:125014302	*ZNF148*	4063	3	8	high
chr3:125062101	*ZNF148*	3029	4	8	high
chr3:125086159	*ZNF148*	3080;3009	27	43	high
chr3:125100716	*ZNF148*	3612	9	14	high
chr3:125102495	*ZNF148*	5933	12	49	high
chr3:125114197	Int	603	2	7	low
chr3:125120338	Int	5980;1440	697	1541	high
chr3:125124585	Int	3960	3	4	high
chr3:125124720	Int	3960	1	6	low
chr3:125124863	Int	3960	1	4	low
chr3:125127659	Int	962	3	3	high
chr3:125139102	Int	5933	8	45	high
chr3:125143564	Int	4726	5	7	high
chr3:125150392	Int	943	3	5	high
chr3:125202167	*SNX4*	7753	6	6	high
chr3:168565534	Int	1677	1	4	low
chr3:182496148	Int	6025	4	10	high
chr4:16933044	Int	3973	8	8	high
chr4:179530965	Int	1294	5	12	high
chr4:181953470	Int	5644;3547	3	3	high
chr5:122086492	Int	5087	4	6	high
chr5:165910098	Int	5552	5	11	high
chr5:19010612	*RP11-124N3.3*	2594	2	6	low
chr5:35396304	Int	5098	2	11	low
chr6:104216242	Int	1166	2	5	low
chr6:104246619	Int	5736	3	8	high
chr6:12630462	Int	904	1	6	low
chr6:47640614	*GPR111*	6994	6	17	high
chr6:9560308	Int	982	2	4	low
chr7:126441830	*GRM8*	4099;3781	15	23	high
chr7:45348769	Int	1536	3	6	high
chr8:125728800	*MTSS1*	3593;3647	851	2647	high
chr8:128693537	Int	547	6	9	high
chr8:128747276	*MYC*	6647;793	1	765	low
chr8:128747548	*MYC*	6647;821	1173	3438	high
chr8:128752628	*MYC*	4535	13	16	high
chr8:95537452	*VIRMA*	4898;2599	26	74	high
chr9:115989198	*SLC31A1*	1307	4	4	high
chr9:13135999	*MPDZ*	3458	4	6	high
chr9:98459616	*RP11-180I4.2*	911;815	612	1644	high
chr9:98459788	*RP11-180I4.2*	940;815	578	1876	high

Continued on next page

chr10:29399148	Int	1359	4	6	high
chr10:4287865	Int	2570	20	54	high
chr11:120297939	*ARHGEF12*	5662	16	26	high
chr11:133712310	*SPATA19*	3088	3	4	high
chr11:27916059	Int	5607	4	18	high
chr12:132188289	Int	4597	3	4	high
chr12:41882591	*PDZRN4*	5755	4	7	high
chr12:82739625	*CCDC59*	436;3463	10	24	high
chr12:99239651	*ANKS1B*	1972	2	4	low
chr13:72999194	Int	1875;1875	170	248	high
chr13:73006471	Int	273;5807	372	949	high
chr13:73018748	Int	5039	3	4	high
chr14:79246541	*NRXN3*	5678;5779	5	10	high
chr15:55379458	*RP11-548M13.1*	7512;7512	40	82	high
chr17:30607816	*RHBDL3*	5169	3	4	high
chr17:66394369	*ARSG*	1305;1296	4	8	high
chr18:40967390	Int	5671;5160	273	731	high
chr19:8795646	Int	2585	4	4	high

Table S4.3: Detailed integration sites in PDX-294R called by SearcHPV.

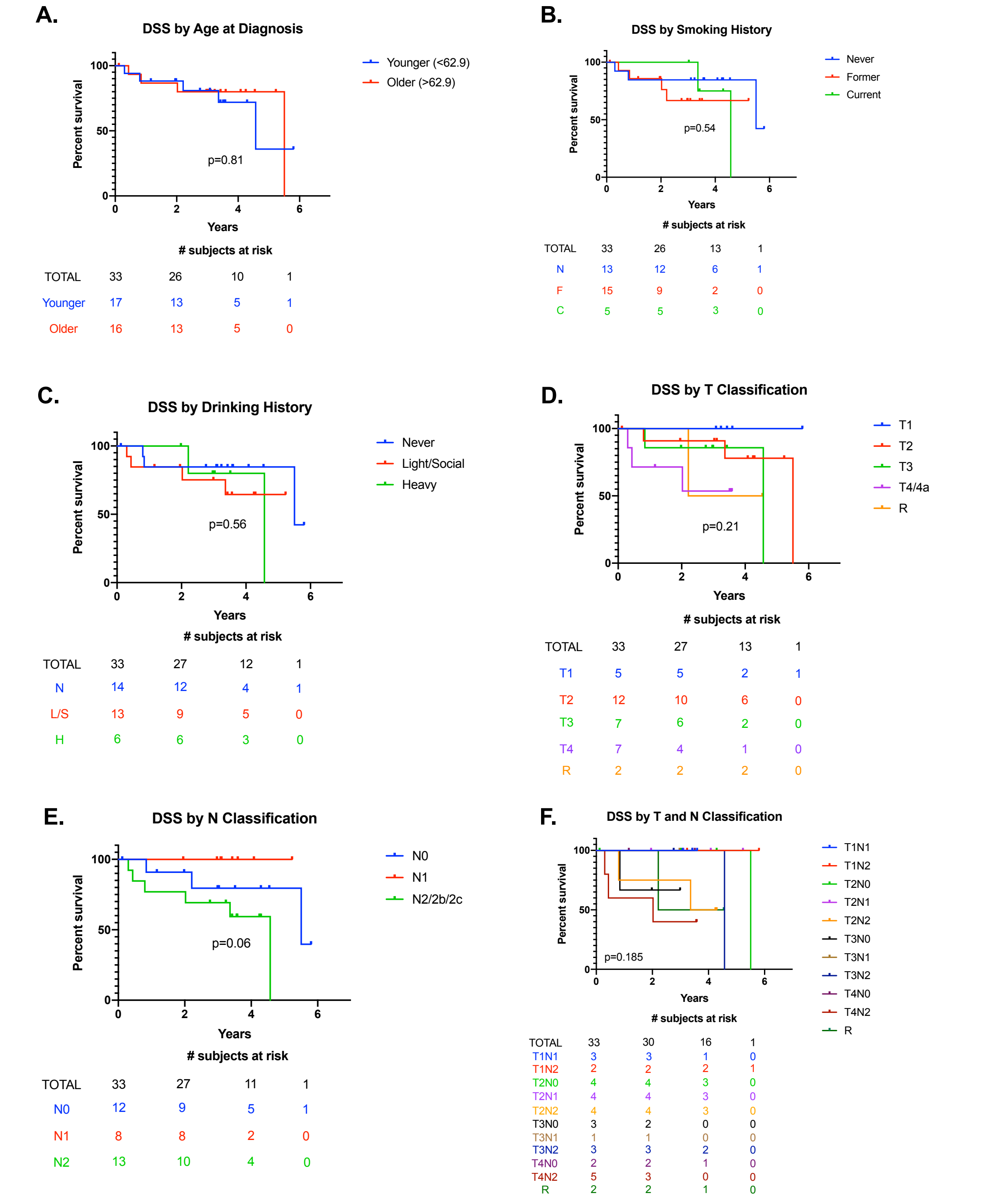
Abbreviations: HumPos, position of integration site in the human genome. Int, intergenic. HPVPos, position of integration site in the HPV genome. SP, number of split reads. PE, number of paired-end reads. Conf, confidence of integration site.

Name	UM-SCC-47	PDX-294R
GEMs Detected	1,397,989	1,497,818
N50 Linked-Reads per Molecule (LPM)	83	11
Mean DNA per GEM	453,260 bp	727,420 bp
SNPs Phased	98.90%	97.80%
Longest Phase Block	28,941,610 bp	3,891,048 bp
N50 Phase Block	7,568,760 bp	500,018 bp
DNA in Molecules >20kb	94.30%	48.90%
DNA in Molecules >100kb	39.80%	4.23%
Corrected Estimated of DNA Loaded	1.06 ng	1.71 ng
Large Structural Variant Calls	444	126
Short Deletion Calls	4,398	4,665
Number of Reads	1,047,322,794	1,018,363,956
Median Insert Size	362 bp	362 bp
Mean Depth	45.2 X	39.5 X
Zero Coverage	0.14%	0.40%
Mapped Reads	95.40%	87.40%
PCR Duplication	4.86%	5.92%
Q30 bases, Read 1	89.40%	88.10%
Q30 bases, Read 2	86.10%	85.70%

Table S4.4. Linked read sequencing statistics.

Abbreviations: GEM, gel beads-in-emulsion.

Appendix 2: Supplemental Figures



A. # subjects at risk

TOTAL	33	26	10	1
Younger	17	13	5	1
Older	16	13	5	0

B. # subjects at risk

TOTAL	33	26	13	1
N	13	12	6	1
F	15	9	2	0
C	5	5	3	0

C. # subjects at risk

TOTAL	33	27	12	1
N	14	12	4	1
L/S	13	9	5	0
H	6	6	3	0

D. # subjects at risk

TOTAL	33	27	13	1
T1	5	5	2	1
T2	12	10	6	0
T3	7	6	2	0
T4	7	4	1	0
R	2	2	2	0

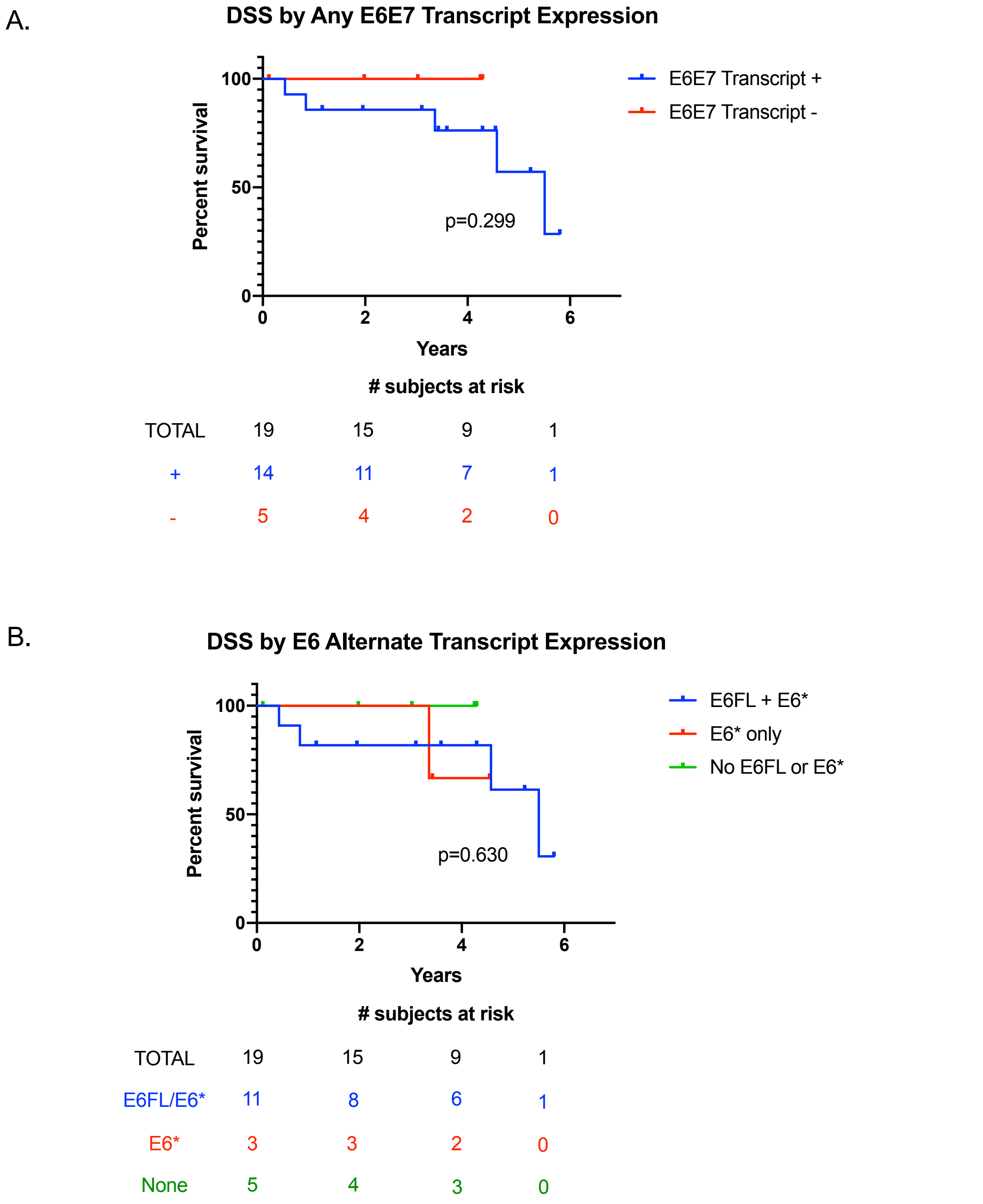
E. # subjects at risk

TOTAL	33	27	11	1
N0	12	9	5	1
N1	8	8	2	0
N2	13	10	4	0

F. # subjects at risk

TOTAL	33	30	16	1
T1N1	3	3	1	0
T1N2	2	2	2	1
T2N0	4	4	3	0
T2N1	4	4	3	0
T2N2	4	4	3	0
T3N0	3	2	0	0
T3N1	1	1	0	0
T3N2	3	3	2	0
T4N0	2	2	1	0
T4N2	5	3	0	0
R	2	2	1	0

Figure S2.1. Kaplan Meier curves separated by clinical variables. A. Age (above vs below average). B-C. Smoking and drinking histories. D-F. T and N stage.



A. DSS by Any E6E7 Transcript Expression

# subjects at risk

TOTAL	19	15	9	1
+	14	11	7	1
-	5	4	2	0

B. DSS by E6 Alternate Transcript Expression

# subjects at risk

TOTAL	19	15	9	1
E6FL/E6*	11	8	6	1
E6*	3	3	2	0
None	5	4	3	0

Figure S2.2. Kaplan Meier curves separated by A) any E6E7 transcript expression determined by qRT-PCR or RT-PCR and B) E6 alternate transcript expression.

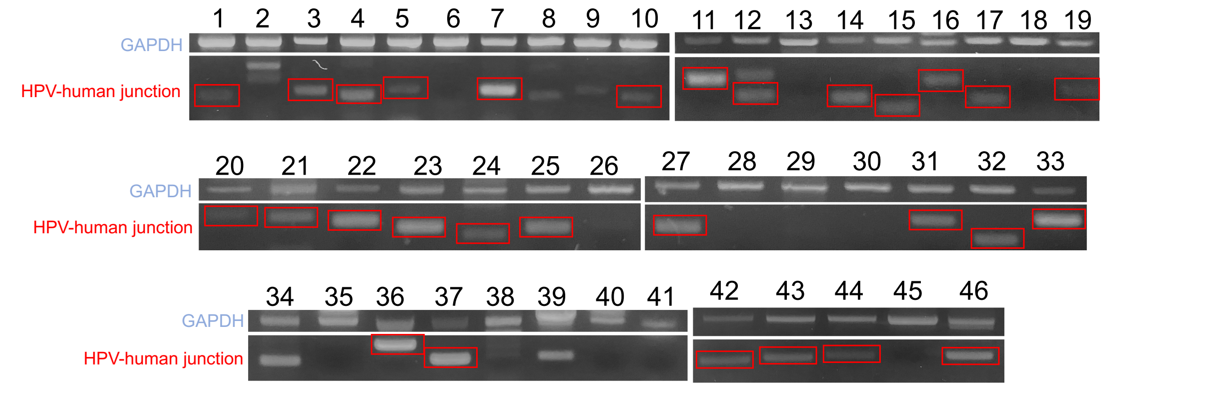


Figure S4.1: PCR validation gel electrophoresis. Top band of each row shows GAPDH (535 bp), bottom bands represent predicted HPV-human junctions (ranging from 70-250 bp). Red boxes demonstrate bands that appeared at the correct molecular weight and were validated by Sanger sequencing.

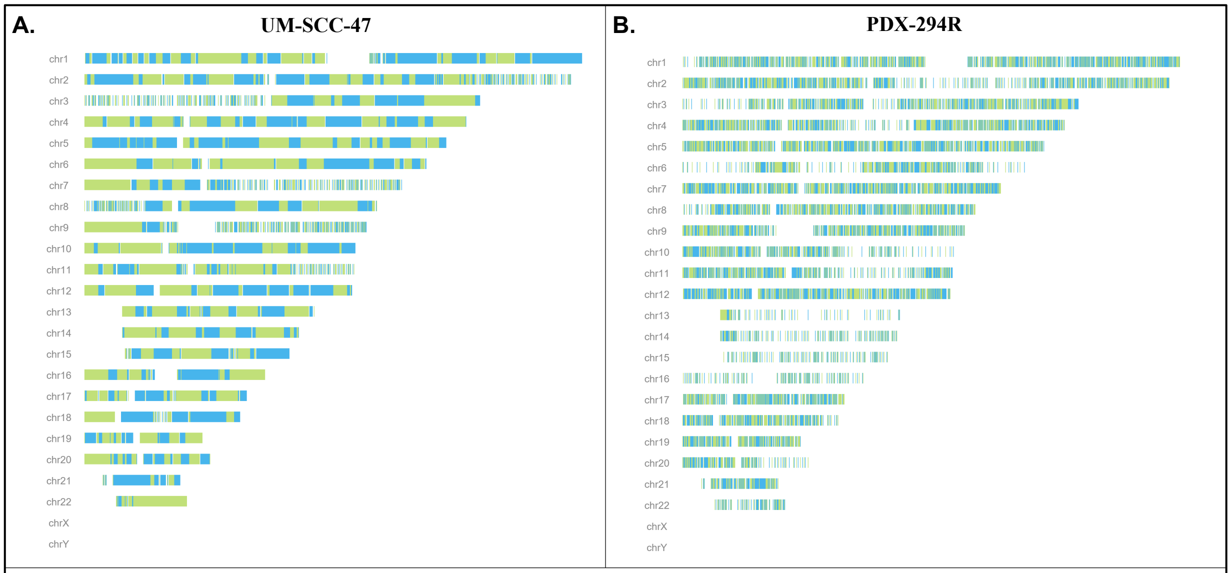


Figure S4.2: Linked read SNP phase plots for UM-SCC-47 (A) and PDX-294R (B) genomes. Alternating colors represent different phase blocks, which are contiguous blocks of DNA from the same allele based on differential SNP phasing performed by LongRanger software.

Appendix 3: Author Contributions

Chapter 1 (Introduction) is adapted from a review article published in the *Journal of Dental Research* in June 2018 entitled "Human papillomavirus genome integration and head and neck cancer." The authors are Lisa M. Pinatti[1,2], Heather M. Walline[2], and Thomas E. Carey[2].

Each author contributed to conception, design, data acquisition, analysis, and interpretation. L.M. Pinatti drafted the manuscript. All authors critically revised the manuscript.

Chapter 2 (HPV Genomic Integration and Survival of HNSCC Patients) was published online in *Head and Neck* in October 2020 as a research article entitled "Association of human papillomavirus integration with better patient outcomes in oropharyngeal squamous cell carcinoma". The authors are Lisa M. Pinatti[1,2], Hana N. Sinha[2], Collin V. Brummel[2], Christine M. Goudsmit[2], Timothy J. Geddes[3], George D. Wilson[3,4], Jan A. Akervall[3,5], J. Chad Brenner[2], Heather M. Walline[2*], and Thomas E. Carey[2*]. *([*]Co-senior authors)*

L.M. Pinatti, H.M. Walline, and T.E. Carey designed and directed the study. J.C. Brenner, C.V. Brummel, G.D. Wilson, J.A. Akervall, and T.E. Carey coordinated acquisition of samples and clinical data. C.M. Goudsmit and T.J. Geddes assisted with preparation of samples. H.M. Walline performed and analyzed the HPV PCR-MassArray. L.M. Pinatti designed and performed the remaining experiments with assistance from H.N. Sinha and analyzed the clinical data. L.M. Pinatti drafted the manuscript with critical revision from T.E. Carey and H.M. Walline.

Chapter 3 (Clonality of Bilateral Tonsillar Carcinomas) is adapted from a research article published in the *Archives of Clinical and Medical Case Reports* in June 2020 entitled "Viral integration analysis reveals likely common clonal origin of bilateral HPV16-positive, p16-positive tonsil tumors." The authors are Lisa M. Pinatti[1,2], Heather M. Walline[2], Thomas E. Carey[2], Jens P. Klussmann[6], and Christian U. Huebbers[6]. Additional patient samples and clinical information in this chapter not originally included in this publication were provided by Ernst-Jan Speel[7].

L.M. Pinatti and H.M. Walline contributed to study conception and design, acquisition of data, analysis and interpretation of data, and drafting of manuscript; C.U. Huebbers, J.P. Klussman and E.J. Speel contributed to sample acquisition and clinical data, as well as study conception and design; T.E. Carey contributed to analysis and interpretation of data. All authors contributed to critical revision of the manuscript.

A version of **Chapter 4 (SearcHPV: Novel viral integration detection methodology)** has been submitted for consideration for publication at *Cancer* as a research article entitled "SearcHPV: a novel approach to identify and assemble human papillomavirus-host genomic integration events in cancer". The authors are Lisa M. Pinatti[1,2†], Wenjin Gu[8†], Yifan Wang[9], Apurva D. Bhangale[2], Collin V. Brummel[2], Thomas E. Carey[2], Ryan E. Mills[8,9*], and J. Chad Brenner[2*]. *([†]Co-first authors, [*]Co-senior authors)*

L.M. Pinatti, T.E. Carey, R.E. Mills and J.C. Brenner contributed to study conception and project administration. W. Gu, Y. Wang, A.D. Bhangale, R.E. Mills and J.C. Brenner contributed to the development of new methodology and software. L.M. Pinatti, W. Gu, Y. Wang, A.D. Bhangale, and C.V. Brummel contributed to data curation and formal data analysis. L.M. Pinatti and W. Gu contributed to data validation, visualization, and drafting of the manuscript. All authors contributed to critical revision of the manuscript.

Affiliations:

[1]Cancer Biology Program, Program in the Biomedical Sciences, Rackham Graduate School, University of Michigan, Ann Arbor, MI, USA.

[2]Dept. of Otolaryngology/Head and Neck Surgery, University of Michigan, Ann Arbor, MI, USA.

[3]Beaumont BioBank, Beaumont Hospital, Royal Oak, MI, USA.

[4]Dept. of Radiation Oncology, Beaumont Hospital, Royal Oak, MI, USA.

[5]Dept. of Otolaryngology, Saint Joseph Mercy Hospital, Ypsilanti, MI, USA.

[6]Dept. of Otorhinolaryngology/Head and Neck Surgery, University of Cologne, Cologne, Germany.

[7]Dept. of Pathology, Maastricht University Medical Center, Maastricht, The Netherlands.

[8]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA.

[9]Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA.