Development of Human Papillomavirus Integration Analysis Technologies for Human Papillomavirus-Associated Cancer Research

by

Wenjin Gu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2024

Doctoral Committee:

Associate Professor Ryan E. Mills, Chair
Associate Professor Alan Boyle,
Associate Professor J. Chad Brenner,
Professor Hui Jiang,
Professor Maureen Sartor

Wenjin Gu

wenjingu@umich.edu

ORCID ID: 0000-0002-0028-1645

Dedication

I dedicate this thesis to my grandfather, Qiguo Gu

Acknowledgements

I would first thank to my mentor, Dr. Ryan E. Mills, for his mentorship, support, and great help through my Ph.D. studies. Besides guiding me to the world of bioinformatics, I would also like to thank him for introducing me to the world of video games. Being competitive in game is the same as in research. The great joy brough by Ryan supported me in my most difficulty time during the pandemic happened to my hometown. I truly appreciate Ryan's instruction on all fields he is good at.

I then would like to all members in my committee who gave me great suggestions on each milestone of this dissertation. Specifically, I would like to thank Dr. Chad Brenner, our closest collaborator, and my co-mentor. I thank for his generous help and suggestions on both our project and my career.

I would like to thank all my colleagues and friends in Mills lab. I am grateful to my friend and peer mentor, Yifan Wang, who guided me to the project of Chapter2. I would like to thank Weichen Zhou, for all the technical supports and company. I appreciated the help from Steven Ho. I still remember he mentioned being consistent on research is the most important tip. I would also like to thank Marcus Sherman on his advice helping me publish my tool in Chapter2. I truly thank to Chen Sun on bringing me to the lab and Alex Weber and her husband's help on my

careers. Finally, I am thankful to having worked with Xiaomeng Du, Jinghao Wang, Brandt Bessel and Irfan Darwish.

I would also like to take the moment and thank all my friends. Fang Fang, Yufeng Zhang, Jiahui Ji, Ruohan Liao, Jieru Shi, Jiaqi Zhang, Liying Chen, Jiacong Du, Yaqi Dai, Zheng Li and Mingyu Du. And I would like to specifically thank my two bridesmaids, Yuxing Huang and Shiyu Wang, who supported me during my busiest time of wedding.

I would also like to spend a second to thank the video games that accompany with me during my PhD times, Genshin Impact. My “husbandos” and “wifus” brought me great joy.

Finally I would like to thank my dearest family members: my parents Bo Gu and Hui Mei, who support me studying aboard; my grandparents Fengqiu Chen and Qiguo Gu, who taught me to be a scientist when I was a child. Last but not least, I would like to thank my beloved husband, Mengping Zhu, for his tremendous understanding, support and belief in me. I cannot accomplish this without him.

Table of Contents

List of Tables

List of Figures

Abstract

Viruses associated with human cancers, known as "tumor viruses," can induce cellular transformation or immortalization, representing a crucial step in cancer initiation. In past decades, the association between viruses and cancer has been a prominent focus in cancer research. Many tumor viruses, such as human papillomavirus (HPV) and hepatitis B virus (HBV), possess the capability to integrate their genomic DNA or RNA into the target host cell, whereas others, like hepatitis C virus (HCV), rarely integrate into the host genome. Recent studies propose that virus integration may introduce additional oncogenic mechanisms. For instance, in cervical and head and neck cancer, HPV integration directly influences cancer-related gene expression, leading to the generation of hybrid viral-host fusion transcripts. Therefore, detecting viral integration sites in the host genome is crucial for further understanding their oncogenic mechanisms in cancer development.

HPV is a well-established driver of malignant transformation in various cancers. However, the impact of HPV integration into the human genome remains largely unresolved due to sample size limitations and existing informatics challenges in identifying viral-host breakpoints from low-read-coverage sequencing data, especially in the presence of complex structural variations around fusion points. In response to these challenges, we developed SearcHPV, a novel method using targeted capture sequencing (TCS) to identify and assemble HPV integration sites in the genome. Our analysis of three HPV+ models demonstrated that SearcHPV detected HPV-host

integration sites with higher sensitivity and specificity than two other commonly used methods. Additionally, we validated the junction assembly of SearcHPV, aiding in the accurate identification of viral-host junction breakpoint sequences. Our findings indicated that viral integration occurs through diverse DNA repair mechanisms, including microhomology-mediated repair, etc.

We expanded our study to 291 head and neck squamous cell carcinoma (HNSCC) patients, employing TCS, RNA-Seq, and nanopore sequencing. We devised a novel approach to locally resolve complex HPV integrations using nanopore sequencing. Using statistical models, we labeled complex structures as "Type2", characterized by multiple integrations clustered with high copy numbers, and less complex integrations as "Type1." We revealed that Type2 events exhibited significantly more non-canonical splicing sites and were more likely to be transcribed, suggesting a complex transcription pattern. Additionally, RNA expression levels of oncogenes around Type2 events were significantly higher than Type1, indicating potential differences in oncogenic mechanism alterations induced by different types of integrations. In a subset of 78 patients with recurrent or metastatic samples, we explored the heterogeneity of HPV integration structures, revealing unique HPV integrations and varying copy numbers in different tumor sites of the same patients. Using nanopore sequencing on one cell line with primary and recurrent samples, we uncovered potential clonal selections of HPV integrations during tumor progression. Our findings emphasize the heterogeneous and complex nature of HPV integration associated with genome rearrangement, potentially contributing to distinct tumorigenic consequences.

We broadened our methodology to other viral-associated cancers and investigated 48 Mucoepidermoid carcinoma (MEC) patients with both TCS and RNA-Seq. We detected one patient with HPV integrated into 13 host genes and exhibiting high expression of HPV16 oncogenes E6 and E7. The genetic mechanisms of host genome integration were found to be similar to our previous findings in HNSCC. This study provided insights into the role of HPV in tumorigenesis of MEC.

Chapter 1 The Landscapes of Viral Integrations in Virus-Associated Cancer

As indicated by the World Health Organization (WHO), approximately 20% of global cancer cases are believed to stem from persistent infections, with 15% of these having a viral etiology that exhibits a higher prevalence in developing nations[1]. These viruses linked to human cancer are referred to as "tumor viruses." When the viral DNA or RNA induces cellular transformation or immortalization, it can initiate the initial stages of cancer development.

During the past centuries, the association between viruses and cancer was always one of the focal topics in cancer research. It has been reported that most of the tumor viruses have the ability to insert their own DNA (or RNA) into that of the target host cell, e.g. HPV and HBV, while others rarely integrate into the host genome, e.g. HCV. Recent studies have suggested that virus integration may represent additional oncogenic mechanisms. For example, in cervical and head and neck cancer, HPV integration has direct effects on cancer-related gene expression and generation of hybrid viral-host fusion transcripts [2,3] . The development of treatment and prevention strategies to address virus associated cancer relies crucially on our comprehension of cancer cells and the mechanisms underlying their development. In the era of sequencing, bioinformatics methods allow people to scale such kinds of studies on large cohorts and various types of viruses with higher base-resolution results compared with traditional PCR based approaches. Consequently, study of virus integration sites in the host genome and developing bioinformatics novel pipelines to fill the technology gaps in the field are essential in the further investigation of the oncogenic mechanisms in cancer development and potential improvement of the current treatments.

In this context, we have provided an overview of the foundational aspects of contemporary landscape studies in viral-associated cancers, along with an exploration of the prevailing bioinformatics technologies employed in research on viral integration. Specifically, we have delved into the findings associated with HPV integration as a distinct topic, laying the groundwork for the subsequent chapters (Chapter 2 to Chapter 4).

1.1 Introduction of tumor virus and virus associated cancer

1.1.1 History of tumor viruses

The concept of cancer as an infectious condition dates back to classical times, as indicated by historical records mentioning "cancer houses," where numerous residents developed specific types of cancer [4]. Additional support for the idea of an infectious origin of tumors came from observations that married couples occasionally experienced similar cancer types, and there were instances where cancer seemed to be transmitted from mother to child. In the 19th century, despite extensive research, no evidence was found to support the idea that bacteria, fungi, or parasites play a role in causing cancer. This led to the widespread belief that cancer is not caused by an infectious agent [4].

These views began to change towards the end of the century, however, in 1898, M'Fadyan and Hobday reported the transmission of oral dog warts without cells, using cell-free extracts [5], and in 1907, Ciuffo conducted similar transmission studies with human warts [6]. As warts are non-cancerous growths and not malignant tumors, these discoveries were not recognized as evidence

for tumor viruses. In 1909, Ellermann and Bang reported that leukemia could also be transmitted to healthy chickens through a cell-free filtrate obtained from cells of affected birds [7]. In 1911, Peyton Rous generated solid tumors in chickens by employing cell-free extracts from a transplantable sarcoma [8]. Because cancers in birds caused by infection were not regarded as reliable models for human cancers at that time, the significance of this investigation was not completely recognized until the discovery that viruses could induce murine leukemias [9,10]. This investigation resulted in the identification of the first oncogenic virus, the Rous sarcoma virus (RSV), and Rous received the Nobel Prize in 1966 for this significant contribution. In the following forty years following the discovery of RSV, additional tumor viruses were revealed. In addition to Gross and Stewart's works, in 1935, Rous and Beard illustrated that the cottontail rabbit papillomavirus (CRPV), which had been discovered a few years earlier, could trigger skin carcinomas in domestic cottontail rabbits [7,11]. Moloney and others identified a mouse virus (mouse polyomavirus) in 1953 capable of inducing a range of solid tumors [12–14].

Following achievements in the study of animal tumor viruses, researchers initiated efforts to identify viruses associated with human tumors. In 1962, Eddy, Hilleman, and collaborators demonstrated the tumorigenic potential of simian virus 40 (SV40) in primates [15,16]. Interestingly, Trentin and colleagues reported, for the first time, that viruses could be associated with cancer development in humans, at least under experimental conditions [17].

The identification of Epstein–Barr virus (EBV) through electron microscopy (EM) in cells cultured from Burkitt's lymphoma (BL) in 1964 revealed the first known human tumor virus [18]. Moreover, Epstein–Barr virus (EBV) infection has later been connected with nasopharyngeal

carcinoma, post-transplant lymphomas, and certain cases of Hodgkin's lymphomas (HL) [4]. In that year, Blumberg, while investigating inherited traits and disease patterns in various global regions, discovered that a blood sample from an Australian aborigine contained an antigen that reacted with the serum from an American hemophilia patient, which was named as Australia (Au) antigen later. In 1968, Blumberg, Okochi, Murakami, and Prince reported influential studies revealing that the blood of hepatitis patients contained the Au antigen. This antigen, identified as the surface antigen of a hepadnavirus known as HBV, serves as the causative agent for serum hepatitis disease [19,20]. Later in 1975, Blumberg and his colleagues established a connection between chronic HBV infection and hepatocellular carcinoma (HCC), one of the most common cancers in the world [21].

In 1974, Harald zur Hausen was the first to suggest that the HPV might be the causative agent for cervical cancer [22,23]. In 1983 and 1984, the same group isolated HPV 16 and 18 from human cervical cancer specimens, confirming the existence of these novel two types of HPV DNA [24,25]. Moreover, HPVs have been associated with additional anogenital cancers and a subset of head and neck cancers. In fact, HPVs are connected to more human cancers than any other virus. As a result, HPV has become a highly significant risk factor for human cancer.

The next major milestone was the identification of the human T-cell leukemia virus (HTLV-I) from patients with T-cell lymphoma/leukemia. In 1980, Gallo observed retroviral reverse transcriptase activity and retroviral particles in cultured human T-cell lymphoma cells. These particles were immunologically distinct from other known viruses [26] and named as HTLV-1. In

1981, the identification of retroviral particles in cell lines derived from patients provided evidence supporting a causal role for HTLV-1 in adult T-cell leukemia (ATL) [27].

In 1989, Houghton and his colleagues identified an antigen encoded by a previously unknown RNA virus, which was subsequently named HCV [28]. Additionally, through the novel serologic test for HCV, Houghton affirmed that HCV was the etiologic agent for post-transfusion hepatitis, distinct from both HBV and hepatitis A virus. The latter is another hepatitis virus transmitted through the fecal-oral route, usually via contaminated food or drinking water [29,30]. Furthermore, he discovered a connection between persistent HCV infection and HCC, in addition to the well-established association observed with HBV [31,32].

Kaposi's sarcoma (KS) is a tumor that spreads in the Mediterranean basin and Africa [33]. It is generally not life-threatening, typically affecting elderly males and primarily localizing to the skin. However, in individuals with acquired immune deficiency syndrome (AIDS), KS often extends to extracutaneous sites, particularly the lungs and gastrointestinal tract, leading to severe complications. The significantly elevated risk of KS in AIDS patients, about 20,000 times higher, prompted the consideration of an infectious cause for the tumor [34]. Despite ruling out the involvement of human immunodeficiency virus 1 (HIV-1) through epidemiological and experimental evidence, researchers focused on identifying new sexually transmitted infectious agents. Using a modern molecular biological approach called representational difference analysis, Chang, Moore, and colleagues identified DNA fragments in 90% of KS tissues from AIDS patients that were distantly homologous to the herpesvirus EBV [35]. This newly discovered virus was then named as Kaposi's sarcoma-associated herpesvirus (KSHV) or human herpesvirus

type 8 (HHV-8). Studies conducted over the past two decades strongly support the idea of an etiologic role for KSHV in the development of KS [33], although it is clear that viral infection alone is not sufficient to induce the disease, and other contributing factors have been proposed [1].

In 2008, a novel polyomavirus called Merkel cell polyomavirus (MCPyV) was identified in samples obtained from individuals afflicted by a severe form of human skin cancer known as Merkel cell carcinoma [36,37]. Over the past decade, multiple analyses have shown that McPyV was strongly associated with the highly aggressive Merkel cell carcinoma (MCC) [38–40]. However, there is still limited understanding of various aspects of MCPyV biology and the mechanisms through which it induces cancer.

Table 1.1 provides a concise overview of the history context of human viruses associated with the development of cancer.

1.1.2 Defining an etiologic role for tumor viruses

It is important to emphasize that oncogenesis is a rare outcome of viral infection, typically emerging following a prolonged and persistent chronic infection [41]. Also, in numerous instances, viral carcinogenesis is linked with an incomplete, non-productive infection [4].
Furthermore, isolating a virus in cancerous tissue does not automatically imply a causal relationship [41]. Consequently, establishing a viral cause for human cancer is often challenging. Hence, various guidelines have been suggested to assist in establishing a causal link between viruses and human cancers (refer to Table 1.2) [4,42–44]. While meeting some of these guidelines

can be challenging and not all are applicable to every virus, they remain valuable tools in assessing a potential association between virus and carcinomas.

Tumor viruses that are widely acknowledged include HPV, HBV, HCV, EBV, KSHV (also known as human herpesvirus 8), HTLV-1 and MCPyV[41,45,46,47]. The International Agency for Research on Cancer (IARC) identifies several tumor viruses as human carcinogens. HBV, HCV, EBV, KSHV, HPVs (particularly type 16), and HTLV-1, are categorized as "carcinogenic to humans". MCPyV is labeled as "probably carcinogenic to humans", and there has been an accumulation of evidence supporting the carcinogenicity of MCV in recent years. Additionally, HIV does not appear to cause cancers directly but does increase the risk of getting several types of cancers. The details of these eight tumor viruses and their associated carcinomas are summarized in Table 1.3, showing highly diversity of features for human tumor viruses[41,45,48]. Human oncoviruses can thus trigger a diverse range of cancers, and cancer types associated with viral infection show significant variability (See Table 1.3)[49–51].

1.1.3 Diversity of tumor viruses

Human tumor viruses exhibit remarkable diversity, encompassing viruses with large double-stranded DNA genomes such as EBV and KSHV, those with small double-stranded DNA genomes including HPV, HBV, and MCPyV, positive-sense single-stranded RNA genomes like HCV, and retroviruses exemplified by HTLV-1. Enveloped virions are characteristic of certain viruses like HBV, HCV, EBV, KSHV, and HTLV-1, while others, namely HPV and MCPyV, possess naked icosahedral virions (refer to Table 1.2). The mechanisms driving oncogenesis in

oncoviruses vary extensively, as outlined below. Nevertheless, oncogenesis is an infrequent outcome within the regular viral life cycle in all cases. Virus-induced cancers typically emerge as monoclonal events resulting from chronic infections, often manifesting many years after the initial infection. This pattern suggests that infection is merely one component in a multi-step process of carcinogenesis [41]. An exceptional case is observed in KSHV-induced Kaposi's sarcoma, which can develop as a polyclonal tumor within months of infection in immunosuppressed individuals [52].

The carcinogenic mechanisms of oncoviruses also exhibit considerable diversity [53]. However, they typically entail sustained expression of specific viral oncogenes that govern proliferative and antiapoptotic activities, the disruption of cellular genomic stability through the integration of viral DNA into the host genome, and viral facilitation of DNA damage along with immune evasion strategies [39]. Zur Hansen and colleagues summarized that human viral oncogenesis shares the following common characteristics [54,55]: (1) Oncoviruses are essential but not standalone factors for cancer development, resulting in a much lower cancer incidence compared to the prevalence of the virus in human populations. (2) Viral cancers emerge within the framework of persistent infections and manifest many years to decades after the initial acute infection. (3) The immune system can exert either detrimental or protective effects, with certain human virus-associated cancers escalating with immunosuppression and others arising in the context of chronic inflammation [49].

1.1.4 Virus-driven carcinogenesis

Tumor viruses are broadly categorized into two groups based on whether an RNA or DNA genome is enclosed within the infectious viral particle. In addition to disparities in replication and life cycle, RNA and DNA viruses also differ in their fundamental mechanisms for inducing cellular transformation and immortalization, which constitutes the initial step in tumor development. RNA tumor viruses, particularly animal retroviruses, such as mouse leukemia virus, are often characterized by their capacity to carry and/or modify crucial cellular growth-regulatory genes, specifically the oncogenes. The proteins produced by these cellular genes are non-essential for viral replication but typically play a pivotal role in controlling the cell cycle. Conversely, DNA tumor viruses such as SV40, mouse polyomavirus, adenovirus, and papillomavirus induce cell transformation by encoding proteins exclusively of viral origin, which are essential for viral replication [1].

Virus-driven carcinogenesis may be categorized into direct and indirect modes of carcinogenicity. Direct carcinogenicity of tumor viruses arises from insertional mutagenesis and the presence of viral oncogenes [56]. Carcinogenesis is attributed to virus integration, a distinct process from viral contamination. Insertional mutagenesis involves the modification of gene structure or transcript levels through successive alterations stemming from the integration of viral DNA. This phenomenon is applicable to the constitutive expression of both host and viral genes. In contrast, the expression of viral oncogenes serves as cancer driver genes within infected cells [45]. Indirect carcinogenicity of tumor viruses encompasses chronic inflammation and an immunosuppressive state originating from the infected cells. Persistent inflammation contributes to the accumulation of DNA damage in tissue stem cells through repetitive tissue

injury and regeneration [57,58]. Specific virus-associated cancers exhibit distinct DNA damage patterns, often referred to as mutational signatures [59,60].

Tumor viruses exhibit several "hallmarks of cancer". The conceptual framework known as the Hallmarks of Cancer, formulated by Weinberg and Hanahan, facilitates the analysis of the malignant phenotype by delineating specific cellular capabilities acquired during the carcinogenic process [49–51]. Each designated cancer "hallmark" represents a biological consequence resulting from oncogenic alterations, contributing to the distinctive phenotypic characteristics of the tumor. Moreover, the Hallmarks of Cancer framework aids in elucidating the multistep nature of human carcinogenesis [61]. This model outlines the time-dependent progression of cancer development, requiring the sequential acquisition of all essential cellular hallmarks that collectively constitute a malignant phenotype. The hallmarks of cancer includes ten aspects: (1) Resisting cell death; (2) Deregulating cellular energetics; (3) Sustaining proliferative signaling; (4) Evading growth suppressors; (5) Avoiding immune destruction; (6) Enabling replicative immortality; (7) Tumor promoting inflammation; (8) Activating invasion and metastasis; (9) inducing angiogenesis; (10) genome instability and mutation.

The cancer hallmark model serves as a potent tool for organizing and comprehending the process of carcinogenesis associated with human viruses and aids in distinguishing the impact of viral genes, the host response to infection, and acquired somatic mutations in the oncogenic process. Many recent landscape studies have employed the hallmarks of cancer framework to elucidate the oncogenic processes associated with viruses [49,62].

Here, we provide a comprehensive summary of the distinct oncogenic mechanisms associated with each known tumor virus.

EBV

EBV is a widespread double-stranded DNA virus belonging to the γ herpesviruses subfamily within the Lymphocryptovirus (LCV) genus. Globally, over 95% of the population is affected by EBV [63,64]. Most EBV infections occur in childhood without manifesting overt symptoms. When primary infection occurs in adolescence or adulthood, it manifests as infectious mononucleosis [65]. Following the primary infection, individuals become asymptomatic carriers. EBV is transmitted through saliva, infecting the oropharyngeal epithelial cells which serve as the initial site of infection [66]. After binding to the C21 receptor, EBV internalizes into the cell and enters a latent state. EBV is ethologically linked to nasopharyngeal carcinoma (NPC), Burkitt's lymphoma (BL), post-transplant lymphomas, and gastric carcinomas. Geographical variations and elevated frequencies of EBV-associated malignancies in specific racial groups suggest potential influences of host genetic factors on disease risk [67,68]. The Epstein-Barr virus (EBV) genome encompasses genes that code for six nuclear proteins known as Epstein-Barr nuclear antigens: EBNA-1, EBNA-2, EBNA-3A, EBNA-3B, EBNA-3C, and the Epstein-Barr nuclear antigen leader protein (EBNA-LP). Additionally, it encodes three latent membrane proteins (LMP-1, LMP-2A, LMP-2B) and Epstein-Barr non-polyadenylated early RNAs (EBERs), which are abundant in latent cells. EBERs serve as markers for detecting Epstein-Barr virus (EBV) infection [69]. EBNA1 encodes a DNA-binding protein crucial for facilitating replication of the viral episomal genome. It is expressed in latent infected B cells and in all EBV-associated malignancies [70,71]. EBNA2 functions as a transcription factor that upregulates genes encoding

LMP1-LMP2 in B cells [72]. LMP-1, a viral oncogene, exhibits similarities to a cell-surface receptor. It impedes apoptosis in EBV-infected cells by inducing anti-apoptotic proteins such as BCL-2, A20, and MCL-1 [73,74]. LMP-1 activates the NF-κB transcription factor in B-lymphocytes and modulates the epidermal growth factor receptor in epithelial cells [75,76]. Additionally, LMP-1 is implicated in three other signaling pathways—c-Jun N-terminal kinase (JNK)-AP-1, mitogen-activated protein kinase (p38/MAPK), and Janus kinase (JAK)-STAT—that regulate cell proliferation and apoptosis [77,78]. The role of EBNA3A, EBNA3B, and EBNA3C is to regulate the expression of LMP-1. The expression patterns of these genes differ based on the malignancy [79]. This variability in gene expression provides valuable insights into the molecular characteristics of each condition [47].

The initial documentation of the integration of the EBV genome into host genomes can be traced back to the 1980s [80–83]. Subsequent investigations substantiated the common occurrence of full-length EBV genomes and DNA fragments being integrated into EBV-positive lymphomas and epithelial carcinomas, such as NPC and gastric carcinoma [84–89]. These observations indicate the coexistence of integrated and episomal EBV DNA in tumor cells both in vivo and in vitro. EBV has the capability to integrate into specific genes associated with tumor suppression and inflammation, including PARK2, CDK15, and TNFAIP3, which play roles in the regulation of TNF-alpha-induced apoptosis/NF-κB pathways [90]. Integration into these genes by EBV can compromise their functions, disrupt TNF-alpha-induced apoptosis/NF-κB pathways, and contribute to the development of cancer [91]. Additionally, integration into the DNA repair-related gene NHEJ1 allows EBV to potentially impair the gene's function and the associated DNA repair pathway, leading to genomic instability in the host cell and promoting malignant transformation.

Although studies on EBV integration are on the rise, the precise implications of viral integrations, particularly their association with host genome abnormalities and the eventual development of cancer, remain inadequately comprehended.

HBV and HCV

HBV and HCV are primary etiological factors for HCC. HBV is implicated in 50–80% of HCC cases associated with viruses, whereas HCV is accountable for 10–25% of documented cases [92,93]. HCC ranks as the fifth most prevalent tumor type globally and represents the third leading cause of cancer-related deaths [94]. Chronic infections are established by both viruses, and in the presence of liver inflammation (hepatitis), the destruction of hepatocytes prompts regeneration and subsequent scarring (fibrosis). This process can progress to cirrhosis and, ultimately, the development of HCC. The development of HCC is influenced by a combination of direct and indirect mechanisms, stemming from chronic oxidative damage that fosters the emergence of mutations. Key driver events in HCC include the dysregulation of p53, TERT, and WNT pathways, primarily attributed to mutations in TP53, the TERT promoter, and CTNNB1, respectively [95–98].

HBV possesses a circular, partially double-stranded DNA genome featuring four overlapping open reading frames (ORFs) responsible for encoding the envelope (preS/S), core (preC/C), polymerase (P), and X proteins [99]. HBV replication relies on reverse transcription. The frequent occurrence of genomic integration of HBV into host DNA suggests that insertional mutagenesis plays a significant role as an oncogenic event in HBV-related carcinogenesis [100]. Although integration of the viral genome into the host chromosome is not indispensable for viral

replication, it does enhance the persistence of the viral genomes [101]. The HBV integration breakpoints are typically random, with a few notable hotspots [95]. The dynamics of HBV integration involve chronic inflammation and increased hepatocyte proliferation, which may lead to the rearrangement of integrated viral and adjacent cellular sequences [102]. Integration events can further result in chromosomal deletions and transpositions of viral sequences across chromosomes [103]. Consequently, HBV integration has the potential to contribute to genomic instability and the activation of proto-oncogenes. HCC associated with HBV often experiences insertional mutagenesis, leading to mutations in KMT2B [104,105], KMT2D, CCND1, CCNE, and TERT [98]. Notably, one-third of TERT dysregulation events are attributed to HBV integration. However, there is no distinct evidence supporting copy number amplification associated with HBV integration [96]. Furthermore, HBV-related HCC exhibits a high frequency of TP53 mutations [97].

It is essential to note, however, that the integration of the HBV genome is not an obligatory prerequisite for malignant progression, as approximately 20% of patients with HBV-associated HCC do not show evidence of integration [106]. Analysis of viral DNA sequences found in HCC has offered insights into additional oncogenic mechanisms of HBV. The majority of HCC tumor cells express sequences encoding the HBV X protein (HBx) and/or truncated envelope PreS2/S viral proteins. Furthermore, a novel viral hepatitis B spliced protein (HBSP) has been identified in HBV-infected patients [107]. Nevertheless, the mere expression of these proteins does not substantiate their involvement in HCC development, and additional research is required to elucidate their potential contributions to the development of HCC [4].

HCV is characterized as a single-stranded RNA virus. HCV encodes three structural proteins (Core, E1, and E2) that undergo cleavage by cellular proteases and seven nonstructural proteins (p7, NS2, NS3, NS4A, NS4B, NS5A, and NS5B) that cleaved by virus-encoded proteases [108,109]. The development of HCC is closely linked to inflammation, and the complement system plays a crucial role as an integral component of the inflammatory response, contributing to various stages of cancer progression. In chronic HCV infection, factors associated with the host, environment, or a combination of both play a significant role than viral factors in influencing the progression to HCC. HCV comprises seven genotypes and numerous subtypes. Specific HCV genotypes are linked to an increased risk of HCC [110]. Research has been conducted to investigate the oncogenic characteristics of HCV in hepatic cells. There are several viral proteins involved in cellular proliferation and survival pathways. The most researched HCV proteins include Core and NS5A. The HCV Core protein plays a pivotal role in modulating several key cell signaling pathways [111–114]. The HCV NS5A protein functions as a transcription factor activator and interacts with various signaling pathways, including those related to the cell cycle, apoptosis, and lipid metabolism [115–118]. Moreover, HCV Core, NS2, NS5A, and NS5B proteins interact with the tumor suppressors p53 and Rb, leading to the dysregulation of their functions [116,119]. Other mechanisms of HCV associated with HCC include inflammation and fibrosis, oxidative damage and genomic instability and epigenetic alterations [120]. HCV infection could also induce a general state of immune suppression [121,122]. Persistent HCV infection disrupts host metabolic pathways, particularly impacting glucose and lipid metabolisms, which contribute to the progression of HCC. The dysregulation of glucose metabolism leads to insulin resistance and diabetes, while HCV-induced steatosis results from increased lipogenesis, impaired degradation, and export [120]. Unlike HBV, the HCV genome does not integrate into the human genome, but it can induce

epigenetic changes that promote its replication through oncogenic events linked to the development of liver cancer [120].

KSHV

The KSHV genome, comprising around 165 Kb of double-stranded DNA, encompasses approximately 90 open reading frames, 12 precursor microRNAs (pre-miRNAs) that undergo splicing to yield at least 25 mature miRNAs, along with various non-coding and antisense RNAs [123,124].

By considering the expression patterns of viral genes, the life cycle of KSHV can be delineated into two distinct phases: latent and lytic [125]. Latency represents a non-productive phase marked by limited gene expression, enabling the virus to evade host immune recognition and ensuring prolonged viral persistence [126]. The lytic phase is marked by the sequential expression of viral genes, facilitating effective replication of viral DNA and its encapsulation into new virions [127–129]. Infection with KSHV is associated with various malignancies in humans [130], Kaposi's Sarcoma (KS), a highly vascularized tumor of endothelial origin [33], and two lymphoproliferative disorders named Primary effusion lymphoma (PEL) and plasmablastic variant of multicentric Castleman's Disease (MCD).

Among KSHV genes, viral IL-6 promotes proliferation and angiogenesis in Kaposi sarcoma [131]. Latency-associated nuclear antigen-1 (LANA-1), derived from open reading frame 73 (ORF73), is pivotal for the replication and stabilization of episomal KSHV [132], contributing to the acquisition of oncogenic properties [133,134]. Essentially, KSHV-infected cells exhibit latent

proliferation under the expression of LANA-1 without active virus replication. There are no reports demonstrating KSHV integrations in the host genomes.

HTLV-1

HTLV-1 was the first human retrovirus identified in association with malignancy [135]. Being T-cell tropic, HTLV-1 induces T-cell activation and proliferation upon infection [27]. It leads to adult T-cell leukemia (ATL) and progressive myelopathy (HAM). Transmission of HTLV-1 primarily takes place through sexual contact and mother-to-infant infection. HTLV-1 is an RNA virus with a single-stranded genome that harbors retroviral genes responsible for core proteins (Gag), reverse transcriptase (Pol), surface glycoprotein for receptor binding (Env), and transcriptional activator (Tax, Rex, p12, p21, p31, p30 and HBZ) genes. The sense genes, including p30, are transcribed by the 5′ LTR, while the regulatory gene HBZ is transcribed by the 3′ LTR in the antisense direction [136]. Tax and HBZ are recognized as two distinct oncogenes in HTLV-1, each operating through different mechanisms of oncogenesis. Tax has the capability to activate crucial cellular pathways, including those involved in T-cell activation and expansion, making it a key contributor to HTLV-1 persistence [137]. Simultaneously, Tax serves as a potent antigen, eliciting a cytotoxic T cell (CTL) response that effectively eliminates HTLV-1 infected cells. Hence, the presence of Tax acts as a double-edged sword. While it is necessary for viral replication and activation of infected T cells, it also makes HTLV-1 susceptible to immune clearance. Faced with the selective pressure from the host immune system, HTLV-1 likely adapts in vivo to suppress or lose Tax expression [138,139]. HBZ stands as the sole viral gene expressed continuously throughout HTLV-1 infection [140], providing a clear distinction between infected and uninfected cells [141]. Remarkably, HBZ demonstrates activities that are often contrary to those of Tax in

several aspects [142]. Many cellular pathways, including the nuclear factor-κB (NFκB) pathway that could be activated by Tax, are suppressed by HBZ [143].

As a RNA retrovirus, integration is part of the lifecycle of HTLV-1. HTLV-1 undergoes reverse transcription of its RNA genome into DNA upon entry, subsequently integrating into the chromosome, ensuring its persistence as an intracellular provirus [136]. The concept of insertional mutagenesis as a mechanism for HTLV-1 persistence was not widely acknowledged until recent revelations showed that HTLV-1 integration sites are often found in proximity to transcriptionally active genomic regions [144–146], including those associated with cancer driver genes [142]. These findings indicate a mechanism for sustained HTLV-1 infection through cis-perturbation of host genes by the provirus.

MCPyV

Merkel cell polyomavirus (MCPyV) is the only polyomavirus that was scientifically proven to cause oncogenesis in humans. The genome is a double-stranded circular DNA with about 5.4 kilobases in length [37]. MCPyV possesses a minimum of five viral genes, namely VP1, VP2, VP3, as well as small and large T antigens (LT). LT is the key protein in carcinogenesis [147,148], playing a dual role in inactivating the Rb pathway through its Rb binding site [149,150] and facilitating viral replication via the helicase domain at the C-terminus [151]. Merkel cell carcinoma (MCC) is an uncommon, highly aggressive neuroendocrine skin neoplasm with an uncertain cell origin, typically appearing on sun-exposed areas. MCC is marked by low survival rates, rapid metastasis, and a rising incidence, and current therapeutic options are deemed insufficient [39]. In the case of MCPyV in MCC, frequent nonsense mutations in the LT gene suggest that the mutant

MCPyV is unable to replicate due to the absence of the helicase domain. MCC development associated with MCPyV is confirmed to hinge on two factors: (1) integration of viral DNA into the host genome; (2) the expression of viral oncoproteins, including a truncated form of the LT antigen. Additional frequently observed genomic alterations in MCC encompass mutations in TP53, Rb, and PIK3CA [152], along with amplification of L-myc [153]. Samples regarding the viral biology are inadequate, and the exact oncogenic mechanism still needs to be elucidated [45].

HPV

HPV is a small, double-stranded DNA virus that is transmitted primarily through sexual contact and infects human epithelium in anogenital and oral mucosa. Currently, more than 200 HPV types have been identified and can be categorized into genera, species, and types through the comparison of their viral genomes[154]. Among them, there are 13-15 high-risk HPV that are associated with oncogenic risk. According to the record from the center for disease control and prevention (CDC), among sexually transmitted infections, HPV exhibits the highest prevalence and incidence rates. From 2011 to 2014 in United States, the prevalence of any oral HPV among adults aged 18–69 was 7.3%, with high-risk HPV at 4.0% [155]. The majority of HPV infections are typically cleared within 6–10 months. However, persistent infections with high-risk HPV types constitute a significant risk factor for the development of HPV-associated cancer. HPV is responsible for nearly all cases of cervical cancers and a substantial number of cancers affecting the vagina, vulva, penis, anus, rectum, and head and neck. Annually in the United States, approximately 46,711 new cancer cases arise in areas frequently associated with HPV. Among women, cervical cancer stands out as the most prevalent HPV-associated cancer, while among

men, oropharyngeal cancers, a subtype of head and neck cancer, are the most common as reported by CDC [156].

High-risk HPV types 16 and 18 are recognized as the primary cause behind the majority of cervical cancers and numerous head and neck squamous cell cancers [157,158]. The HPV16 genome comprises a circular genome of 7.9 kilobases, structured into three components: (1) the early gene region (E), consists of 6 genes, E1, E2, E4, E5, E6 and E7. E3, E5, and E8 have been recognized, although their expression is not consistently observed across the Papillomaviridae [159]. (2) the late gene region (L), consists of 2 genes, L1 and L2. (3) The upstream regulatory region (URR) [159] (Figure 1.1). Among them, E6 and E7 are oncogenes that disrupt the cell cycle regulation by blocking the function of key cell cycle regulators, TP53 and RB1. E6 protein interacts with cellular TP53, leading to ubiquitin-mediated degradation [160,161]. The E6 gene has different isoforms including full-length and alternatively spliced variants known as E6*I, E6*II, or E6*III. They play a role in driving oncogenic transformation. E7 protein binds to the RB1 pocket, preventing its interaction with the transcription factor E2F. This interference results in the unscheduled transcription of genes associated with cell cycle entry. It results in persistent transcription of S-phase genes, facilitated by alternative cyclin-CDK complexes in the cell cycle. Additionally, there is irregular transcription and expression of p16INK4a, serving as a valuable surrogate histological marker for HPV infection [162]. The interaction of E7 with RB induces continuous cell cycle entry, progression, and cellular proliferation [163]. The high-risk HPV E1 and E2 genes are crucial for viral replication, with E2 serving as a transcriptional repressor of E6 and E7. Other early genes, the E1, E2, E4 and E5 genes play crucial roles in the replication of the viral genome. E1 and E2 are dependent on the host DNA polymerase and replication machinery

for their functions. E1 serves as an ATP-dependent helicase, facilitating the beaks of host double-stranded DNA, thereby activating the DNA damage response pathway. On the other hand, E2 encodes a transcription factor that binds to viral DNA, acting as a repressor of the early gene transcription, particularly E6 and E7, thereby modulating key aspects of the viral life cycle [154,159]. The URR is responsible for the transcription of early genes, viral amplification, and cellular tropism and consists of DNA recognition sites for both viral and host transcription factors, including the early gene promoter (p97). Another gene promoter, designated p670, is located within the E7 open reading frame (ORF), regulating late gene expression [159]. L1 and L2 proteins function in virion capsid structural proteins. They are not expressed in the basal cells that harbor infection of late stages of precancer or cancer [159].

The mechanism and functional implications of HPV integration into the genome is a primary focus of Chapters 2-4 and will be comprehensively elucidated further in Chapter 1.2 below.

1.2 Landscape findings of HPV integration research

As discussed in Chapter 1.1, HPV has the potential to induce various types of cancers. Within these malignancies, the viral genome, initially in episomal form, may undergo linearization and integration into the host genome. Despite this, HPV integration does not constitute a required component of the papillomavirus life cycle. In earlier years, numerous studies with small sample sizes and limited technology for detecting HPV integration yielded controversial results. In recent years, larger-scale projects have provided insights into the role of HPV integration in oncogenic mechanisms, particularly in cervical cancers and head and neck cancers, presenting several inspiring models. This section aims to summarize key findings from past research on

HPV integration, specifically in the context of cervical cancer, head and neck cancer, and other cancers induced by HPV.

1.2.1 HPV integration research in cervical cancer

1.2.1.1 Introduction of cervical cancer

Cervical cancer ranks as the second most prevalent cancer globally, registering approximately 604,000 new cases and 342,000 deaths worldwide in 2020 [164]. This tumor represents a gradual cellular alteration in the cervix that arises subsequent to HPV infection. The two primary pathological types of cervical cancer are squamous carcinoma (SCC) and adenocarcinoma (AC). Adenocarcinomas constitute around 25% of cervical cancers, and unlike cervical SCC, which is predominantly associated with HPV, adenocarcinomas in the cervix form a more diverse group of tumors, with approximately 15% having no connection to HPV infection [165]. In cases of cervical SCC, the integration of HPV into the host genome is commonly observed. Indeed, studies reported that more than 83% of cervical cancers showed positive for HPV integration in HPV-infected groups [166,167]. In cervical adenocarcinoma (AC), fewer studies have been conducted. HPV integration was reported to be detected in 36.8% of HPV-infected patients in a recent study [168].

1.2.1.2 HPV integration and survival outcomes

The correlation between HPV integration status and patient survival in cervical cancer has been established through several studies, yielding overall consistent results. A study involving 121 cervical cancer patients in 2009 suggested that individuals with HPV16 integrated forms exhibited better disease-free survival compared to those with non-integrated forms, although statistical significance was not reached [169]. In 2017, another study with 98 locally advanced cervical cancer patients demonstrated that the episomal HPV group showed the most favorable disease-free survival outcome, while the integrated group had a poorer outcome, and the HPV-negative group had the worst prognosis [170]. A more comprehensive study in 2018, with a sample size of 108, reported that the HPV-positive group (overall survival, 123 months) exhibited better survival outcomes compared to the HPV-negative group (overall survival, 37 months). The episomal form of the virus was identified as a favorable predictive factor (overall and relapse-free survival, 100%), while the integrated form was a significantly unfavorable predictive factor (overall survival, 25 months; relapse-free survival, 7 months). The survival of patients with the integrated form was notably lower than that of HPV-positive patients and individuals with a mixed form of the virus (relapse-free survival, 52 months), considering both overall and relapse-free survival [171].

1.2.1.3 Location of integration

The first noted integration of HPV into the human genome was elucidated in 1987, specifically between KLF5 and KLF12 in the SiHa cell line [172]. Early studies of the role of integration in cervical lesions indicated a stochastic nature or a favor for common fragile sites, regions with microhomology, highly transcriptionally active regions, or proximity to microRNAs (miRNAs)

[166,173]. Subsequent investigations, however, revealed multiple hotspots on the human genome for HPV integration. Notably, integrations were detected in close proximity to the cMYC locus, suggesting a potential preference for integration of HPV18 at this locus [174]. A study by [175] demonstrated that a majority of integration events occur within known or predicted genes or in the vicinity of miRNAs [175] . In a more extensive analysis involving 104 patients and 5 cell lines, [3]identified genomic hotspot regions where integration events occurred, leading to elevated protein expression from MYC and HMGA2 when HPV integrated into flanking regions [3]. In 2016, [176] examined over 1,200 integration events in cervical cancers, revealing that integration most frequently transpired at three loci: 3q28, 8q24.21, and 13q22.1. These regions, rich in genes, encompass crucial tumor suppressors such as TP63, TPRG1, MYC, KLF5, and KLF12. Additionally, they observed a higher frequency of integration into genes than expected by chance, suggesting a potential for functional alteration of critical genes [176] . In 2021, Kamal et al. identified the MACROD2 gene as the most frequent HPV integration hotspot, followed by MIPOL1/TTC6 and TP63, in a cohort of 242 cervical cancer patients [177]. Concurrently, Warburton and colleagues investigated a separate cohort of 584 cervical cancers, reporting that HPV integration breakpoints were enriched at both FANCD2-associated fragile sites and enhancer-rich regions, often accompanied by adjacent focal DNA amplification [178].

1.2.1.4 Mechanism of integration

Although the mechanism of HPV integration is not fully understood, there are several models presented to explain it in different aspects. HPV integrated only one copy of a segment and presented few HPV fusions in some cases and were also observed to be focal amplified in a

clustered region with multiple fusions [178–180]. In the era of short read data, several papers demonstrated diverse models to depict such different types of HPV integration events. In 2017, McBride and Warburton reviewed previous research and described HPV integration as two types: Type 1 involves the integration of a single genome into cellular DNA, while Type 2 entails multiple tandem head-to-tail repeats of the HPV genome, sometimes with intervening cellular flanking sequences, at a singular genomic locus [181]. Subsequently, in 2018, Akagi et al. examined five HPV16-positive HNSCC cell lines and two HPV16-positive cervical cancer cell lines, introducing a "looping model" to describe focal duplication with both HPV and human segments forming DNA concatemers around integration events [182]. This research was further demonstrated by the same group recently using long read sequencing on 5 primary oropharyngeal cancers and 4 cell lines (3 cervical cancers, 1 head and neck cancers) [183]. In this paper they tried to resolve the recombination of host DNA and virus rearrangements and suggested this structural variation as “heterocateny”. They presented that this structure arose from extrachromosomal circular DNA (ecDNA) insertion with repetitive HPV and human segments, which can be excised from chromosomes, amplified, and undergo recombination events between host and/or HPV segments within the same cells (Figure 1.2). Another study in 2022 employed long-read sequencing in cervical cancer to elucidate clonal human papillomavirus (HPV) integration events. They suggested this was due to inter-chromosomal translocations and extrachromosomal circular (ECC) DNA structure. The researchers categorized four types of HPV integration: Type A, characterized by a truncated HPV genome containing E6/E7; Type B, featuring a truncated HPV genome lacking E6/E7; Type C, marked by an overflowing continuous segment containing the intact HPV genome; and Type D, which represents a combination of Type A, Type B, or Type C [179]. Earlier this year, the existence of

ecDNA structure in HPV integration events was further affirmed by Tian et al, utilizing real time CRISPER FISHer technology [184].

Another aspect of HPV integration mechanism involves cellular repair processes, such as nonhomologous end joining (NHEJ) and homologous recombination. In cervical cancer, this mechanism was addressed in a large-scale study in 2015. Hu, et al reported that microhomologous sequence between the human and HPV genomes was significantly enriched near integration breakpoints. This observation suggests that the fusion between viral and human DNA may have occurred through a process mediated by microhomology. Additionally, they also highlighted that the genomic elements including SINE-Alu were enriched significantly with integration sites and proposed a model induced by connecting breakage at the HPV fusion around SINE-Alu [3] (Figure 1.3).

1.2.1.5 HPV integration and viral oncogenes

The elevated expression of viral oncogenes E6 and E7 in cervical cancer has been documented in many studies, first observed by zur Hausen in 1989 [185]. The mechanism underlying this phenomenon was initially elucidated as a consequence of the disruption of E2, the repressor of E6/E7, facilitated by human papillomavirus (HPV) integration during the early stages of investigation in this field [186–190]. This was supported by some DNA-level research that demonstrated that HPV breakpoints enriched significantly in E2/E1 region and inserted less in E6/E7 genes.

Recent research has advanced our understanding by revealing that the focal duplication structure associated with HPV integration also contributes significantly to the amplification of E6 and E7 copy numbers [179,191,192]. This dual regulatory mechanism results in the overexpression of these critical oncogenes. However, it is now thought that the cancer development can occur independently of these genes [3,193–196] without necessarily depending on the E6 and E7 oncogenes expression or different distributed HPV breakpoints in E6/E7 and E2 gene region [3,197]. Several researchers have proposed additional mechanisms that HPV uses to induce cervical cancer development, including alterations on host genes, complex rearrangement and generation of chimeric transcriptions, etc.

1.2.1.6 HPV integration and its effect on host genome

Modifications and alterations generated by the HPV integration into the host genome that can lead to carcinogenesis and could be demonstrated in four main aspects: (1) loss of function of tumor suppressor genes, (2) increase in oncogene expression, (3) fusion transcription and (3) Epigenetic level regulations. These pathways are described below.

Loss of function of tumor suppressor genes

HPV exhibits the capability to integrate its genome into tumor suppressor genes, resulting in their inactivation, potentially inducing a selective advantage within the host cell [197]. Integrations within an intron of the RAD51B gene have been reported. The RAD51B gene plays a crucial role in the DNA repair pathway, and its functional loss may contribute to genomic instability [196,198]. In additional investigations, the impact of HPV fusion was observed to extend to another

set of tumor suppressor genes, including TP63, P3H2, GMDS-DT, and the pseudogene CMAHP[190]. Moreover, the fusion events involving HPV were associated with downregulation in the RNA expression of PROS1 when integration occurred in its upstream region, as reported by Zeng et al. in 2023 [199].

Increase in oncogene expression

Ojesina et al. [196] elucidated that the integration of HPV upstream of the NR4A2 gene is associated with its overexpression. This overexpression extends to various genes potentially implicated in the NR4A2 pathway. Additionally, diverse investigations have reported the amplification of oncogenes, such as FOXE1, PIM1, and SLC47A2 [174,200]. In recent studies, LENG9 was amplified in HPV-integrated cervical tumors with highest expression in 103 cervical tumors [179]. RNA expression of MIR205HG was observed increasing upon integration of HPV into its enhancer region [199].

Fusion transcription

The integration of HPV into the host genome has the potential to induce the transcription of fusion proteins combining viral and host elements, thereby influencing the transcriptional activity of adjacent host genes. Such fusion transcripts have also been named chimeric transcripts and have been reported in numerous studies. In 2017, The Cancer Genome Atlas Research Network (TCGA) identified viral-cellular fusion transcripts in 141 of 169 (83%) HPV-positive cancers, including all HPV18-positive cancers. 70% of these fusion transcripts included known or predicted genes while others included intergenic regions [167]. Another study also reported the evidence of chimeric transcripts in 8/11 HPV16+, 9/10 HPV18+ samples and CaSki cells [201].

The viral-cellular fusion transcripts were observed to be more stable compared to the episome-derived transcripts and were devoid of the instability core motif "AUUUA". The longer existence of the fusion transcripts may lead to elevate the expression of viral oncoproteins and drive the carcinogenic process [202]. Recent studies using RNA long-read sequencing resolved 9 fusion transcripts in 5 cervical cancers. They identified a consistent transcription pattern in fusion events, where structurally analogous fusion transcripts are generated through specific splicing in E6 and a canonical splicing donor site in E1, connecting to diverse human splicing acceptors. HPV16 E6I-E7-E1SD880-human gene was the highly expressed HPV-human fusion transcript, acting as a pivotal driver in cervical carcinogenesis, which induced the overexpression of E6I and E7, consequently leading to the transcriptional suppression of tumor suppressor genes, including CMAHP, TP63, and P3H2. They also demonstrated the existence of a novel read-through fusion gene mRNA, E1-CMAHP (E1C), resulting from the integration of HPV58 E1 with CMAHP. This fusion transcript has demonstrated its capability to facilitate the malignant transformation of cervical epithelial cells by modulating downstream oncogenes, thereby participating in diverse biological processes [190].

Epigenetic level regulations

In late stages of cervical cancer there is an alteration in methylation patterns, a phenomenon linked to the viral state [203]. The modified methylation patterns were evident in various regions of the viral genome, including increased methylation in the E2 binding sites within the URR, as well as in the L1 and L2 regions. The correlation between elevated methylation and the progression of cervical lesions is notable. These alterations contribute to the deregulation of E6/E7 expression. The correlation between methylation patterns and the quantity of integrated

viral copies within the host genome has been demonstrated [204]. Furthermore, significant differences exist in the methylation levels in samples where the HPV exists only in the episomal form, compared to samples with HPV integrated into the host genome. These observations are applicable in a similar manner to HPV 18 [204]. As described above, the RNA expression of MIR205HG exhibited an increase upon the integration of HPV into its enhancer, while PROS1 was downregulated. The promoter methylation levels of both PROS1 and MIR205HG were inversely correlated with their respective gene expressions [199]. Notably, Tian et al reported the integration of HPV gives rise to cellular super enhancers (SE), which operate as extrachromosomal circular DNA (ecDNA), and regulate unconstrained transcription [184].

1.2.1.7 HPV integration and other recent findings

In a study conducted earlier this year, it was revealed that HPV integration sites may exhibit either transcriptional silent or active transcription, resulting in the generation of viral-host fusion transcripts. Using multi-omics datasets, they reported that tumors characterized by productive HPV integration are linked to elevated levels of E6/E7 proteins, which enhanced tumor aggressiveness and immunoevasion [184].

1.2.2 HPV integration research in head and neck cancer

1.2.2.1 Introduction of head and neck cancer

HNSCC ranks as the sixth most prevalent cancer globally, with an annual occurrence of 630,000 new cases and over 350,000 fatalities [205]. Primarily originating from the mucosal linings of the upper aerodigestive tract, HNSCC predominantly manifests in the oral cavity, oropharynx, and larynx, often linked to oncogenic activation of epidermal growth factor receptor (EGFR) and mutations in tumor-suppressor genes like TP53 and CDKN2A [205,206]. HNSCC carcinogenesis is broadly categorized into high-risk HPV-mediated and HPV-negative subtypes, primarily associated with tobacco and alcohol consumption [207]. Notably, the incidence of HPV-positive HNSCC, particularly oropharyngeal squamous cell carcinomas (OPSCCs), has markedly increased in the Western world over the past decade [206]. Up to 90% of OPSCCs are now linked to HPV, surpassing the incidence of HPV-positive cervical squamous cell carcinomas in the USA [154,208].

As described before, cervical SCCs show HPV positivity in 95–100% of cases, with different HPV subtypes exhibiting varied integration frequencies. Specifically, HPV16 integrates in 50–80% of cases, while HPV18 integrates in over 90% of cases [209–211]. In OPSCCs, HPV positivity varies from 20–90% across studies due to geographical differences, sample preparation methods, and detection techniques. Moreover, 90–95% of HPV-positive OPSCCs are infected with HPV16 [212,213].

Various methodologies, such as fluorescence in situ hybridization (FISH) and polymerase chain reaction (PCR) and bioinformatics methods on sequencing techniques, have been employed to assess HPV integration rates in HNSCC, revealing percentages ranging from 5% to 70% [211,214–218]. Tonsillar squamous cell carcinomas (TSCC) was reported to have integration incidence of 40-100% [211,215]. However, challenges in directly comparing studies arise from differences in patient cohorts, tumor locations, and the mixed form of integrated and episomal HPV DNA. Additionally, variations in bioinformatic pipelines used for viral integration detection contribute to divergent integration rates across studies.

1.2.2.2 HPV integration and survival outcomes

In HPV-positive HNSCC, certain characteristics such as less genetic alterations, impaired DNA repair response, and better response to radiotherapy contribute to a favorable prognosis [206,219,220]. This is particularly evident in younger, healthier individuals with fewer comorbidities. However, factors like smoking, EGFR overexpression, advanced nodal stage, and chromosomal instability can lead to a poorer prognosis [221,222].

The association between HPV integration and patient prognosis has been debated for years [209,223]. Recent studies suggest that HPV integration may be linked to an unfavorable prognosis. Some findings indicate that patients with fully episomal or mixed forms of HPV16 exhibit better survival than those with integrated HPV16 or HPV-negative tumors [217,224–227]. However, conflicting results exist [228–230], and the method used to detect viral integration is crucial in

interpreting outcomes. For instance, techniques like PCR for E2 and E6/7 expression may overestimate mixed physical status of HPV [224].

Specifically, studies have shown varying results regarding the correlation between viral integration and patient prognosis in HPV-positive HNSCCs. The technique used for detecting viral integration, variations in anatomical locations of tumors, and relatively small sample sizes in some studies contribute to the inconsistency in findings. Another recent research (sample size = 15) in cervical cancer suggested the HPV integration events with multiple fusion sites might display worse outcomes compared to single HPV integration events [179]. This result suggested that HPV integration types need to be better identified instead of merely grouped to integrated versus episomal groups.

In summary, while some studies indicate a connection between HPV integration and unfavorable prognosis in HNSCC, the overall evidence is inconsistent. Methodological variations and the inclusion of diverse tumor locations and small patient cohorts in studies contribute to the complexity of interpreting these results.

1.2.2.3 Location of integration

Molecular investigations reveal that HPV-positive cancers often have one or more integration sites as described in the cervical cancer section [176,209,223]. These sites are distributed throughout the human genome, frequently aligning with fragile regions. Notable integration hotspots include specific locations on various chromosomes 2q22.3, 3p14.2, 3q28, 8q24.22, 9q22, 13q22.1,

14q24.1, 17p11.1, and 17q23.1–17q23.2 [176,223,231]. Walline et al. discovered differences in integration sites between responsive and recurrent oropharyngeal tumors. In responsive cases, HPV tends to integrate in non-coding regions, while recurrent tumors display complex integration patterns in cancer-related genes [232].

HPV integration predominantly occurs in gene-rich regions, particularly in cancer-related genes like oncogenes (e.g., TP63, MYC, ERBB2) and tumor suppressor genes (e.g., BCL2, FANCC, HDAC2, RAD51B, CSMD1) , and to a lesser extent in microRNA regions [210,233]. Studies, such as those by Parfenov et al. (sample size = 279) [217] and Olthof et al. (sample size = 75) [211], highlight the non-random nature of HPV integration, with a preference for less protected and more accessible chromosomal regions, including actively transcribed cancer genes. Pinatti et al. used HPV integration detection to study the clonal relationship between bilaterally developing TSCCs [234]. They identified multiple integration events, including those in genes CD36 and LAMA3.

In a recent investigation utilizing whole-genome sequencing (WGS), 105 cases of HPV-positive oropharyngeal cancers were examined, revealing virus integration in 77% of the cases. The study identified five statistically significant recurrent integration sites proximal to genes involved in the regulation of epithelial stem cell maintenance, including SOX2, TP63, FGFR, MYC, and immune evasion-related gene CD274 [2].

1.2.2.4 Mechanism of integration

Most findings related to the mechanism of HPV integration in HNSCC have been demonstrated in the section of cervical cancer as this question was often investigated using a combined cohort of both cervical cancer and HNSCC. Here, we only included the additional research related to HNSCC.

As mentioned before, the two aspects of integration of mechanisms were the looping model for focal genomic instability and cellular repair mechanisms, including NHEJ and microhomology mediated end-joining (MMEJ). MMEJ was demonstrated in a cervical cohort[3] as well as in HNSCC. In 2019, Leeman and colleagues conducted a comprehensive analysis of double-strand break (DSB) repair pathways in isogenic paired HNSCC cell lines using three different methodologies. Their findings revealed that the HPV16 E7 oncoprotein inhibits the canonical NHEJ pathway while promoting the error-prone MMEJ pathway. This mechanistic insight provides a rationale for the clinical radiosensitivity observed in these cancers. Comparing HPV positive HNSCC to HPV negative, the study observed a significant increase in the proportion of deletions with flanking microhomology, a signature indicative of the backup, error-prone DSB repair pathway known as MMEJ [235]. The distributions of microhomology around HPV integration were also examined in three HPV positive HNSCC samples in 2018 [236] and 34 HPV positive OPSCC patients in 2022 [237]. The majority of integrations identified had some degree of microhomology between 0-10bp.

1.2.2.5 HPV integration and viral oncogene

Similar to HPV integration in cervical cancer, in HNSCC, viral integration often involves the opening of the episome within the E2 leading to the deletion of E4 and E5 and parts of E2 and L2 [231]. The deletion of E2 disrupts its role as a transcriptional repressor in the URR, resulting in increased expression of E6 and E7 genes. Consequently, this disruption affects cellular signaling pathways, promotes heightened cellular proliferation, and inhibits the normal process of cell death (apoptosis) [210,220].

Huebbers et al. and Zhang et al. demonstrated a significant upregulation of HPV16-E6*I expression in OPSCCs with integrated viral genomes [238,239] . In both studies, the correlation between viral integration and E6*I overexpression was observed, along with associations with keratinocyte differentiation signatures. Paget-Bailly et al. also noted that the ectopic expression of HPV16 E6*I led to the deregulation of cellular genes involved in ROS metabolism, contributing to viral integration by inducing genome instability [240]. The presence of E6 partially mitigates the impact of E6*I. This observation is further supported by a clinical cohort, where tumors overexpressing E6*I were associated with cancer pathways linked to ROS metabolism [241]. However, additional studies are needed to elucidate how E6*I regulates genes related to oxidative stress and its implications for HPV-driven tumorigenesis [240].

Nevertheless, multiple studies on primary tumors indicate that the disruption of E2 upon viral integration alone does not necessarily result in increased expression of E6 and E7

oncogenes[211,217]. This suggests that similar to cervical cancer, HPV integration might play additional oncogenic mechanisms besides the E6/E7 viral oncogenesis pathway.

1.2.2.6 HPV integration effects on cellular genes

We used the same four aspects in the cervical cancer section to describe the alteration induced by HPV integration in HNSCC in host genes here.

Loss of function of tumor suppressor genes

The functional loss of a tumor suppressor occurs through HPV integration into specific genes. the deletion of gene regions and the production of truncated and potentially nonfunctional transcripts, as well as host-viral fusion transcripts.

Parfenov et al. identified HPV integration into RAD51, causing a 28-fold extrachromosomal amplification and the generation of alternate transcripts, likely yielding a nonfunctional RAD51 protein. Integration into ETS2 resulted in the deletion of exons 7 and 8, leading to decreased transcription of these exons and likely producing a truncated protein without affecting the overall gene expression [217].

Pannone et al. observed a correlation between HPV integration (detected by ISH) and downregulation of Toll-like receptor 4 (TLR4). TLR4 is crucial in innate immune responses to pathogens, including HPV, recognizing pathogen-associated molecular patterns (PAMPs). Decreased TLR4 expression in uterine cervical carcinomas and HPV-positive OPSCCs is

associated with altered signaling cascades, reduced interferon production, and compromised immune responses, partly facilitated by viral proteins E6 and E7 interference with innate immunity [242,243].

Episomal HPV DNA presence is linked to deregulation of immune response and cell survival pathways. Tumors with mutations in TRAF3 and CYLD genes often contain episomal HPV, suggesting a potential mechanism for maintaining episomes in HNSCCs and promoting mutations in these genes, leading to constitutive activation of NF-κB and impaired innate immunity [225,234].

Increase in oncogene expression

The investigation by Huebbers et al. explored gene expression differences in oropharyngeal tumors with and without HPV integration, revealing elevated protein expression of AKR1C1 and AKR1C3 in cases with HPV integration [238]. Upregulation of AKRs was also noted in HPV-negative tumors, likely due to oxidative stress response induced by Keap1/Cul3/NRF2 system mutations. AKRs play roles in various metabolic pathways and detoxification of chemotherapeutic drugs [244]. Feedback loops between oxidative stress response and AKR1C expression, as well as interactions with the viral isoform HPV16-E6*I, contribute to increased AKR1C1 expression. Moreover, elevated AKR1C1 and AKR1C3 levels result in reduced concentrations of retinoic acids, activating NRF2, and subsequent signaling pathways related to cell proliferation, metabolic reprogramming, chemotherapy resistance, and impaired DNA damage response [244].

Additionally, HPV integration upstream of an oncogene, as reported by Parfenov et al. can lead to oncogene overexpression through amplification of the downstream region, exemplified by the 250-fold amplification and overexpression of NR4A2 [217]. They also suggested that CD274 as a recurrent hotspot in enhancer region with higher expression level in HPV integrated patients. Interchromosomal translocations were also reported, causing overexpression of key oncogenes such as KLF5, TP63, and TPRG1 [245]. Likewise, Symer et al. observed a 16-fold enrichment in genomic copy number hyperamplification in close proximity to HPV integrations within 105 cases of HPV-positive OPSCCs. The degree of focal host genomic instability rises in correlation with the local density of integrants. Furthermore, there is an 86-fold increase in the frequency of genes expressed at exceptionally high or low levels within a range of ±150 kb around integrants [2].

Fusion transcription

Akagi et al. examined the impact of HPV integration on cell-virus fusion transcripts and gene expression in 10 HNSCC cell lines and 1 primary tumor. They identified virus-host fusion transcript expression in all samples, validating whole-genome sequencing findings. Examples of gene disruption were observed, such as HPV integration in UD-SCC-2 leading to DIAPH2 segment deletions and rearrangements, resulting in viral-fusion transcripts without native transcripts or functional protein. In UM-SCC-47, aberrant TP63 expression resulted from HPV integration-mediated amplification, generating viral-host transcripts and a truncated TP63 protein. Additional instances of gene disruption involved FOXE1 and PIM1 oncogene amplification in UPCI:SCC090 cells [182].

Hassounah et al. demonstrated HPV integration into the CD274 gene (PD-L1), generating a truncated isoform of PD-L1 that, when secreted, maintained its ability to bind PD-1, negatively regulating T cell function outside the cell[246]. Koneva et al. identified three tumors with CD274 as an HPV integration site, correlating with increased PD-L1 expression [227]. Broutian et al. observed HPV insertions flanking a 16-fold somatic amplification of the PIM1 gene in UPCI:SCC090, accompanied by increased PIM1 transcripts. PIM1 overexpression, associated with poor survival, involves PIM kinase phosphorylation of substrates in the PIK3CA/AKT/mTOR pathway, promoting increased cell metabolism and growth [182,222,247–249].

A case report by Huebbers et al. described a rare malignant transformation in juvenile-onset recurrent respiratory papillomatosis, involving low-risk HPV type 6 integration into the AKR1C3 gene, chromosomal region deletion, and loss of AKR1C3 protein expression [250,251]. Walline et al. reported that in UM-SCC-47, HPV integration into TP63 generated a viral-host fusion transcript between HPV16 E2 and exon 14 of TP63, producing a truncated ΔNTP63 protein. Other cell lines did not exhibit viral-host fusion transcripts, possibly due to in-frame integration into introns that were subsequently spliced out [252].

In 2022, another investigation identified the expression of chimeric transcripts in 147 genes within the 105 OPSCC tumors, causing outlier expression in 35% of them. Unlike the typical splicing of transcripts from amplified host genes, chimeric transcripts exhibit significantly altered structures when expressed at outlier levels. These chimeric transcripts contribute to the disruption of the 147 host genes, partly through readthrough expression and/or utilization of host

splice donor, splice acceptor, and/or cryptic splice sites. For instance, a single tumor displayed the detection of 31 distinct chimeric transcripts at CASC8 [2].

Epigenetic level regulations

In a study of 57 HPV-positive OPSCC patients, Reuschenbach et al. discovered that among 16 samples with an intact E2 gene, there was an association with methylation of specific binding sites (E2BS3 and E2BSx4) in the URR. This methylation resulted in the loss of protein expression, resembling the effect of E2 gene deletion. Interestingly, in most cases, the URR was not methylated [253]. Recent studies suggest that the methylation of the viral genome is not necessarily linked to the physical status of HPV. For example, although hypermethylation within the URR was observed in two cell lines (UM-SCC-47 and CaSki), two other cell lines (UM-SCC-104 and SiHa) with a mixed physical status of the HPV genome exhibited an unmethylated URR [253]. Hatano et al. observed a correlation between the methylation status of the integrated HPV genome in certain HNSCC cell lines and the methylation status of the host genome surrounding the integration breakpoints. Consequently, they suggested that the expression of viral oncogenes might depend on the specific location of viral integration [254].

A study elucidated the relationship between chromatin structure changes and the viral state in HPV-positive HNSCC samples. This included the H3K27ac histone mark, distinguishing tumor samples with and without HPV integration, and its enrichment colocalized with varying numbers of HPV integration sites. Additionally, these differential enrichments of the H3K27ac histone mark were found in upstream genes implicated in HNSCC tumorigenesis, such as TP63, FOXE1, NOTCH1, and EGFL7, where their expression is enhanced [231].

Furthermore, an additional investigation employing ChIP-seq demonstrated that the conserved CTCF binding site within the HPV genome associates with CTCF in four HPV-positive cancer cell lines. Notably, significant alterations in the CTCF binding pattern and chromatin accessibility were observed exclusively within a 100 kbp proximity to HPV integration sites. The outcomes indicated a novel CTCF binding site resulting from HPV integration induces a restructuring of chromatin state and the upregulation of genes crucial for tumor viability in HPV-positive tumors [255].

1.2.2.7 HPV integration and other recent findings

HPV-positive and HPV-negative HNSCCs exhibit distinct miRNA expression patterns, with specific miRNA subsets significantly associated with overall survival, disease-free survival, and distant metastasis in HPV-positive HNSCCs [256,257]. Hui et al. identified 128 differentially expressed miRNAs between tumor and normal tissue in OPSCCs and proposed that HPV integration near these miRNAs might contribute to their dysregulation [257]. Wald et al. observed altered expression of a subset of miRNAs in HPV16-positive HNSCC cell lines compared to both HPV-negative HNSCC cell lines and immortalized normal keratinocytes [258]. Given that the HPV16-positive cell lines in this study contained integrated HPV, the findings suggest a potential role of integration in the dysregulation of miRNAs.

1.3 Bioinformatics Methods in viral integration research

1.3.1 Detection technology of viral integration

Various approaches have been used to detect HPV integration in tumor tissues. Initially, techniques like in situ hybridization (ISH) or fluorescence in situ hybridization (FISH) were employed to visualize HPV DNA or RNA, enabling the identification of viral integration at the single-cell level in cells and tissues. Alternatively, PCR-based methods, such as quantitative PCR (qPCR) for E6/E7 copy number determination relative to E2, Detection of Integrated Papillomavirus Sequences (DIPS) PCR for detecting virus-human DNA sequences, and Amplification of Papillomavirus Oncogene Transcripts (APOT) PCR for detecting virus–human RNA transcripts, have been developed. Furthermore, NGS techniques like WGS, Whole Exome Sequencing (WES), and RNA-Seq have advanced and can identify HPV-human nucleic acid sequences.

New techniques are being developed to investigate viral integration along with HPV sequences, using custom-made RNA probes specific to HPV. This enables DNA enrichment for viral sequences, increasing the likelihood of detecting HPV integration. This enrichment step is followed by amplification and NGS [259–261]. Examples of emerging techniques for HPV integration detection include nanopore or Pacbio sequencing on DNA/RNA isolated from fresh frozen tissues, and HPV capturing with long read sequencing, as well as Targeted Locus Amplification (TLA) on DNA isolated from FFPE tissues, combining HPV capturing with circularization of DNA fragments and amplification. An overview of the currently employed

techniques for identifying HPV integration, along with their advantages and disadvantages, is presented in Table 1.4 [223].

1.3.2 Bioinformatics viral integration detection

A growing number of studies employ NGS techniques to identify HPV integration in the human host genome. Reliable NGS data requires an optimal bioinformatic pipeline for the rapid and exclusive detection of the viral genome from large-scale genome-wide DNA sequencing. This is typically achieved by detecting virus-host chimeric fusions or paired-end reads [210]. Several bioinformatic approaches for identifying viral integration sites have been developed, including SeqMap2 [262], VirusSeq [263], VirusFinder1-2 [264,265], ViralFusionSeq [266], Virus-Clip [267], Vy-PER [268] , BATVI [269], , HGT-ID [212] , VirTect [270] and HIVID2 [271], specifically used for detecting integrated HPV genomes.

The variation in reported HPV integration sites (0–600) in cervical cancers may be attributed to the diversity in bioinformatics tools, with suggestions that high integration rates result from low-stringency approaches [210 223]. When mapping integration sites, potential artifacts in bioinformatic data, such as misidentification of fusion transcripts, contamination, difficulty distinguishing between episomal and linearized sequences, repetitive regions, miss alignment and homologous sequence, should be considered. Quality control and confidence of integration sites using established techniques are essential [210,223,272].

Newly developed bioinformatic tools have emerged. Viral integration and Fusion identification (ViFi) detect viral integrations from WGS data and human–virus fusion mRNA from RNAseq data. Unlike reference-based alignment mapping, ViFi combines this with a phylogenetic model for better detection of evolutionarily divergent viruses [273]. SurVirus, an improved virus integration caller that was aware of integrations in repetitive regions and corrects the false alignment of reads which were crucial for the discovery of integrations [274]. Capture-based sequencing methods, like nanopore sequencing, enable sequencing of long DNA molecules, and specific bioinformatic methods are developed for accurate analysis. SearcHPV, a pipeline for targeted capture sequencing data, operates more accurately and efficiently than existing pipelines, performing local assembly around the junction site and simplifying confirmation experiments [236]. VIRUSBreakend, a virus-centric approach, detects viral integration in regions of low mappability, such as centromeres and telomeres, using single breakends [275]. The characteristics and benchmarking of these tools are summarized in Table 1.5.

As outlined earlier, numerous aspects of HPV and its genomic integration remain unclear. This dissertation aims to introduce a pioneering technology for detecting HPV integration in targeted sequencing data, detailed in Chapter 2. Subsequently, Chapter 3 will feature my investigations into identifying distinct types of HPV integrations and examining the correlation between their complexity and heterogeneity with tumor characteristics and genetic consequences, utilizing multi-omics data from a cohort exceeding 200 patients. To broaden the scope of our research to other rare cancers, Chapter 4 will showcase my efforts in exploring the role of HPV in mucoepidermoid carcinomas. The comprehensive content of this dissertation introduces several

innovative tools that contribute to the viral-associated cancer research community, providing valuable insights into additional oncogenic mechanisms induced by HPV integration.

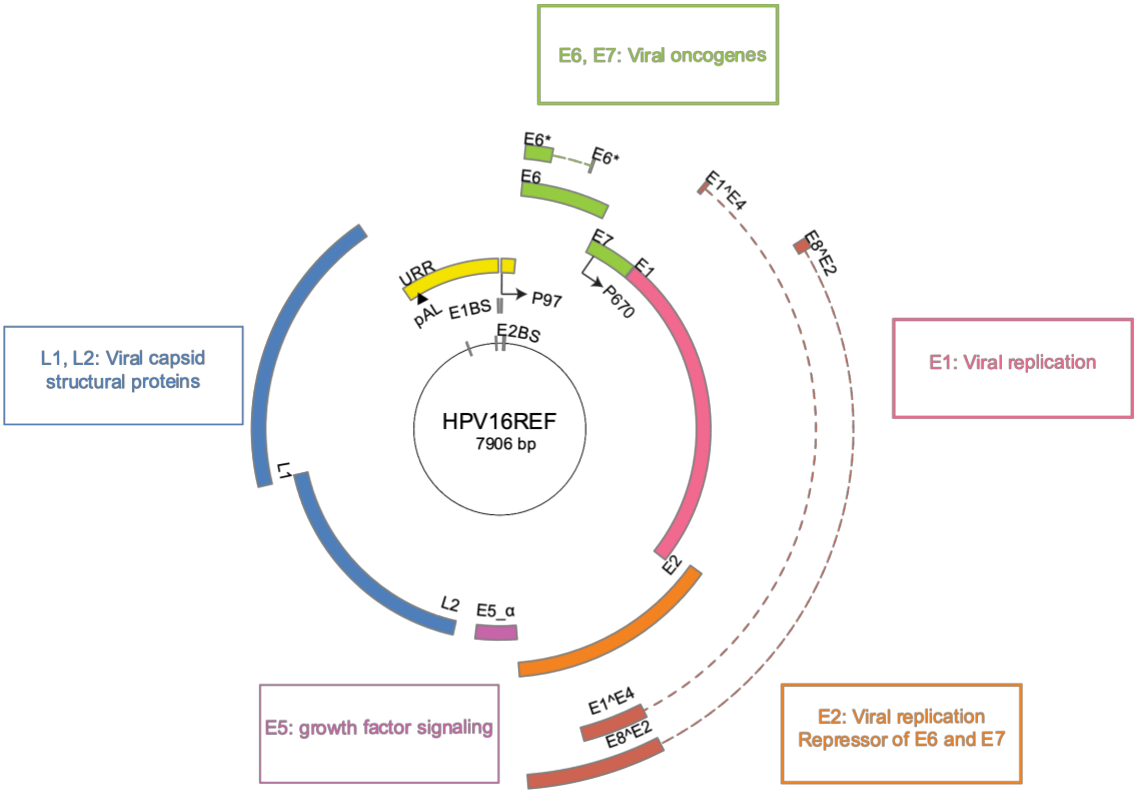


Figure 1.1 HPV16 genome structure. The main elements were from PAVE database[286]

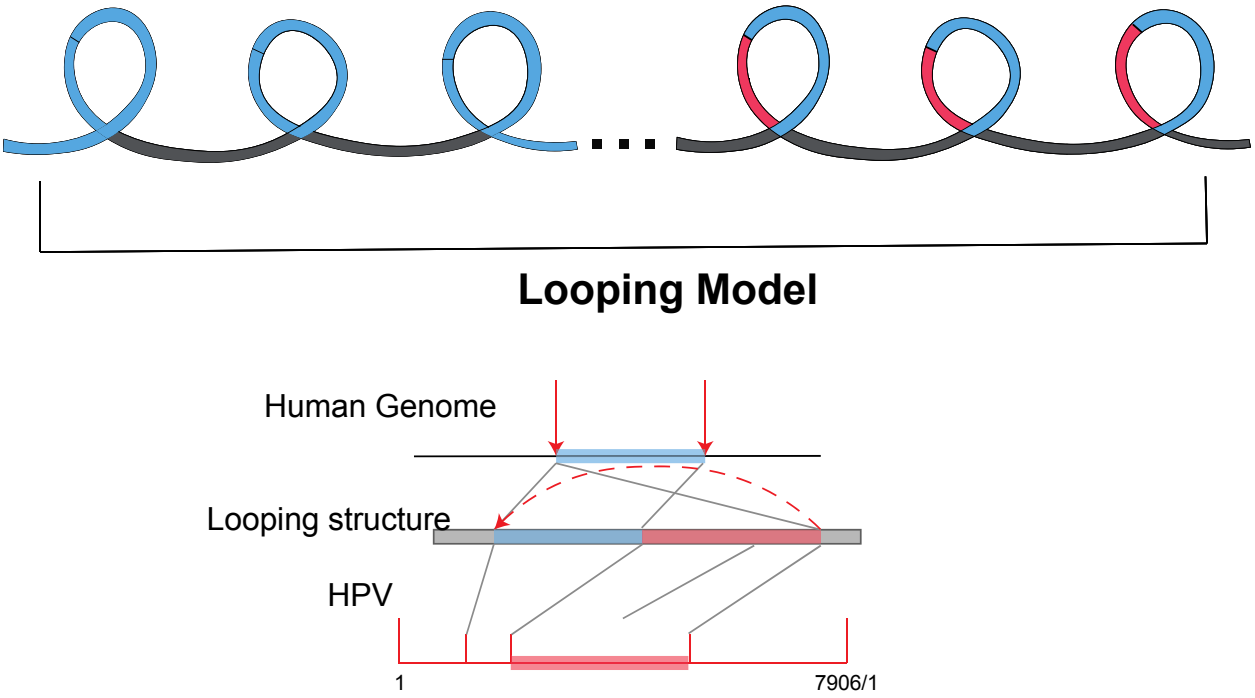


Figure 1.2 The schematic of the Looping Model [182]

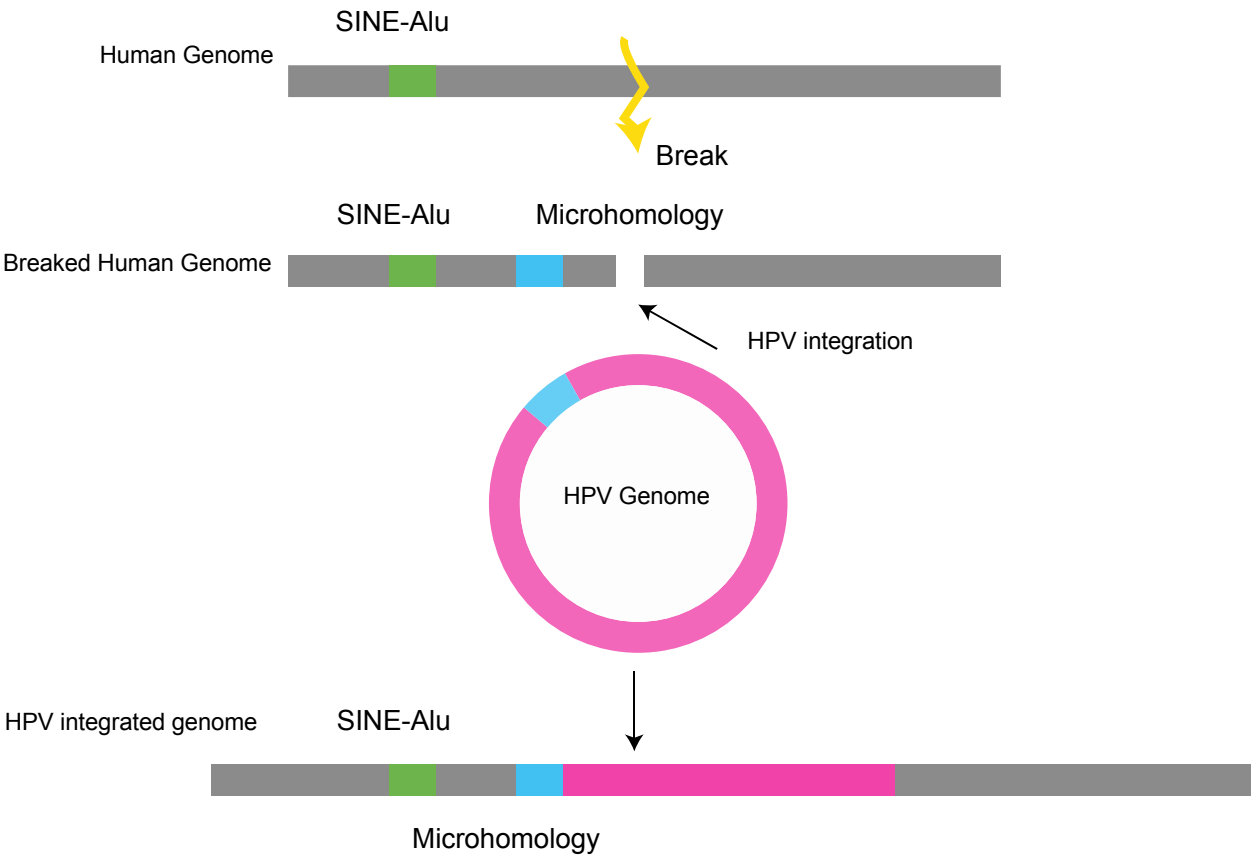


Figure 1.3 The schematic of a mechanism connecting breakpoints around SINE-Alu and using MMEJ [3]

Table 1.1 Summary of the history context of tumor virus. This table summarized the major findings and the timeline for all the tumor viruses.

Virus	Timeline	Authors	Major findings
RSV	1911	Rous, et al	the first oncogenic virus
CRPV	1935	Rous, Bread	induce skin carcinomas in rabbit model
Mouse Virus	1953	Moloney, et al	induce solid tumors in mouse model
SV40	1962	Hilleman, et al; Trentin, et al	viruses linked to primates model and human
EBV	1965	Epstein, Barr, et al	EBV was detected in BL. The first known human tumor virus
HBV	1967-1968	Blumberg, et al	HBV was linked to HCC
HPV	1974	zur Hausen, et al	HPV was linked to cervical cancer
HTLV-1	1980	Gallo, et al	HTLV-1 detected
	1981	Hinuma, et al	HTLV-1 linked to ATL
HCV	1989	Houghton, et al	HCV was detected and linked to HCC
KSHV	1994	Chang, Moore, et al	KSHV was detected from KS tissues from AIDS patients DNA fragments
MCPyV	2008	zur Hausen, et al; Moore, et al	MCPyV was detected in MCC

Table 1.2 Criteria for define a tumor virus

zur Hausen criteria [44]
1. The consistent presence and continuity of viral DNA in biopsies of tumors and in cell lines originating from the same tumor type. 2. The confirmation of growth-promoting activity exhibited by specific viral genes or host cell genes altered by the virus, demonstrated in tissue culture systems or relevant animal models. 3. The verification that the malignancy's characteristics rely on the ongoing expression of viral oncogenes or the modification of host cell genes containing viral sequences. 4. Epidemiological evidence supporting the notion that infection with the respective virus constitutes a significant risk factor for the development of cancer.
Evans and Mueller criteria [424,4342]
Epidemiologic
1. The geographical prevalence of viral infection aligns with that of the tumor, taking into account the presence of established co-factors. 2. Case subjects exhibit elevated levels of viral markers compared to matched control subjects. 3. Tumor development follows the presence of viral markers, with a higher frequency of tumors observed in individuals with markers compared to those without. 4. Preventing viral infection decreases the incidence of tumors.
Virologic guidelines 1. The virus can transform cells in vitro. 2. The viral genome is present in tumor cells but absent in normal cells. 3. The virus induces tumors in experimental animals.
Hill criteria [4,43]
1. Strength of correlation (how frequently is the virus linked to the tumor?) 2. Consistency (has the connection been consistently observed?) 3. Specificity of correlation (is the virus uniquely linked to the tumor?) 4. Temporal relationship (does virus infection precede the development of tumors?) 5. Biological gradient (is there a dose-response relationship with viral load?) 6. Biological plausibility (is it biologically feasible for the virus to cause the tumor?) 7. Coherence (does the connection align with existing knowledge about the tumor?) 8. Experimental evidence (are there corroborating laboratory findings?)

Table 1.3 Representative list of tumor viruses. Abbreviations: ds, double strands; ss, single strands; ATLA, anti-adult T-cell leukemia/lymphoma antibody; EBER, EBV-encoded small RNA; LANA, Latency-associated nuclear antigen; +, positive sense. (the American Cancer Society at http://www.cancer.org/) *HIV doesn't appear to cause cancers directly but increases the risk of getting several types of cancers.

Virus	HPV	HBV	HCV	EBV	KSHV	MCV	HTLV-1	HIV*
family	*Papillomaviridae*	*Hepadnaviridae*	*Flaviviridae*	*Herpesviridae*	*Herpesviridae*	*Polyomaviridae*	*Retroviridae*	*Retroviridae*
virus genome	dsDNA	ss/dsDNA	+ssRNA	linearsdDNA	circularsdDNA	dsDNA	+RNA	ssRNA
genome size (kb)	7.9	3	9.5-12.5	170	170	5.4	9	9.2-9.8
virus size (nm)	52-60	52-55	40-60	200	100-150	40-55	100	100
envelope capsid	Absent Isosahedral	Present Isosahedral	Present Isosahedral	Present Isosahedral	Present Isosahedral	Absent Isosahedral	Present Isosahedral	Present Isosahedral
vaccination	Accessible	Accessible	Inaccessible	Inaccessible	Inaccessible	Inaccessible	Inaccessible	Inaccessible
anti-viral treatment	Not established	Effective	Effective	Occasionally Effective	Not established	Not established	Not established	Not established
diagnostic molecule	p16	HBs antigen	anti-HCV antibody	EBER	LANA	CM2B4	ATLA	p24
Integration	Yes	Yes	No	Yes	No	Yes	Yes	Yes

Table 1.4 Tumor viruses and their associated carcinomas

Virus	Target	Cancer type
High-risk HPVs	Uterine Cervix Head and neck (Oropharynx) Vagina Vulva Penis	Squamous cell carcinoma Adenocarcinoma
HBV	Liver	Hepatocellular carcinoma Cholangiocellular carcinoma
HCV	Liver Hematopoietic system	Hepatocellular carcinoma Cholangiocellular carcinoma Malignant lymphoma
EBV/HHV-4	Stomach Nasopharnx Hematopoietic system Soft tissue	Adenocarcinoma Nasopharyngeal carcinoma EBV-associated smooth muscle tumor
KSHV/HHV-8	Soft tissue Hematopoietic system	Kaposi sarcoma
MCV	Skin	Merkel cell carcinoma
HTLV-1	Hematopoietic system	Adult T-cell leukemia/lymphoma

Table 1.5 Technology for viral integrations

Technology[223]		Pros	Cons	Suitable cases	Relative Cost (1-4, low to high)
In-situ hybridization (ISH) [276,277]	(Fluorescence) in-situ hybridization ((F) ISH)	High sensitive; Fast results; Expensive than PCR but cheap than sequencing;	Cannot identify site of integration if only virus probe is used	RNA/DNA; Number of integration sites per nucleus	1
PCR	Quantitative or real time PCR (qPCR, RT-PCR) [278,279]	High specific; Extremely sensitive; Cheap than sequencing;	Cannot identify the site of integration; E2/E7 ratio not always could be used as cut-off;	Fresh frozen material; Detect viral load	1
	Detection of Integrated Papillomavirus Sequences PCR (DIPS-PCR) [211,238,280,281]	Cheap than sequencing; Identify the site of integration	Only for E2 fractures; Less suitable for Formalin-Fixed Paraffin-Embedded (FFPE) material	Fresh frozen material;	1
	Amplification of Papillomavirus Oncogene Transcripts PCR (APOT-PCR) [211,238,280,281]	Cheap than sequencing; Identify the site of integration; High accurate; High sensitive; Identify the viral copy number	Less suitable for FFPE materials; Require stable RNA;	RNA; Fresh frozen material;	1
NGS	RNA-Seq [210,276]	Deep profile the transcriptome; Identify the site of integration	Cannot find 5'end of HPV breakpoints; Only expressed fusions	RNA; Blood, fresh-frozen, FFPE, fine needle aspirates, core needle biopsies and single cells	2

	WGS [210,282,283]	High accurate; High sensitive; Identify the site of integration; Identify the viral copy number;	Need good coverage; Expensive than PCR and FISH; Time consuming	DNA;	3
	WES[210,282,283]	Highly accurate; Extremely sensitive; Cheap than WGS; Identify the site of integration; Identify the viral copy number;	Less suitable for FFPE materials; Need good coverage; Only in exome	DNA; Blood and fresh-frozen biopsy;	2
	Capture-based assay [210,284]	Identify the site of integration; Identify the viral copy number; Increases chance of finding HPV integration sites; High coverage	Need good coverage;	DNA/RNA; Blood, fresh-frozen, FFPE, fine needle aspirates, core needle biopsies	2
Emerging Techniques	Targeted Locus Amplification [285]	Long read; Identify the site of integration; Identify the viral copy number; Increases chance of finding HPV integration sites;		Fresh-frozen, FFPE	4
	Long read sequencing [216]	Long read; More comprehensive for large, complex, duplicated, HPV integration events; Identify the site of integration; Identify the viral copy number;	Less suitable for FFPE materials; Not suitable for single nucleotide variation detection; High base-calling rate; Need good coverage	DNA/RNA	4

Table 1.6 Summary for Viral integration callers. Abbreviations: PE: pair-end; WGS: whole genome sequencing; HCC: hepatocellular carcinoma; HNSCC: head and neck squamous cell carcinoma; HPV: human papillomavirus; HBV: hepatitis B virus

Tool	Sequencing data	Benchmarking data	Alignment	Informative reads	Additional feature
SeqMap2 [262]	Roche 454	NA	BLAT	Split	Web server platform
VirusSeq [263]	Illumina	17 HCC cancers PE RNA-Seq data; 239 HNSCC PE RNA-Seq	BWA-SW	Paired-end	Detect Virus
VirusFinder1-2 [264,265]	Illumina	VirusFinder1: 10 PE WGS samples with HBV virus VirusFinder2: 13 HCC cancers PE WGS; 4 HCC cell lines RNA-Seq; 2 targeted sequencing Merkel cell carcinomas	BWA-ALN	Paired-end & Split	Detect Virus
ViralFusionSeq [266]	Illumina	1 HCC cell line RNA-Seq; Simulated WGS; 2 HCC WGS	BWA-SW	Paired-end & Split	Reconstructing transcript sequences

Virus-Clip [267]	Illumina	2 HCC RNA-Seq	Host:BLASTN Virus: BWA-MEM	Split	
Vy-PER [268]	Illumina	1 Simulated WGS; One liver cancer PE RNA-Seq; 2 PE WGS liver cancers	Host: BWA-ALN Virus: BLAT	Paired-end	
BATVI [269]	Illumina	Simulated data; 2 PE WGS HCC cancers	Host: BLASTN Virus: BatMis& BLATN	Paired-end & Split	
HGT-ID [212]	Illumina	Simulated data; 4 HPV positive samples WGS PE; 13 HBV positive HCC samples PE WGS ; 7 PE WGS, RNA-Seq paired HCC samples	BWA-mem	Paired-end & Split (only soft-clipped)	
ViFi [273]	Illumina	68 WGS, RNA-Seq paired CESC samples, 6 RNA-Seq HCC samples; 20 HCC WGS samples; 96 HPV WGS simulated samples	BWA-MEM	Paired-end	
VirTect [270]	Illumina	Simulated data; 1 WES HBV HCC sample;	BWA-SW	Paired-end & Split (only soft-clipped	

		9 WGS HCC samples			
HIVID2 [271]	Illumina	Simulated data; 134 cervical cancers PE WGS ; 852 PE WGS HCC samples	BWA-MEM	Paired-end & Split	
SurVirus [274]	Illumina	Simulated data; 135 cervical cancers WGS PE, 10 paired PE RNA-Seq; 426 WGS HCC patients, 12 paired RNA-Seq; 88 HCC WGS PE	BWA-MEM	Paired-end	
SearcHPV [236]	Illumina	3 targeted capture sequencing samples; 2 paired WES	BWA-MEM	Paired-end & Split	Assemble local contig
VIRUSBreakend [275]	Illumina	Simulated data; 22 WGS HBV; 5191 WGS tumor samples	BWA	Paired-end & Split	Detect viral integrations including regions such as centromeres and telomeres

Bibliography

1. Bergonzini V, Salata C, Calistri A, Parolin C, Palù G. View and review on viral oncology research. Infect Agent Cancer. 2010;5:11.

2. Symer DE, Akagi K, Geiger HM, et al. Diverse tumorigenic consequences of human papillomavirus integration in primary oropharyngeal cancers. Genome Res. 2022;32(1):55-70.

3. Hu Z, Zhu D, Wang W, et al. Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism. Nature Genetics. 2015;47(2):158-163. doi:10.1038/ng.3178

4. McLaughlin-Drubin ME, Munger K. Viruses associated with human cancer. Biochim Biophys Acta. 2008;1782(3):127-150.

5. M'Fadyean J, Hobday F. Note on the Experimental Transmission of Warts in the Dog. J Comp Pathol Ther. 1898;11:341-344

6. Ciuffo G. Innesto positivo con filtrato di verruca volgare. Giorn Ital Mal Venereol. Published online 1907.

7. Bang O, Ellermann V. Experimentelle Leukämie bei Hühnern. Zentralbl Bakteriol Parasitenkd Infectionskr Hyg Abt. Published online 1909.

8. Rous P. Transmission of a Malignant New Growth by Means of a Cell-Free Filtrate. JAMA. 1983;250(11):1445-1446.

9. Gross L. Susceptibility of Newborn Mice of an Otherwise Apparently "Resistant" Strain to Inoculation with Leukemia. Proc Soc Exp Biol Med. 1950;73(2):246-248.

10. Stewart SE, Eddy BE, Borgese N. Neoplasms in mice inoculated with a tumor agent carried in tissue culture. J Natl Cancer Inst. 1958;20(6):1223-1243.

11. Rous P, Beard JW. THE PROGRESSION TO CARCINOMA OF VIRUS-INDUCED RABBIT PAPILLOMAS (SHOPE). J Exp Med. 1935;62(4):523-548.

12. Friend C. Cell-free transmission in adult Swiss mice of a disease having the character of a leukemia. J Exp Med. 1957;105(4):307-318.

13. Graffi A. Chloroleukemia of mice. Ann N Y Acad Sci. 1957;68(2):540-558.

14. Moloney JB. Biological studies on a lymphoid-leukemia virus extracted from sarcoma 37. I. Origin and introductory investigations. J Natl Cancer Inst. 1960;24:933-951.

15. Eddy BE, Borman GS, Grubbs GE, Young R. Identification of the oncogenic substance in rhesus monkey kidney cell cultures as simian virus 40. Virology. 1962;17(1):65-75.

16. Girardi AJ, Sweet BH, Slotnick VB, Hilleman MR. Development of Tumors in Hamsters Inoculated in the Neonatal Period with Vacuolating Virus, SV40. Proc Soc Exp Biol Med. 1962;109(3):649-660.

17. Trentin JJ, Yabe Y, Taylor G. The quest for human cancer viruses. Science. 1962;137(3533):835-841.

18. Epstein MA, Achong BG, Barr YM. VIRUS PARTICLES IN CULTURED LYMPHOBLASTS FROM BURKITT'S LYMPHOMA. Lancet. 1964;1(7335):702-703.

19. Okochi K, Murakami S. Observations on Australia antigen in Japanese. Vox Sang. 1968;15(5):374-385.

20. Prince AM. An antigen detected in the blood during the incubation period of serum hepatitis. Proc Natl Acad Sci U S A. 1968;60(3):814-821.

21. Blumberg BS, Larouzé B, London WT, et al. The relation of infection with the hepatitis B agent to primary hepatic carcinoma. Am J Pathol. 1975;81(3):669-682.

22. zur Hausen H. Condylomata acuminata and human genital cancer. Cancer Res. 1976;36(2 pt 2):794.

23. zur Hausen H, Meinhof W, Scheiber W, Bornkamm GW. Attempts to detect virus-secific DNA in human tumors. I. Nucleic acid hybridizations with complementary RNA of human wart virus. Int J Cancer. 1974;13(5):650-656.

24. Boshart M, Gissmann L, Ikenberg H, Kleinheinz A, Scheurlen W, zur Hausen H. A new type of papillomavirus DNA, its presence in genital cancer biopsies and in cell lines derived from cervical cancer. EMBO J. 1984;3(5):1151-1157 - 1157.

25. Dürst M, Gissmann L, Ikenberg H, zur Hausen H. A papillomavirus DNA from a cervical carcinoma and its prevalence in cancer biopsy samples from different geographic regions. Proc Natl Acad Sci U S A. 1983;80(12):3812-3815.

26. Poiesz BJ, Ruscetti FW, Gazdar AF, Bunn PA, Minna JD, Gallo RC. Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. Proc Natl Acad Sci U S A. 1980;77(12):7415-7419.

27. Hinuma Y, Nagata K, Hanaoka M, et al. Adult T-cell leukemia: antigen in an ATL cell line and detection of antibodies to the antigen in human sera. Proc Natl Acad Sci U S A. 1981;78(10):6476-6480.

28. Choo QL, Kuo G, Weiner AJ, Overby LR, Bradley DW, Houghton M. Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. Science. 1989;244(4902):359-362.

29. Alter HJ, Purcell RH, Shih JW, et al. Detection of antibody to hepatitis C virus in prospectively followed transfusion recipients with acute and chronic non-A, non-B hepatitis. N Engl J Med. 1989;321(22):1494-1500.

30. Kuo G, Choo QL, Alter HJ, et al. An assay for circulating antibodies to a major etiologic virus of human non-A, non-B hepatitis. Science. 1989;244(4902):362-364.

31. Colombo M, Kuo G, Choo QL, et al. Prevalence of antibodies to hepatitis C virus in Italian patients with hepatocellular carcinoma. Lancet. 1989;2(8670):1006-1008.

32. Tan A, Yeh SH, Liu CJ, Cheung C, Chen PJ. Viral hepatocarcinogenesis: from infection to cancer. Liver Int. 2008;28(2):175-188.

33. Ganem D. KSHV infection and the pathogenesis of Kaposi's sarcoma. Annu Rev Pathol. 2006;1:273-296.

34. Beral V, Peterman TA, Berkelman RL, Jaffe HW. Kaposi's sarcoma among persons with AIDS: a sexually transmitted infection? Lancet. 1990;335(8682):123-128.

35. Chang Y, Cesarman E, Pessin MS, et al. Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. Science. 1994;266(5192):1865-1869.

36. Zur Hausen H. Novel human polyomaviruses--re-emergence of a well known virus family as possible human carcinogens. Int J Cancer. 2008;123(2):247-250.

37. Feng H, Shuda M, Chang Y, Moore PS. Clonal integration of a polyomavirus in human Merkel cell carcinoma. Science. 2008;319(5866):1096-1100.

38. Yang A, Wijaya WA, Yang L, He Y, Cen Y, Chen J. The impact of merkel cell polyomavirus positivity on prognosis of merkel cell carcinoma: A systematic review and meta-analysis. Front Oncol. 2022;12:1020805.

39. Dimitraki MG, Sourvinos G. Merkel Cell Polyomavirus (MCPyV) and Cancers: Emergency Bell or False Alarm? Cancers . 2022;14(22). doi:10.3390/cancers14225548

40. Wijaya WA, Liu Y, Qing Y, Li Z. Prevalence of Merkel Cell Polyomavirus in Normal and Lesional Skin: A Systematic Review and Meta-Analysis. Front Oncol. 2022;12:868781.

41. Schiller JT, Lowy DR. An Introduction to Virus Infections and Human Cancer. Recent Results Cancer Res. 2021;217:1-11.

42. Evans AS, Mueller NE. Viruses and cancer. Causal associations. Ann Epidemiol. 1990;1(1):71-92.

43. Hill AB. THE ENVIRONMENT AND DISEASE: ASSOCIATION OR CAUSATION? Proc R Soc Med. 1965;58(5):295-300.

44. zur Hausen H. Oncogenic DNA viruses. Oncogene. 2001;20(54):7820-7823.

45. Hatano Y, Ideta T, Hirata A, et al. Virus-Driven Carcinogenesis. Cancers . 2021;13(11). doi:10.3390/cancers13112625

46. Bergonzini V, Salata C, Calistri A, Parolin C, Palù G. View and review on viral oncology research. Infectious Agents and Cancer. 2010;5(1). doi:10.1186/1750-9378-5-11

47. Carrillo-Infante C, Abbadessa G, Bagella L, Giordano A. Viral infections as a cause of cancer (Review). International Journal of Oncology. Published online 2007. doi:10.3892/ijo.30.6.1521

48. German Advisory Committee Blood (Arbeitskreis Blut), Subgroup “Assessment of Pathogens Transmissible by Blood.” Human Immunodeficiency Virus (HIV). Transfus Med Hemother. 2016;43(3):203-222.

49. Mesri EA, Feitelson MA, Munger K. Human viral oncogenesis: a cancer hallmarks analysis. Cell Host Microbe. 2014;15(3):266-282.

50. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-674.

51. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100(1):57-70.

52. Cesarman E, Damania B, Krown SE, Martin J, Bower M, Whitby D. Kaposi sarcoma. Nat Rev Dis Primers. 2019;5(1):9.

53. Schiller JT, Lowy DR. Virus infection and human cancer: an overview. Recent Results Cancer Res. 2014;193:1-10.

54. Bouvard V, Baan R, Straif K, et al. A review of human carcinogens--Part B: biological agents. Lancet Oncol. 2009;10(4):321-322.

55. Zur Hausen H. The search for infectious causes of human cancers: where and why. Virology. 2009;392(1):1-10.

56. Chen X, Kost J, Sulovari A, et al. A virome-wide clonal integration analysis platform for discovering cancer viral etiology. Genome Res. 2019;29(5):819-830.

57. Beachy PA, Karhadkar SS, Berman DM. Tissue repair and stem cell renewal in carcinogenesis. Nature. 2004;432(7015):324-331.

58. Hatano Y, Fukuda S, Hisamatsu K, Hirata A, Hara A, Tomita H. Multifaceted Interpretation of Colon Cancer Stem Cells. Int J Mol Sci. 2017;18(7). doi:10.3390/ijms18071446

59. Knijnenburg TA, Wang L, Zimmermann MT, et al. Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. Cell Rep. 2018;23(1):239-254.e6.

60. Alexandrov LB, Kim J, Haradhvala NJ, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578(7793):94-101.

61. Vogelstein B, Kinzler KW. The multistep nature of cancer. Trends Genet. 1993;9(4):138-141.

62. Zapatka M, Borozan I, Brewer DS, et al. The landscape of viral associations in human cancers. Nat Genet. 2020;52(3):320-330.

63. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, International Agency for Research on Cancer. Epstein-Barr Virus and Kaposi's Sarcoma Herpesvirus/human Herpesvirus 8. IARC; 1997.

64. Evans AS. Viral Infections of Humans: Epidemiology and Control. Springer Science & Business Media; 2013.

65. Ebell MH. Epstein-Barr virus infectious mononucleosis. Am Fam Physician. 2004;70(7):1279-1287.

66. Sixbey JW, Nedrud JG, Raab-Traub N, Hanes RA, Pagano JS. Epstein–Barr Virus Replication in Oropharyngeal Epithelial Cells. N Engl J Med. 1984;310(19):1225-1230.

67. Chan SH. Aetiology of nasopharyngeal carcinoma. Ann Acad Med Singapore. 1990;19(2):201-207.

68. Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet. 1996;58(6):1347-1363.

69. Camilleri-Broët S, Camparo P, Mokhtari K, et al. Overexpression of BCL-2, BCL-X, and BAX in primary central nervous system lymphomas that occur in immunosuppressed patients. Mod Pathol. 2000;13(2):158-165.

70. Wilson JB, Bell JL, Levine AJ. Expression of Epstein-Barr virus nuclear antigen-1 induces B cell neoplasia in transgenic mice. EMBO J. 1996;15(12):3117-3126 - 3126.

71. Levitskaya J, Coram M, Levitsky V, et al. Inhibition of antigen processing by the internal repeat region of the Epstein-Barr virus nuclear antigen-1. Nature. 1995;375(6533):685-688.

72. Abbot SD, Rowe M, Cadwallader K, et al. Epstein-Barr virus nuclear antigen 2 induces expression of the virus-encoded latent membrane protein. J Virol. 1990;64(5):2126-2134.

73. Gregory CD, Dive C, Henderson S, et al. Activation of Epstein–Barr virus latent genes protects human B cells from death by apoptosis. Nature. 1991;349(6310):612-614.

74. Laherty CD, Hu HM, Opipari AW, Wang F, Dixit VM. The Epstein-Barr virus LMP1 gene product induces A20 zinc finger protein expression by activating nuclear factor kappa B. J Biol Chem. 1992;267(34):24157-24160.

75. Mosialos G, Birkenbach M, Yalamanchili R, VanArsdale T, Ware C, Kieff E. The Epstein-Barr virus transforming protein LMP1 engages signaling proteins for the tumor necrosis factor receptor family. Cell. 1995;80(3):389-399.

76. Eliopoulos AG, Stack M, Dawson CW, et al. Epstein-Barr virus-encoded LMP1 and CD40 mediate IL-6 production in epithelial cells via an NF-kappaB pathway involving TNF receptor-associated factors. Oncogene. 1997;14(24):2899-2916.

77. Eliopoulos AG, Young LS. Activation of the cJun N-terminal kinase (JNK) pathway by the Epstein-Barr virus-encoded latent membrane protein 1 (LMP1). Oncogene. 1998;16(13):1731-1742.

78. Gires O, Kohlhuber F, Kilger E, et al. Latent membrane protein 1 of Epstein–Barr virus interacts with JAK3 and activates STAT proteins. EMBO J. 1999;18(11):3064-3073 - 3073.

79. Cen H, Williams PA, McWilliams HP, Breinig MC, Ho M, McKnight JL. Evidence for restricted Epstein-Barr virus latent gene expression and anti-EBNA antibody response in solid organ transplant recipients with posttransplant lymphoproliferative disorders. Blood. 1993;81(5):1393-1403.

80. Henderson A, Ripley S, Heller M, Kieff E. Chromosome site for Epstein-Barr virus DNA in a Burkitt tumor cell line and in lymphocytes growth-transformed in vitro. Proc Natl Acad Sci U S A. 1983;80(7):1987-1991.

81. Matsuo T, Heller M, Petti L, O'Shiro E, Kieff E. Persistence of the entire Epstein-Barr virus genome integrated into human lymphocyte DNA. Science. 1984;226(4680):1322-1325.

82. Lawrence JB, Villnave CA, Singer RH. Sensitive, high-resolution chromatin and chromosome mapping in situ: presence and orientation of two closely integrated copies of EBV in a lymphoma line. Cell. 1988;52(1):51-61.

83. Anvret M, Karlsson A, Bjursell G. Evidence for integrated EBV genomes in Raji cellular DNA. Nucleic Acids Res. 1984;12(2):1149-1161.

84. Delecluse HJ, Bartnizke S, Hammerschmidt W, Bullerdiek J, Bornkamm GW. Episomal and integrated copies of Epstein-Barr virus coexist in Burkitt lymphoma cell lines. J Virol. 1993;67(3):1292-1299.

85. Wolf J, Pawlita M, Klevenz B, et al. Down-regulation of integrated Epstein-Barr virus nuclear antigen 1 and 2 genes in a Burkitt lymphoma cell line after somatic cell fusion with autologous EBV-immortalized lymphoblastoid cells. Int J Cancer. 1993;53(4):621-627. doi:10.1002/ijc.2910530416

86. Hurley EA, Agger S, McNeil JA, et al. When Epstein-Barr virus persistently infects B-cell lines, it frequently integrates. J Virol. 1991;65(3):1245-1254.

87. Kripalani-Joshi S, Law HY. Identification of integrated Epstein-Barr virus in nasopharyngeal carcinoma using pulse field gel electrophoresis. Int J Cancer. 1994;56(2):187-192.

88. Chang Y, Cheng SD, Tsai CH. Chromosomal integration of Epstein-Barr virus genomes in nasopharyngeal carcinoma cells. Head Neck. 2002;24(2):143-150.

89. Cao S, Strong MJ, Wang X, et al. High-throughput RNA sequencing-based virome analysis of 50 lymphoma cell lines from the Cancer Cell Line Encyclopedia project. J Virol. 2015;89(1):713-729.

90. Peng RJ, Han BW, Cai QQ, et al. Genomic and transcriptomic landscapes of Epstein-Barr virus in extranodal natural killer T-cell lymphoma. Leukemia. 2019;33(6):1451-1462.

91. Chakravorty S, Yan B, Wang C, et al. Integrated Pan-Cancer Map of EBV-Associated Neoplasms Reveals Functional Host–Virus Interactions. Cancer Res. 2019;79(23):6010-6023.

92. European Association for the Study of the Liver. Electronic address: easloffice@easloffice.eu, European Association for the Study of the Liver. EASL Clinical Practice Guidelines: Management of hepatocellular carcinoma. J Hepatol. 2018;69(1):182-236.

93. Stella L, Santopaolo F, Gasbarrini A, Pompili M, Ponziani FR. Viral hepatitis and hepatocellular carcinoma: From molecular pathways to the role of clinical surveillance and antiviral treatment. World J Gastroenterol. 2022;28(21):2251-2281.

94. de Martel C, Ferlay J, Franceschi S, et al. Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. Lancet Oncol. 2012;13(6):607-615.

95. Sung WK, Zheng H, Li S, et al. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. Nat Genet. 2012;44(7):765-769.

96. Totoki Y, Tatsuno K, Covington KR, et al. Trans-ancestry mutational landscape of hepatocellular carcinoma genomes. Nat Genet. 2014;46(12):1267-1273.

97. Schulze K, Imbeaud S, Letouzé E, et al. Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. Nat Genet. 2015;47(5):505-511.

98. Cancer Genome Atlas Research Network. Electronic address: wheeler@bcm.edu, Cancer Genome Atlas Research Network. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell. 2017;169(7):1327-1341.e23.

99. Seeger C, Mason WS. Hepatitis B virus biology. Microbiol Mol Biol Rev. 2000;64(1):51-68.

100. Jiang Z, Jhunjhunwala S, Liu J, et al. The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. Genome Res. 2012;22(4):593-601.

101. Feitelson MA, Lee J. Hepatitis B virus integration, fragile sites, and hepatocarcinogenesis. Cancer Lett. 2007;252(2):157-170.

102. Bréchot C. Pathogenesis of hepatitis B virus-related hepatocellular carcinoma: old and new paradigms. Gastroenterology. 2004;127(5 Suppl 1):S56-S61.

103. Dandri M, Burda MR, Bürkle A, et al. Increase in de novo HBV DNA integrations in response to oxidative DNA damage or inhibition of poly(ADP-ribosyl)ation. Hepatology. 2002;35(1):217-223.

104. Li X, Zhang J, Yang Z, et al. The function of targeted host genes determines the oncogenicity of HBV integration in hepatocellular carcinoma. J Hepatol. 2014;60(5):975-984.

105. Zhao LH, Liu X, Yan HX, et al. Genomic and oncogenic preference of HBV integration in hepatocellular carcinoma. Nat Commun. 2016;7:12992.

106. Bréchot C, Gozuacik D, Murakami Y, Paterlini-Bréchot P. Molecular bases for the development of hepatitis B virus (HBV)-related hepatocellular carcinoma (HCC). Semin Cancer Biol. 2000;10(3):211-231.

107. Soussan P, Garreau F, Zylberberg H, Ferray C, Brechot C, Kremsdorf D. In vivo expression of a new hepatitis B virus protein encoded by a spliced RNA. J Clin Invest. 2000;105(1):55-60.

108. Lindenbach BD, Rice CM. The ins and outs of hepatitis C virus entry and assembly. Nat Rev Microbiol. 2013;11(10):688-700.

109. Paul D, Madan V, Bartenschlager R. Hepatitis C virus RNA replication and assembly: living on the fat of the land. Cell Host Microbe. 2014;16(5):569-579.

110. Kanwal F, Kramer JR, Ilyas J, Duan Z, El-Serag HB. HCV genotype 3 is associated with an increased risk of cirrhosis and hepatocellular cancer in a national sample of U.S. Veterans with HCV. Hepatology. 2014;60(1):98-105.

111. Banerjee A, Ray RB, Ray R. Oncogenic potential of hepatitis C virus proteins. Viruses. 2010;2(9):2108-2133.

112. Mahmoudvand S, Shokri S, Taherkhani R, Farshadpour F. Hepatitis C virus core protein modulates several signaling pathways involved in hepatocellular carcinoma. World J Gastroenterol. 2019;25(1):42-58.

113. Meyer K, Basu A, Saito K, Ray RB, Ray R. Inhibition of hepatitis C virus core protein expression in immortalized human hepatocytes induces cytochrome c-independent increase in Apaf-1 and caspase-9 activation for cell death. Virology. 2005;336(2):198-207.

114. Ray RB, Ray R. Hepatitis C virus manipulates humans as its favorite host for a long-term relationship. Hepatology. 2019;69(2):889-900.

115. Choi SH, Hwang SB. Modulation of the Transforming Growth Factor-β Signal Transduction Pathway by Hepatitis C Virus Nonstructural 5A Protein *. J Biol Chem. 2006;281(11):7468-7478.

116. Han Y, Niu J, Wang D, Li Y. Hepatitis C Virus Protein Interaction Network Analysis Based on Hepatocellular Carcinoma. PLoS One. 2016;11(4):e0153882.

117. Jiang YF, He B, Li NP, Ma J, Gong GZ, Zhang M. The oncogenic role of NS5A of hepatitis C virus is mediated by up-regulation of survivin gene expression in the hepatocellular cell through p53 and NF-κB pathways. Cell Biol Int. 2011;35(12):1225-1232.

118. Majumder M, Ghosh AK, Steele R, Ray R, Ray RB. Hepatitis C virus NS5A physically associates with p53 and regulates p21/waf1 gene expression in a p53-dependent manner. J Virol. 2001;75(3):1401-1407.

119. Irshad M, Gupta P, Irshad K. Molecular basis of hepatocellular carcinoma induced by hepatitis C virus infection. World J Hepatol. 2017;9(36):1305-1314.

120. Khatun M, Ray R, Ray RB. Hepatitis C virus associated hepatocellular carcinoma. Adv Cancer Res. 2021;149:103-142.

121. Moorman JP, Zhang CL, Ni L, et al. Impaired hepatitis B vaccine responses during chronic hepatitis C infection: Involvement of the PD-1 pathway in regulating CD4+ T cell responses. Vaccine. 2011;29(17):3169-3176.

122. Shi L, Wang JM, Ren JP, et al. KLRG1 impairs CD4+ T cell responses via p16ink4a and p27kip1 pathways: role in hepatitis B vaccine failure in individuals with hepatitis C virus infection. J Immunol. 2014;192(2):649-657.

123. Jones T, Ye F, Bedolla R, et al. Direct and efficient cellular transformation of primary rat mesenchymal precursor cells by KSHV. J Clin Invest. 2012;122(3):1076-1081.

124. Chen CJ, Hsu WL, Yang HI, et al. Epidemiology of virus infection and human cancer. Recent Results Cancer Res. 2014;193:11-32.

125. Miller G, Heston L, Grogan E, et al. Selective switch between latency and lytic replication of Kaposi's sarcoma herpesvirus and Epstein-Barr virus in dually infected body cavity lymphoma cells. J Virol. 1997;71(1):314-324.

126. Guito J, Lukac DM. KSHV reactivation and novel implications of protein isomerization on lytic switch control. Viruses. 2015;7(1):72-109.

127. Lukac DM, Yuan Y. Reactivation and lytic replication of KSHV. In: Arvin A, Campadelli-Fiume G, Mocarski E, et al., eds. Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis. Cambridge University Press.

128. Giffin L, Damania B. KSHV: pathways to tumorigenesis and persistent infection. Adv Virus Res. 2014;88:111-159.

129. Hughes DJ, Wood JJ, Jackson BR, Baquero-Pérez B, Whitehouse A. NEDDylation is essential for Kaposi's sarcoma-associated herpesvirus latency and lytic reactivation and represents a novel anti-KSHV target. PLoS Pathog. 2015;11(3):e1004771.

130. Kalt I, Masa SR, Sarid R. Linking the Kaposi's sarcoma-associated herpesvirus (KSHV/HHV-8) to human malignancies. Methods Mol Biol. 2009;471:387-407.

131. Aoki Y, Jaffe ES, Chang Y, et al. Angiogenesis and Hematopoiesis Induced by Kaposi's Sarcoma-Associated Herpesvirus-Encoded Interleukin-6: Presented in part at the 40th Annual American Society of Hematology Meeting, December 7, 1998 (Miami Beach, FL). Blood. 1999;93(12):4034-4043.

132. Ballestas ME, Chatis PA, Kaye KM. Efficient persistence of extrachromosomal KSHV DNA mediated by latency-associated nuclear antigen. Science. 1999;284(5414):641-644.

133. Friborg J Jr, Kong W, Hottiger MO, Nabel GJ. p53 inhibition by the LANA protein of KSHV protects against cell death. Nature. 1999;402(6764):889-894.

134. Fujimuro M, Wu FY, apRhys C, et al. A novel viral mechanism for dysregulation of β-catenin in Kaposi's sarcoma–associated herpesvirus latency. Nat Med. 2003;9(3):300-306.

135. Gallo RC. The first human retrovirus. Sci Am. 1986;255(6):88-98.

136. Zuo X, Zhou R, Yang S, Ma G. HTLV-1 persistent infection and ATLL oncogenesis. J Med Virol. 2023;95(1):e28424.

137. Zhao T, Wang Z, Fang J, et al. HTLV-1 activates YAP via NF-κB/p65 to promote oncogenesis. Proc Natl Acad Sci U S A. 2022;119(9). doi:10.1073/pnas.2115316119

138. Rende F, Cavallari I, Corradin A, et al. Kinetics and intracellular compartmentalization of HTLV-1 gene expression: nuclear retention of HBZ mRNAs. Blood. 2011;117(18):4855-4859.

139. Satou Y, Utsunomiya A, Tanabe J, Nakagawa M, Nosaka K, Matsuoka M. HTLV-1 modulates the frequency and phenotype of FoxP3+CD4+T cells in virus-infected individuals. Retrovirology. 2012;9(1):46.

140. Satou Y, Yasunaga JI, Yoshida M, Matsuoka M. HTLV-I basic leucine zipper factor gene mRNA supports proliferation of adult T cell leukemia cells. Proc Natl Acad Sci U S A. 2006;103(3):720-725.

141. Koya J, Saito Y, Kameda T, et al. Single-Cell Analysis of the Multicellular Ecosystem in Viral Carcinogenesis by HTLV-1. Blood Cancer Discov. 2021;2(5):450-467.

142. Accolla RS, Jacobson S, Willems L, Watanabe T. Human T Cell Leukemia Virus-1 (HTLV-1) Infection, Associated Pathology and Response of the Host. Frontiers Media SA; 2023.

143. Barski MS, Minnell JJ, Hodakova Z, et al. Cryo-EM structure of the deltaretroviral intasome in complex with the PP2A regulatory subunit B56γ. Nat Commun. 2020;11(1):5043.

144. Gillet NA, Malani N, Melamed A, et al. The host genomic environment of the provirus determines the abundance of HTLV-1-infected T-cell clones. Blood. 2011;117(11):3113-3122.

145. Melamed A, Laydon DJ, Gillet NA, Tanaka Y, Taylor GP, Bangham CRM. Genome-wide determinants of proviral targeting, clonal abundance and expression in natural HTLV-1 infection. PLoS Pathog. 2013;9(3):e1003271.

146. Cook LB, Melamed A, Niederer H, et al. The role of HTLV-1 clonality, proviral structure, and genomic integration site in adult T-cell leukemia/lymphoma. Blood. 2014;123(25):3925-3931.

147. Pipas JM. Common and unique features of T antigens encoded by the polyomavirus group. J Virol. 1992;66(7):3979-3985.

148. Shuda M, Feng H, Kwun HJ, et al. T antigen mutations are a human tumor-specific signature for Merkel cell polyomavirus. Proc Natl Acad Sci U S A. 2008;105(42):16272-16277.

149. Ludlow JW, Shon J, Pipas JM, Livingston DM, DeCaprio JA. The retinoblastoma susceptibility gene product undergoes cell cycle-dependent dephosphorylation and binding to and release from SV40 large T. Cell. 1990;60(3):387-396.

150. Houben R, Adam C, Baeurle A, et al. An intact retinoblastoma protein-binding site in Merkel cell polyomavirus large T antigen is required for promoting growth of Merkel cell carcinoma cells. Int J Cancer. 2012;130(4):847-856.

151. Kwun HJ, Guastafierro A, Shuda M, et al. The minimum replication origin of merkel cell polyomavirus has a unique large T-antigen loading architecture and requires small T-antigen expression for optimal replication. J Virol. 2009;83(23):12118-12128.

152. Erstad DJ, Cusack JC Jr. Mutational analysis of merkel cell carcinoma. Cancers . 2014;6(4):2116-2136.

153. Paulson KG, Lemos BD, Feng B, et al. Array-CGH reveals recurrent genomic changes in Merkel cell carcinoma including amplification of L-Myc. J Invest Dermatol. 2009;129(6):1547-1555.

154. Pinatti LM, Walline HM, Carey TE. Human Papillomavirus Genome Integration and Head and Neck Cancer. J Dent Res. 2018;97(6):691-700.

155. McQuillan G, Kruszon-Moran D, Markowitz LE, Unger ER, Paulose-Ram R. Prevalence of HPV in Adults Aged 18-69: United States, 2011-2014. NCHS Data Brief. 2017;(280):1-8.

156. CDC. Centers for disease control and prevention. Centers for Disease Control and Prevention. Published January 5, 2024. Accessed January 8, 2024. https://www.cdc.gov/

157. Syrjänen S. Human papillomaviruses in head and neck carcinomas. N Engl J Med. 2007;356(19):1993-1995.

158. zur Hausen H. Papillomaviruses in the causation of human cancers - a brief historical account. Virology. 2009;384(2):260-265.

159. Harari A, Chen Z, Burk RD. Human papillomavirus genomics: past, present and future. Curr Probl Dermatol. 2014;45:1-18.

160. Scheffner M, Werness BA, Huibregtse JM, Levine AJ, Howley PM. The E6 oncoprotein encoded by human papillomavirus types 16 and 18 promotes the degradation of p53. Cell. 1990;63(6):1129-1136.

161. Sathish N, Abraham P, Peedicayil A, Sridharan G, John S, Chandy G. Human papillomavirus 16 E6/E7 transcript and E2 gene status in patients with cervical neoplasia. Mol Diagn. 2004;8(1):57-64.

162. Boyer SN, Wazer DE, Band V. E7 protein of human papilloma virus-16 induces degradation of retinoblastoma protein through the ubiquitin-proteasome pathway. Cancer Res. 1996;56(20):4620-4624.

163. Wiest T, Schwarz E, Enders C, Flechtenmacher C, Bosch FX. 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):1510-1517.

164. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3):209-249.

165. Stolnicu S, Barsan I, Hoang L, et al. International Endocervical Adenocarcinoma Criteria and Classification (IECC): A New Pathogenetic Classification for Invasive Adenocarcinomas of the Endocervix. Am J Surg Pathol. 2018;42(2):214-226.

166. Wentzensen N, Vinokurova S, von Knebel Doeberitz M. 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):3878-3884.

167. Cancer Genome Atlas Research Network, Albert Einstein College of Medicine, Analytical Biological Services, et al. Integrated genomic and molecular characterization of cervical cancer. Nature. 2017;543(7645):378-384.

168. Li W, Lei W, Chao X, et al. Genomic alterations caused by HPV integration in a cohort of Chinese endocervical adenocarcinomas. Cancer Gene Ther. 2021;28(12):1353-1364.

169. Nambaru L, Meenakumari B, Swaminathan R, Rajkumar T. Prognostic significance of HPV physical status and integration sites in cervical cancer. Asian Pac J Cancer Prev. 2009;10(3):355-360.

170. Joo J, Shin HJ, Park B, et al. Integration Pattern of Human Papillomavirus Is a Strong Prognostic Factor for Disease-Free Survival After Radiation Therapy in Cervical Cancer Patients. Int J Radiat Oncol Biol Phys. 2017;98(3):654-661.

171. Ibragimova M, Tsyganov M, Shpileva O, et al. HPV status and its genomic integration affect survival of patients with cervical cancer. Neoplasma. 2018;65(3):441-448.

172. el Awady MK, Kaplan JB, O'Brien SJ, Burk RD. Molecular analysis of integrated human papillomavirus 16 sequences in the cervical cancer cell line SiHa. Virology. 1987;159(2):389-398.

173. Ziemann F, Arenz A, Preising S, 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. Am J Cancer Res. 2015;5(3):1017-1031.

174. Ferber MJ, Thorland EC, Brink AATP, et al. Preferential integration of human papillomavirus type 18 near the c-myc locus in cervical carcinoma. Oncogene. 2003;22(46):7233-7242.

175. Schmitz M, Driesch C, Jansen L, Runnebaum IB, Dürst M. Non-random integration of the HPV genome in cervical cancer. PLoS One. 2012;7(6):e39632.

176. Bodelon C, Untereiner ME, Machiela MJ, Vinokurova S, Wentzensen N. Genomic characterization of viral integration sites in HPV-related cancers. Int J Cancer. 2016;139(9):2001-2011.

177. Kamal M, Lameiras S, Deloger M, et al. Human papilloma virus (HPV) integration signature in Cervical Cancer: identification of MACROD2 gene as HPV hot spot integration site. Br J Cancer. 2021;124(4):777-785.

178. Warburton A, Markowitz TE, Katz JP, Pipas JM, McBride AA. Recurrent integration of human papillomavirus genomes at transcriptional regulatory hubs. NPJ Genom Med. 2021;6(1):101.

179. Zhou L, Qiu Q, Zhou Q, et al. Long-read sequencing unveils high-resolution HPV integration and its oncogenic progression in cervical cancer. Nat Commun. 2022;13(1):2563.

180. Wang C, Bai R, Liu Y, et al. Multi-region sequencing depicts intratumor heterogeneity and clonal evolution in cervical cancer. Med Oncol. 2023;40(2):78.

181. McBride AA, Warburton A. The role of integration in oncogenic progression of HPV-associated cancers. PLoS Pathog. 2017;13(4):e1006211.

182. Akagi K, Li J, Broutian TR, et al. Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. Genome Res. 2014;24(2):185-199.

183. Akagi K, Symer DE, Mahmoud M, et al. Intratumoral Heterogeneity and Clonal Evolution Induced by HPV Integration. Cancer Discov. 2023;13(4):910-927.

184. Tian R, Huang Z, Li L, et al. HPV integration generates a cellular super-enhancer which functions as ecDNA to regulate genome-wide transcription. Nucleic Acids Res. 2023;51(9):4237-4251.

185. zur Hausen H. Host cell regulation of HPV transforming gene expression. Princess Takamatsu Symp. 1989;20:207-219.

186. Choo KB, Pan CC, Han SH. Integration of human papillomavirus type 16 into cellular DNA of cervical carcinoma: preferential deletion of the E2 gene and invariable retention of the long control region and the E6/E7 open reading frames. Virology. 1987;161(1):259-261.

187. Kahla S, Kochbati L, Chanoufi MB, Maalej M, Oueslati R. HPV-16 E2 physical status and molecular evolution in vivo in cervical carcinomas. Int J Biol Markers. 2014;29(1):e78-e85.

188. Cricca M, Venturoli S, Leo E, Costa S, Musiani M, Zerbini M. Disruption of HPV 16 E1 and E2 genes in precancerous cervical lesions. J Virol Methods. 2009;158(1-2):180-183.

189. Tornesello ML, Buonaguro L, Giorgi-Rossi P, Buonaguro FM. Viral and cellular biomarkers in the diagnosis of cervical intraepithelial neoplasia and cancer. Biomed Res Int. 2013;2013:519619.

190. Liu M, Han Z, Zhi Y, et al. Long-read sequencing reveals oncogenic mechanism of HPV-human fusion transcripts in cervical cancer. Transl Res. 2023;253:80-94.

191. Warburton A, Redmond CJ, Dooley KE, et al. HPV integration hijacks and multimerizes a cellular enhancer to generate a viral-cellular super-enhancer that drives high viral oncogene expression. PLoS Genet. 2018;14(1):e1007179.

192. Dooley KE, Warburton A, McBride AA. Tandemly Integrated HPV16 Can Form a Brd4-Dependent Super-Enhancer-Like Element That Drives Transcription of Viral Oncogenes. MBio. 2016;7(5). doi:10.1128/mBio.01446-16

193. Fan J, Fu Y, Peng W, et al. Multi-omics characterization of silent and productive HPV integration in cervical cancer. Cell Genom. 2023;3(1):100211.

194. Scarpini CG, Groves IJ, Pett MR, Ward D, Coleman N. Virus transcript levels and cell growth rates after naturally occurring HPV16 integration events in basal cervical keratinocytes. J Pathol. 2014;233(3):281-293.

195. Gray E, Pett MR, Ward D, et al. In vitro progression of human papillomavirus 16 episome-associated cervical neoplasia displays fundamental similarities to integrant-associated carcinogenesis. Cancer Res. 2010;70(10):4081-4091.

196. Ojesina AI, Lichtenstein L, Freeman SS, et al. Landscape of genomic alterations in cervical carcinomas. Nature. 2014;506(7488):371-375.

197. Oyervides-Muñoz MA, Pérez-Maya AA, Rodríguez-Gutiérrez HF, et al. Understanding the HPV integration and its progression to cervical cancer. Infect Genet Evol. 2018;61:134-144.

198. Khoury JD, Tannir NM, Williams MD, et al. Landscape of DNA virus associations across human malignant cancers: analysis of 3,775 cases using RNA-Seq. J Virol. 2013;87(16):8916-8926.

199. Zeng X, Wang Y, Liu B, et al. Multi-omics data reveals novel impacts of human papillomavirus integration on the epigenomic and transcriptomic signatures of cervical tumorigenesis. J Med Virol. 2023;95(5):e28789.

200. Dürst M, Croce CM, Gissmann L, Schwarz E, Huebner K. Papillomavirus sequences integrate near cellular oncogenes in some cervical carcinomas. Proc Natl Acad Sci U S A. 1987;84(4):1070-1074.

201. Brant AC, Menezes AN, Felix SP, de Almeida LM, Sammeth M, Moreira MAM. 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):1853-1861.

202. Ehrig F, Häfner N, Driesch C, et al. Differences in Stability of Viral and Viral-Cellular Fusion Transcripts in HPV-Induced Cervical Cancers. Int J Mol Sci. 2019;21(1). doi:10.3390/ijms21010112

203. Fernandez AF, Rosales C, Lopez-Nieva P, et al. The dynamic DNA methylomes of double-stranded DNA viruses associated with human cancer. Genome Res. 2009;19(3):438-451.

204. Chaiwongkot A, Vinokurova S, Pientong C, et al. Differential methylation of E2 binding sites in episomal and integrated HPV 16 genomes in preinvasive and invasive cervical lesions. Int J Cancer. 2013;132(9):2087-2094.

205. Vigneswaran N, Williams MD. Epidemiologic trends in head and neck cancer and aids in diagnosis. Oral Maxillofac Surg Clin North Am. 2014;26(2):123-141.

206. Leemans CR, Braakhuis BJM, Brakenhoff RH. The molecular biology of head and neck cancer. Nat Rev Cancer. 2011;11(1):9-22.

207. Rietbergen MM, Leemans CR, Bloemena E, et al. Increasing prevalence rates of HPV attributable oropharyngeal squamous cell carcinomas in the Netherlands as assessed by a validated test algorithm. Int J Cancer. 2013;132(7):1565-1571.

208. Faraji F, Zaidi M, Fakhry C, Gaykalova DA. Molecular mechanisms of human papillomavirus-related carcinogenesis in head and neck cancer. Microbes Infect. 2017;19(9-10):464-475.

209. Speel EJM. HPV Integration in Head and Neck Squamous Cell Carcinomas: Cause and Consequence. Recent Results Cancer Res. 2017;206:57-72.

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

211. Olthof NC, Speel EJM, Kolligs J, 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):e88718.

212. Baheti S, Tang X, O'Brien DR, et al. HGT-ID: an efficient and sensitive workflow to detect human-viral insertion sites using next-generation sequencing data. BMC Bioinformatics. 2018;19(1):271.

213. Castellsagué X, Alemany L, Quer M, et al. HPV Involvement in Head and Neck Cancers: Comprehensive Assessment of Biomarkers in 3680 Patients. J Natl Cancer Inst. 2016;108(6):djv403. Published 2016 Jan 28. doi:10.1093/jnci/djv403

214. Hafkamp HC, Speel EJM, Haesevoets A, et al. A subset of head and neck squamous cell carcinomas exhibits integration of HPV 16/18 DNA and overexpression of p16INK4A and p53 in the absence of mutations in p53 exons 5-8. Int J Cancer. 2003;107(3):394-400.

215. Tang KD, Baeten K, Kenny L, Frazer IH, Scheper G, Punyadeera C. Unlocking the Potential of Saliva-Based Test to Detect HPV-16-Driven Oropharyngeal Cancer. Cancers . 2019;11(4). doi:10.3390/cancers11040473

216. Kono N, Arakawa K. Nanopore sequencing: Review of potential applications in functional genomics. Dev Growth Differ. 2019;61(5):316-326.

217. Parfenov M, Pedamallu CS, Gehlenborg N, 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):15544-15549.

218. Deng Z, Hasegawa M, Kiyuna A, et al. Viral load, physical status, and E6/E7 mRNA expression of human papillomavirus in head and neck squamous cell carcinoma. Head Neck. 2013;35(6):800-808.

219. Alsahafi E, Begg K, Amelio I, et al. Clinical Update on Head and Neck Cancer: Molecular Biology and Ongoing Challenges. KOPS Universität Konstanz; 2019.

220. Olthof NC, Straetmans JMJAA, Snoeck R, Ramaekers FCS, Kremer B, Speel EJM. Next-generation treatment strategies for human papillomavirus-related head and neck squamous cell carcinoma: where do we go? Rev Med Virol. 2012;22(2):88-105.

221. Elrefaey S, Massaro MA, Chiocca S, Chiesa F, Ansarin M. HPV in oropharyngeal cancer: the basics to know in clinical practice. Acta Otorhinolaryngol Ital. 2014;34(5):299-309.

222. Olthof NC, Huebbers CU, Kolligs J, et al. Viral load, gene expression and mapping of viral integration sites in HPV16-associated HNSCC cell lines. Int J Cancer. 2015;136(5):E207-E218.

223. Balaji H, Demers I, Wuerdemann N, et al. Causes and Consequences of HPV Integration in Head and Neck Squamous Cell Carcinomas: State of the Art. Cancers . 2021;13(16). doi:10.3390/cancers13164089

224. Nulton TJ, Kim NK, DiNardo LJ, Morgan IM, Windle B. Patients with integrated HPV16 in head and neck cancer show poor survival. Oral Oncol. 2018;80:52-55.

225. Hajek M, Sewell A, Kaech S, Burtness B, Yarbrough WG, Issaeva N. TRAF3/CYLD mutations identify a distinct subset of human papillomavirus-associated head and neck squamous cell carcinoma. Cancer. 2017;123(10):1778-1790.

226. Veitía D, Liuzzi J, Ávila M, Rodriguez I, Toro F, Correnti M. Association of viral load and physical status of HPV-16 with survival of patients with head and neck cancer. Ecancermedicalscience. 2020;14:1082.

227. Koneva LA, Zhang Y, Virani S, et al. HPV Integration in HNSCC Correlates with Survival Outcomes, Immune Response Signatures, and Candidate Drivers. Mol Cancer Res. 2018;16(1):90-102.

228. Vojtechova Z, Sabol I, Salakova M, 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):386-395.

229. Lim MY, Dahlstrom KR, Sturgis EM, Li G. Human papillomavirus integration pattern and demographic, clinical, and survival characteristics of patients with oropharyngeal squamous cell carcinoma. Head Neck. 2016;38(8):1139-1144.

230. Pinatti LM, Sinha HN, Brummel CV, et al. Association of human papillomavirus integration with better patient outcomes in oropharyngeal squamous cell carcinoma. Head Neck. 2021;43(2):544-557.

231. Kelley DZ, Flam EL, Izumchenko E, et al. Integrated Analysis of Whole-Genome ChIP-Seq and RNA-Seq Data of Primary Head and Neck Tumor Samples Associates HPV Integration Sites with Open Chromatin Marks. Cancer Res. 2017;77(23):6538-6550.

232. Walline HM, Komarck CM, McHugh JB, et al. Genomic Integration of High-Risk HPV Alters Gene Expression in Oropharyngeal Squamous Cell Carcinoma. Mol Cancer Res. 2016;14(10):941-952.

233. Rusan M, Li YY, Hammerman PS. Genomic landscape of human papillomavirus-associated cancers. Clin Cancer Res. 2015;21(9):2009-2019.

234. Pinatti LM, Walline HM, Carey TE, Klussmann JP, Huebbers CU. Viral Integration Analysis Reveals Likely Common Clonal Origin of Bilateral HPV16-Positive, p16-Positive Tonsil Tumors. Arch Clin Med Case Rep. 2020;4(4):680-696.

235. Leeman JE, Li Y, Bell A, et al. Human papillomavirus 16 promotes microhomology-mediated end-joining. Proc Natl Acad Sci U S A. 2019;116(43):21573-21579.

236. Pinatti LM, Gu W, Wang Y, et al. SearcHPV: A novel approach to identify and assemble human papillomavirus-host genomic integration events in cancer. Cancer. 2021;127(19):3531-3540.

237. Cao Y, Haring CT, Brummel C, et al. Early HPV ctDNA Kinetics and Imaging Biomarkers Predict Therapeutic Response in p16+ Oropharyngeal Squamous Cell Carcinoma. Clin Cancer Res. 2022;28(2):350-359.

238. Huebbers CU, Verhees F, Poluschkin L, 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):2465-2477.

239. Zhang Y, Koneva LA, Virani S, 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):4735-4745.

240. Paget-Bailly P, Meznad K, Bruyère D, et al. Comparative RNA sequencing reveals that HPV16 E6 abrogates the effect of E6*I on ROS metabolism. Sci Rep. 2019;9(1):5938.

241. Qin T, Koneva LA, Liu Y, 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):2375-2389.

242. Pannone G, Bufo P, Pace M, et al. TLR4 down-regulation identifies high risk HPV infection and integration in head and neck squamous cell carcinomas. Front Biosci . 2016;8(1):15-28.

243. Yang W, Liu Y, Dong R, et al. Accurate Detection of HPV Integration Sites in Cervical Cancer Samples Using the Nanopore MinION Sequencer Without Error Correction. Front Genet. 2020;11:660.

244. Wanichwatanadecha P, Sirisrimangkorn S, Kaewprag J, Ponglikitmongkol M. Transactivation activity of human papillomavirus type 16 E6*I on aldo-keto reductase genes enhances chemoresistance in cervical cancer cells. J Gen Virol. 2012;93(Pt 5):1081-1092.

245. Safe S, Jin UH, Hedrick E, Reeder A, Lee SO. Minireview: role of orphan nuclear receptors in cancer and potential as drug targets. Mol Endocrinol. 2014;28(2):157-172.

246. Hassounah NB, Malladi VS, Huang Y, et al. Identification and characterization of an alternative cancer-derived PD-L1 splice variant. Cancer Immunol Immunother. 2019;68(3):407-420.

247. Broutian TR, Jiang B, Li J, et al. Human papillomavirus insertions identify the PIM family of serine/threonine kinases as targetable driver genes in head and neck squamous cell carcinoma. Cancer Lett. 2020;476:23-33.

248. Beier UH, Weise JB, Laudien M, Sauerwein H, Görögh T. Overexpression of Pim-1 in head and neck squamous cell carcinomas. Int J Oncol. 2007;30(6):1381-1387.

249. Chiang WF, Yen CY, Lin CN, et al. Up-regulation of a serine-threonine kinase proto-oncogene Pim-1 in oral squamous cell carcinoma. Int J Oral Maxillofac Surg. 2006;35(8):740-745.

250. Huebbers CU, Preuss SF, Kolligs J, 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):e57207.

251. Penning TM. Aldo-Keto Reductase Regulation by the Nrf2 System: Implications for Stress Response, Chemotherapy Drug Resistance, and Carcinogenesis. Chem Res Toxicol. 2017;30(1):162-176.

252. Walline HM, Goudsmit CM, McHugh JB, 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):840-852.

253. Khanal S, Shumway BS, Zahin M, 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):30419-30433.

254. Hatano T, Sano D, Takahashi H, et al. Identification of human papillomavirus (HPV) 16 DNA integration and the ensuing patterns of methylation in HPV-associated head and neck squamous cell carcinoma cell lines. Int J Cancer. 2017;140(7):1571-1580.

255. Karimzadeh M, Arlidge C, Rostami A, Lupien M, Bratman SV, Hoffman MM. Human papillomavirus integration transforms chromatin to drive oncogenesis. Genome Biol. 2023;24(1):142.

256. Lajer CB, Garnæs E, Friis-Hansen L, 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):1526-1534.

257. Hui ABY, Lin A, Xu W, et al. Potentially prognostic miRNAs in HPV-associated oropharyngeal carcinoma. Clin Cancer Res. 2013;19(8):2154-2162.

258. Wald AI, Hoskins EE, Wells SI, Ferris RL, Khan SA. Alteration of microRNA profiles in squamous cell carcinoma of the head and neck cell lines by human papillomavirus. Head Neck. 2011;33(4):504-512.

259. Lace MJ, Anson JR, Haugen TH, Dierdorff JM, Turek LP. Interferon treatment of human keratinocytes harboring extrachromosomal, persistent HPV-16 plasmid genomes induces de novo viral integration. Carcinogenesis. 2015;36(1):151-159.

260. Visalli G, Riso R, Facciolà A, et al. Higher levels of oxidative DNA damage in cervical cells are correlated with the grade of dysplasia and HPV infection. J Med Virol. 2016;88(2):336-344.

261. Wei L, Gravitt PE, Song H, Maldonado AM, Ozbun MA. Nitric Oxide Induces Early Viral Transcription Coincident with Increased DNA Damage and Mutation Rates in Human Papillomavirus–Infected Cells. Cancer Res. 2009;69(11):4878-4884.

262. Hawkins TB, Dantzer J, Peters B, et al. Identifying viral integration sites using SeqMap 2.0. Bioinformatics. 2011;27(5):720-722.

263. Chen Y, Yao H, Thompson EJ, Tannir NM, Weinstein JN, Su X. VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. Bioinformatics. 2013;29(2):266-267.

264. Wang Q, Jia P, Zhao Z. VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. PLoS One. 2013;8(5):e64465.

265. Wang Q, Jia P, Zhao Z. VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. Genome Med. 2015;7(1):2.

266. Li JW, Wan R, Yu CS, Co NN, Wong N, Chan TF. ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution. Bioinformatics. 2013;29(5):649-651.

267. Ho DWH, Sze KMF, Ng IOL. Virus-Clip: a fast and memory-efficient viral integration site detection tool at single-base resolution with annotation capability. Oncotarget. 2015;6(25):20959-20963.

268. Forster M, Szymczak S, Ellinghaus D, et al. Vy-PER: eliminating false positive detection of virus integration events in next generation sequencing data. Sci Rep. 2015;5:11534.

269. Tennakoon C, Sung WK. BATVI: Fast, sensitive and accurate detection of virus integrations. BMC Bioinformatics. 2017;18(Suppl 3):71.

270. Khan A, Liu Q, Chen X, et al. Detection of human papillomavirus in cases of head and neck squamous cell carcinoma by RNA-seq and VirTect. Mol Oncol. 2019;13(4):829-839.

271. Zeng X, Zhao L, Shen C, Zhou Y, Li G, Sung WK. HIVID2: an accurate tool to detect virus integrations in the host genome. Bioinformatics. 2021;37(13):1821-1827.

272. Dyer N, Young L, Ott S. Artifacts in the data of Hu et al. Nat Genet. 2016;48(1):2-4.

273. Nguyen NPD, Deshpande V, Luebeck J, Mischel PS, Bafna V. ViFi: accurate detection of viral integration and mRNA fusion reveals indiscriminate and unregulated transcription in proximal genomic regions in cervical cancer. Nucleic Acids Res. 2018;46(7):3309-3325.

274. Rajaby R, Zhou Y, Meng Y, et al. SurVirus: a repeat-aware virus integration caller. Nucleic Acids Res. 2021;49(6):e33.

275. Cameron DL, Jacobs N, Roepman P, Priestley P, Cuppen E, Papenfuss AT. VIRUSBreakend: Viral Integration Recognition Using Single Breakends. Bioinformatics. 2021;37(19):3115-3119.

276. Kukurba KR, Montgomery SB. RNA Sequencing and Analysis. Cold Spring Harb Protoc. 2015;2015(11):951-969.

277. Allergy Methods and Protocols. Humana Press

278. Abreu ALP, Souza RP, Gimenes F, Consolaro MEL. A review of methods for detect human Papillomavirus infection. Virol J. 2012;9:262.

279. Morgan IM, DiNardo LJ, Windle B. Integration of Human Papillomavirus Genomes in Head and Neck Cancer: Is It Time to Consider a Paradigm Shift? Viruses. 2017;9(8). doi:10.3390/v9080208

280. Luft F, Klaes R, Nees M, 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):9-17.

281. Ziegert C, Wentzensen N, Vinokurova S, et al. A comprehensive analysis of HPV integration loci in anogenital lesions combining transcript and genome-based amplification techniques. Oncogene. 2003;22(25):3977-3984.

282. Gradíssimo A, Burk RD. Molecular tests potentially improving HPV screening and genotyping for cervical cancer prevention. Expert Rev Mol Diagn. 2017;17(4):379-391.

283. Petersen BS, Fredrich B, Hoeppner MP, Ellinghaus D, Franke A. Opportunities and challenges of whole-genome and -exome sequencing. BMC Genet. 2017;18(1):14.

284. Harlé A, Guillet J, Thomas J, et al. HPV insertional pattern as a personalized tumor marker for the optimized tumor diagnosis and follow-up of patients with HPV-associated carcinomas: a case report. BMC Cancer. 2019;19(1):277.

285. de Vree PJP, de Wit E, Yilmaz M, et al. Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping. Nat Biotechnol. 2014;32(10):1019-1025.

286. Van Doorslaer K, Li Z, Xirasagar S, et al. The Papillomavirus Episteme: a major update to the papillomavirus sequence database. Nucleic Acids Res. 2017;45(D1):D499-D506.

287. Peng Z, Zhang Y, Gu W, et al. Integration of the hepatitis B virus X fragment in hepatocellular carcinoma and its effects on the expression of multiple molecules: a key to the cell cycle and apoptosis. Int J Oncol. 2005;26(2):467-473.

288. Bouchard MJ, Schneider RJ. The enigmatic X gene of hepatitis B virus. J Virol. 2004;78(23):12725-12734.

289. Arbuthnot P, Capovilla A, Kew M. Putative role of hepatitis B virus X protein in hepatocarcinogenesis: effects on apoptosis, DNA repair, mitogen-activated protein kinase and JAK/STAT pathways. J Gastroenterol Hepatol. 2000;15(4):357-368.

290. Pang R, Tse E, Poon RTP. Molecular pathways in hepatocellular carcinoma. Cancer Lett. 2006;240(2):157-169.

291. Huang J, Kwong J, Sun EC, Liang TJ. Proteasome complex as a potential cellular target of hepatitis B virus X protein. J Virol. 1996;70(8):5582-5591.

292. Hu Z, Zhang Z, Doo E, Coux O, Goldberg AL, Liang TJ. Hepatitis B virus X protein is both a substrate and a potential inhibitor of the proteasome complex. J Virol. 1999;73(9):7231-7240.

293. Zhang Z, Torii N, Furusaka A, Malayaman N, Hu Z, Liang TJ. Structural and functional characterization of interaction between hepatitis B virus X protein and the proteasome complex. J Biol Chem. 2000;275(20):15157-15165.

294. Rahmani Z, Huh KW, Lasher R, Siddiqui A. Hepatitis B virus X protein colocalizes to mitochondria with a human voltage-dependent anion channel, HVDAC3, and alters its transmembrane potential. J Virol. 2000;74(6):2840-2846.

295. Huh KW, Siddiqui A. Characterization of the mitochondrial association of hepatitis B virus X protein, HBx. Mitochondrion. 2002;1(4):349-359.

296. Chami M, Ferrari D, Nicotera P, Paterlini-Bréchot P, Rizzuto R. Caspase-dependent Alterations of Ca2+ Signaling in the Induction of Apoptosis by Hepatitis B Virus X Protein *. J Biol Chem. 2003;278(34):31745-31755.

297. Bouchard MJ, Wang LH, Schneider RJ. Calcium signaling by HBx protein in hepatitis B virus DNA replication. Science. 2001;294(5550):2376-2378.

298. Forgues M, Difilippantonio MJ, Linke SP, et al. Involvement of Crm1 in hepatitis B virus X protein-induced aberrant centriole replication and abnormal mitotic spindles. Mol Cell Biol. 2003;23(15):5282-5292.

Chapter 2 SearcHPV: A Novel Approach to Identify and Assemble Human Papillomavirus–Host Genomic Integration Events in Cancer

This chapter was published in 2021 in Cancer (PMID: 34160069). The author of this dissertation served as the co-first author of this paper. The main text and supplementary figures of this paper was presented below. Other supplementary materials could be referred to the published journal.

2.1 Introduction

Human papillomavirus (HPV) is a well-established driver of malignant transformation in a number of cancers, including head and neck squamous cell carcinomas (HNSCC). Although HPV genomic integration is not a normal event in the lifecycle of HPV, it is frequently reported in HPV+ cancers [1–4] and it may be a contributor to oncogenesis. 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 E2, the transcriptional repressor of the HPV oncoproteins, leading to an increase in genetic instability [6]. HPV integration occurs within/near cellular genes more often than expected by chance [7] and has been reported to be associated with structural variations [8]. 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 low sensitivity assays and are limited in their ability to detect the broad spectrum of genomic changes resulting from this process. Next-generation sequencing (NGS) technologies overcome these limitations. Previous groups have assessed HPV integration within HNSCC tumors in The Cancer Genome Atlas (TCGA) and cell lines by whole-genome sequencing (WGS) [2, 3, 8]. There are a variety of viral integration detection tools developed for WGS data, such as VirusFinder2 [12, 13] and VirusSeq[14]. 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 [5, 15–18]. This strategy allows for better coverage of integration sites than an untargeted approach like WGS but 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 to have a relatively low validation rate and limitations on the structural information surrounding the fusion sites, which impairs mechanistic studies. 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.

2.2 Materials and methods

2.2.1 Targeted capture sequencing

DNA from the HPV16-positive UM-SCC-47 cell line [45], a Patient derived xenograft (PDX)-294R (National Cancer Institute Identifier: PDX-932174–294-R) and a frozen HPV+ sample, TumorA, were submitted to the University of Michigan Advanced Genomics Core for targeted capture sequencing. The patient donating TumorA was consented for next generation sequencing under a previously described protocol approved by the University of Michigan Institutional Review Board [41]. 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, and over 200 HNSCC-related genes, which are detailed in Heft Neal et. al 2020 [19]. 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.

2.2.2 Novel integration caller (SearcHPV)

The pipeline of SearcHPV has four main steps which are detailed below: (1) Alignment; (2) Genome fusion point calling; (3) Assembly; (4) HPV fusion point calling (Figure 2.1). The package is available on Github: https://github.com/mills-lab/SearcHPV.

2.2.2.1 Alignment

The customized reference genome used for alignment was constructed by catenating the HPV16 genome (from Papillomavirus Episteme (PAVE) database [20, 21]) 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 [22]. Then we performed an indel realignment by Picard Tools [23] and GATK [24]. Duplications were marked by Picard MarkDuplicates Tool [23] for the filtering in downstream steps.

2.2.2.2 Genome fusion points calling

To identify the fusion points, we extracted reads with regions matched to HPV16 and filtered those reads to meet these criteria: (1) not secondary alignment; (2) mapping quality greater or equal than 50; (3) not duplicated. Genome fusion points were called by split reads (reads spanning both the human and HPV genomes) and the paired-end reads (reads with one end matched to HPV and the other matched the human genome) at the surrounding region (+/−300bp) (Figure 2.1A). 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 nature of read mapping and structural variations.

2.2.2.3 Assembly

To construct longer sequence contigs from individual reals, we extracted supporting split reads and paired-end reads 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 × read length, resulting in overlapping read segments. For such events, we first merged these reads using PEAR [25] and then combined them with other individual reads to perform a local assembly by CAP3 [26] (Figure 2.1).

2.2.2.4 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 1B) and from right side (Contig B in Fig 1B). For each class, if there were high confidence contigs in the class, we selected the contig with maximum length among them, otherwise 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 2.1C). 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 [22] and the analysis was performed with R 3.6.1 [27] and Python [28].

2.3 RESULTS

2.3.1 SearcHPV pipeline

To overcome the limitations of viral integration detection in WGS of detecting rare events, we performed HPV targeted capture sequencing which allows for deeper investigation of these events. Current bioinformatics pipelines available are not designed for this type of data so we developed a novel HPV integration detection tool for targeted capture sequencing data, which we termed "SearcHPV". Two HPV16+ HNSCC models, UM-SCC-47 and Patient derived xenograft (PDX)-294R as well as an HPV16+ HNSCC tumor, TumorA, were subjected to targeted-capture based Illumina sequencing using a custom panel of probes spanning the entire HPV16 genome. The paired end reads then went through the four steps of analysis of SearcHPV: alignment to custom reference genome, genome fusion points calling, local assembly and HPV fusion point calling (Figure 2.1). Analysis of the integration sites in the models using our pipeline SearcHPV showed a high frequency of HPV16 integration with a total of six events in UM-SCC-47, ninety-eight in PDX-294R and eight in TumorA (Figure 2.2, Figure 2.8, Table S1–S3).

2.3.2 Comparison to other integration callers and confirmation of integration sites

In addition to using SearcHPV, we used two previously developed integration callers, VirusFinder2 and VirusSeq to independently call integration events in UM-SCC-47, PDX-294R and TumorA (Figure 2.3, Tables S4–S6). We found that SearcHPV called HPV integration events at a much higher rate than either previous caller (Figure 2.3B). There were a large number of sites that were only identified by SearcHPV (n=82). In order to assess the accuracy of each caller, we performed PCR for PDX-294R and UM-SCC-47 on source genomic DNA followed by Sanger sequencing with primers spanning the HPV-human junction sites predicted by the callers. We tested all integration sites with sufficient sequence complexity for primer design (n=43), twenty-five of which were unique to SearcHPV and five of which were unique to VirusSeq. VirusFinder2 does not allow for local assembly of the integration junctions which rendered us unable to test these sites. UM-SCC-47 was also subjected to Oxford Nanopore GridION sequencing to provide additional supportive evidence of integration sites. We combined the information from PCR and Nanopore sequencing to interrogate a total of 44 integration sites and compared the conformation rates for each caller. (Figure 2.3C. S1, Table S7, S17). Sites unique to SearcHPV had a confirmation rate of 19/26 (73%). The confirmation rate of high confidence SearcHPV sites was higher than that for low confidence sites (25/32 (78%) versus 4/7 (57%)). In contrast, only 1/5 (20%) sites unique to VirusSeq could be confirmed.

To further compare the performance of SearcHPV and the other two callers, we expanded the sequencing requirements by applying them on whole exome sequencing data (WES) for UM-SCC-47 and PDX-294R, which were either previously generated by our lab [41,42] or were publicly available, respectively. VirusSeq did not report any integration results in either sample from the WES data. For UM-SCC-47, SearcHPV and VirusFinder2 both called one integration site. This

site was reported by SearcHPV from targeted capture data. For PDX-294R, SearcHPV identified three integration sites while VirusFinder2 did not identify any sites. Two of three integration sites were also called by SearcHPV from targeted capture data and the other one was not covered in the targeted region of our targeted capture technology (Table S10–13). By examining the location of integration sites called from targeted capture sequencing for these two samples, we found that most (102/104) fell outside of the targeted region of WES, resulting in lower coverage of reads and insufficient evidence to identify the integration events (Table S14–16). Given this limitation of WES on capturing genome-wide HPV integration events, our approach was still more applicable on identifying HPV integration events than VirusSeq and VirusFinder2.

2.3.2.1 Localization of integration sites

We next examined the integration sites detected by SearcHPV. The six integration sites discovered in UM-SCC-47 were clustered on chromosome 3q28 within/near the cellular gene *TP63* and either had breakpoints within the HPV16 genes E1, E2 or L1. The integration sites fell within intron 10, intron 12 and exon 14. One additional integration site was 8.6 kb downstream of the *TP63* coding region.

For TumorA, six of eight integration sites were clustered on chromosome 9q34 within/near gene *TRAF2,* including one integration site that fell within *FBXW5* which was 15.8kb downstream of *TRAF2.* Among them, three integration sites fell within intron 5 of *TRAF2* and one mapped to intron 8.

Within PDX-294R, HPV16 integration sites were identified across 21 different chromosomes, occurring most frequently on chromosome 3. For the 98 integration events of PDX-294R, we identified 142 breakpoints in the HPV genome. The most frequently involved HPV genes were E1 (45/142 (32%)) and L1 (31/142 (22%)). Most of the integration sites mapped to within/near (<50 kb) a known cellular gene (89/98 (91%)). Of the sites that fell within a gene, the majority of integrations took place within an intronic region (3¾2 (78%)). Although the integration sites were scattered throughout the human genome, we saw examples of closely clustered sites around cancer-relevant genes, including *ZNF148* and *SNX4* on chromosome 3q21.2, *MYC* on chromosome 8q24.21 and *FOXN2* on chromosome 2p16.3.

2.3.2.2 Association of integration sites and large-scale duplications

We predicted that the complex integration sites we discovered in UM-SCC-47, PDX-294R and TumorA 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, PDX-294R and TumorA to 10X linked-read sequencing. We generated over 1 billion reads for each sample (Table S8), with phase blocks (contiguous blocks of DNA from the same allele) of up to 28.9M, 3.8M and 15.3M bases in length for UM-SCC-47, PDX-294R and TumorA, respectively (Figure 2.7). This led to the identification of 444 high confidence large structural events in UM-SCC-47, 126 events in the PDX-294R model and 49 events in TumorA. We then performed integrated analysis with our SearcHPV results. There was a 130 kb duplication surrounding the integration events in *TP63* in UM-SCC-47 (Figure 2.4A). In PDX-294R, 38/98 (39%) integration sites were within a region that contained a large-scale duplication, while the

other 50 integration events fell outside regions of large structural variation. This suggested that in this PDX model, 38/126 (30%) 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 also associated with large genomic duplications (Figure 2.4B–C). For TumorA, large duplications were not observed within the surrounding region of the eight integration events (Figure 2.4E).

To further resolve the structure around the clusters of integration sites, we performed local assembly for UM-SCC-47 using Nanopore sequencing data (See Supplementary File, Figure 2.4F). The 60K-bp scaffold indicated a 15K-bp, twice amplified segment that matched against the human genome and a 7.5K-bp, twice amplified segment matched against HPV genome. These segments were potentially amplified from a large 22.5K-bp focal genomic segment that has both human and HPV genomic components (Figure 2.4F, copy1–3) and then parts of one duplication were deleted resulting in the shorter segment in the middle (Figure 2.4F, copy2). These human segments and HPV segments were all bounded by identical or very near breakpoints. The integration sites on the human genome shown by the local assembly kept consistent with results from SearcHPV. Notably, within the focal HPV segments, an HPV-HPV junction structure was also identified showing an HPV internal rearrangement structure (Figure 2.4F, pink and yellow parts). This HPV internal rearrangement occurred twice and resulted in additional breakpoints on the HPV genome. The focal amplification structure resolved by local assembly from Nanopore sequencing confirmed the duplications predicted by 10X linked-read sequencing and indicated the association of HPV integrations and large-scaled duplications.

2.3.2.3 Microhomology at junction sites

Finally, to evaluate possible mechanisms of DNA repair-mediated integration, we examined the degree of sequence overlap between the genomes at junction sites that were covered by contigs. 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 genomes (Figure 2.5A). The majority (59%) of junction sites in the three samples had at least some degree of microhomology (Figure 2.5B–D). 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. There was also a large number of junctions with gaps between the human and HPV genomes ranging from 1 – 54 bp long.

2.4 Discussion

We developed a novel bioinformatics pipeline that we termed "SearcHPV" and show that it operated in a more accurate and efficient manner than existing pipelines on targeted capture sequencing data. 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 and one frozen HNSCC HPV+ sample, 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 Oxford Nanopore sequencing 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, PDX-294R and TumorA. Importantly, UM-SCC-47 has been previously assessed for HPV integration by a variety of methods [8, 29–32], which we leveraged as ground truth knowledge to validate our integration caller. All previous studies were in agreement that HPV16 is integrated 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 [32] and/or WGS [8] with the exception of E1 into intron 10, which was unique to our caller and confirmed by direct PCR. 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 [31]), 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 TumorA and PDX-294R have not been previously analyzed for viral-host integration sites and therefore represented a true discovery case. For TumorA, we identified a cluster of HPV integration sites within/near *TRAF2*. Interestingly, *TRAF2* was previously identified as a potential downstream

effector of E6/E7 [43,44], and due to the role of *TRAF2* in regulating innate immunity, this gene may have a larger role in HPV16-mediated biology than previously recognized.

For PDX-294R, we identified widespread HPV integration sites throughout the host genome and also observed that 66% of integration sites were found within or near genes. This aligns with previous reports that integrations are detected in host genes more frequently than expected by chance [2, 3, 7, 33]. One particularly interesting cluster of integration events surrounded the cellular proto-oncogene *MYC*. Importantly, *MYC* has been identified as a potential hotspot for HPV integration[7, 34] and the junctions we detected in/near this gene had 2–4 bp of microhomology, 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, 35, 36]. There is a high degree of microhomology between HPV16 and this gene. 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[37].

HPV integration sites have been associated with structural variations in the human genome[3, 8, 37], which supports an additional genetic mechanism as to why HPV integration sites may often be detected adjacent to host cancer-related genes. These structural variation events are 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, 38]. 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. It is unclear why only some integration sites were associated with structural variants, but it is possible that an alternative mechanism of integration occurred [38]. Notably, we resolved and identified an HPV-HPV junction that bounded in a large duplication segment and showed the possibility of HPV internal rearrangement to be involved in HPV integration events.

Importantly, this observation that HPV integration events tended to be enriched in cellular genes could result from 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, implicating 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. The majority of breakpoints had some level of microhomology. 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 [39]. 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 these sites [40]. However, given the relatively small number of

events we examined, we expect that future analysis with our pipeline will help resolve the specific role of each DNA repair pathway in HPV-human fusion breakpoints.

Overall, our new HPV detection pipeline SearchHPV overcomes a gap in the field of viral-host integration analysis. While the performance of SearcHPV has only been examined on three samples, in the future, we expect that the application of this pipeline in large HPV+ cancer tissue cohorts will help advance our understanding of the potential oncogenic mechanisms associated with viral integration. With the emerging set of tools such as SearcHPV, 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.

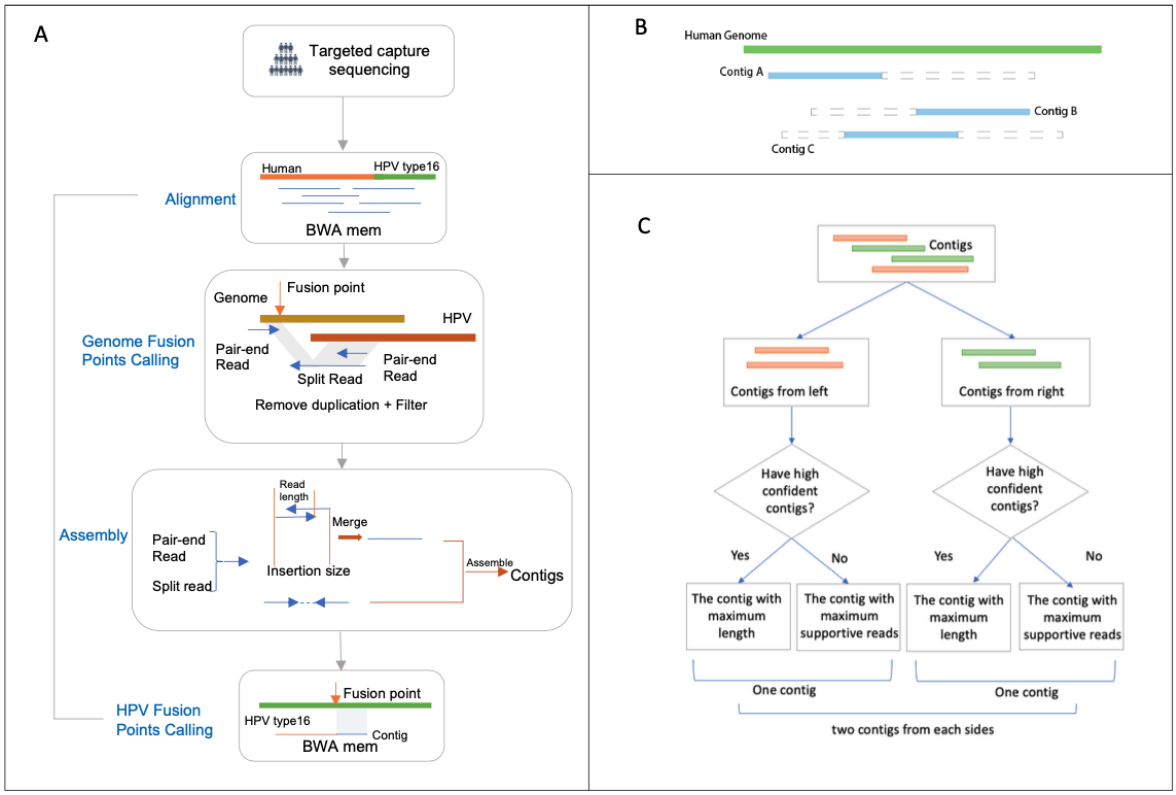


Figure 2.1 Workflow of SearcHPV. A. Paired-end reads from targeted capture sequencing were aligned to a catenated Human-HPV reference genome. After removing duplication and filter, fusion points were identified by split reads and pair-end reads. Informative reads were extracted for local assembly. Reads pairs that have overlaps were merged first before assembly. Assembled contigs were aligned to the HPV genome to identify the breakpoints on HPV. B. Contigs were divided into two classes. Blue solid triangle demonstrates the matched region of the contig. Grey dashed triangle demonstrates the clipped region of the contig. Contig A would be assigned to the left group and Contig B would be assigned to the right group. Contig C would be randomly assigned to the left or right group. C. Workflow for the contig selection procedures for fusion point with multiple candidates contigs. For each fusion point, we report at least one contig and at most two contigs representing two directions.

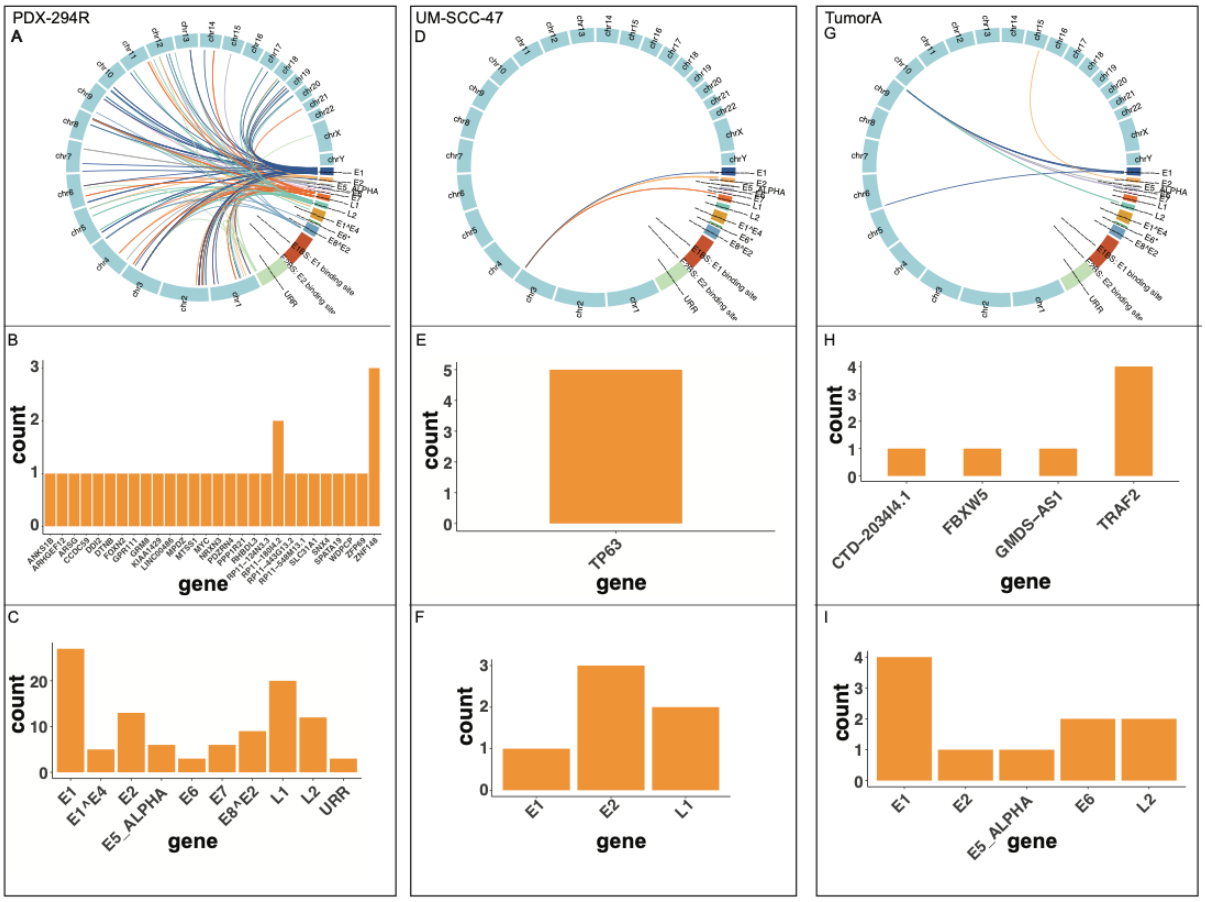


Figure 2.2 Distribution of breakpoints in the human and HPV genomes called by SearcHPV. A-C. Results for PDX-294R. A. Links of breakpoints in the human and HPV16 genomes for PDX-294R. B. Quantification of breakpoint calls in human genes for PDX-294R. C. Quantification of breakpoints calls in the HPV16 genes for PDX-294R. D-F. As described in A-C for UM-SCC-47. G-I. As described in A-C for 4840 TumorA

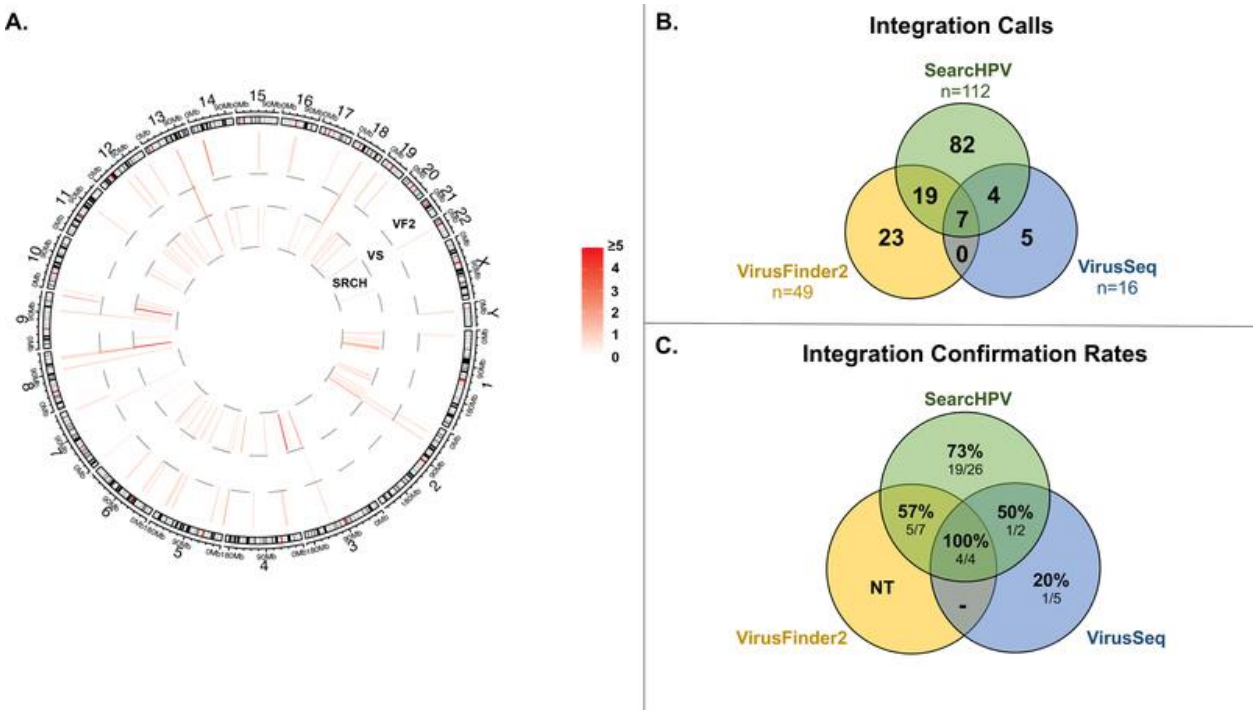


Figure 2.3 Comparison of integration sites called by SearcHPV, VirusSeq and VirusFinder2 in three samples. A. Each bar denotes integration sites within the region. The colormap shows the count of the integration sites. B. Number of integration sites called by each program. Integration sites from VirusSeq and VirusFinder2 were clustered within 100bp to keep consistent with SearcHPV. C. PCR and Nanopore confirmation rate for a subset of B that were chosen to assess accuracy using both PCR and Nanopore sequencing where available. If there is at least one split read from Nanopore sequencing data supporting an integration site, the integration site was regarded as validated by Nanopore sequencing. An integration site was counted as confirmed if it was validated by PCR or Nanopore sequencing.

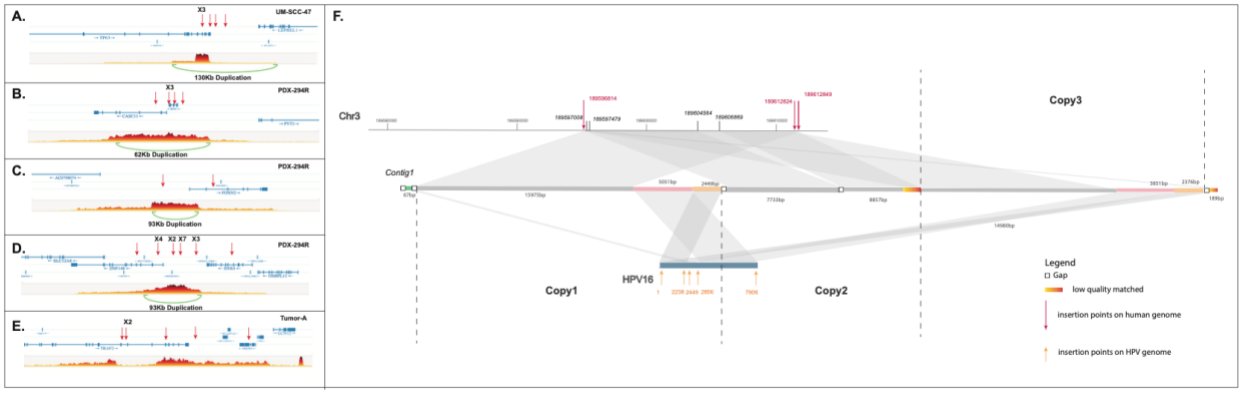


Figure 2.4 Genomic duplications associated with HPV integration. A. UM-SCC-47. B-D. PDX-294R. E. TumorA. Red arrows indicate integration sites. Each plot shows the number of overlapping barcodes observed in sequencing reads of that region. F. Local assembly around the HPV integration sites in UM-SCC-47 using Nanopore sequencing data. The scaffold mapped to different regions was marked by different colors. Gray: match to human genome reference. Green, pink and yellow: match to HPV genome. Potential duplications were marked by the same color.

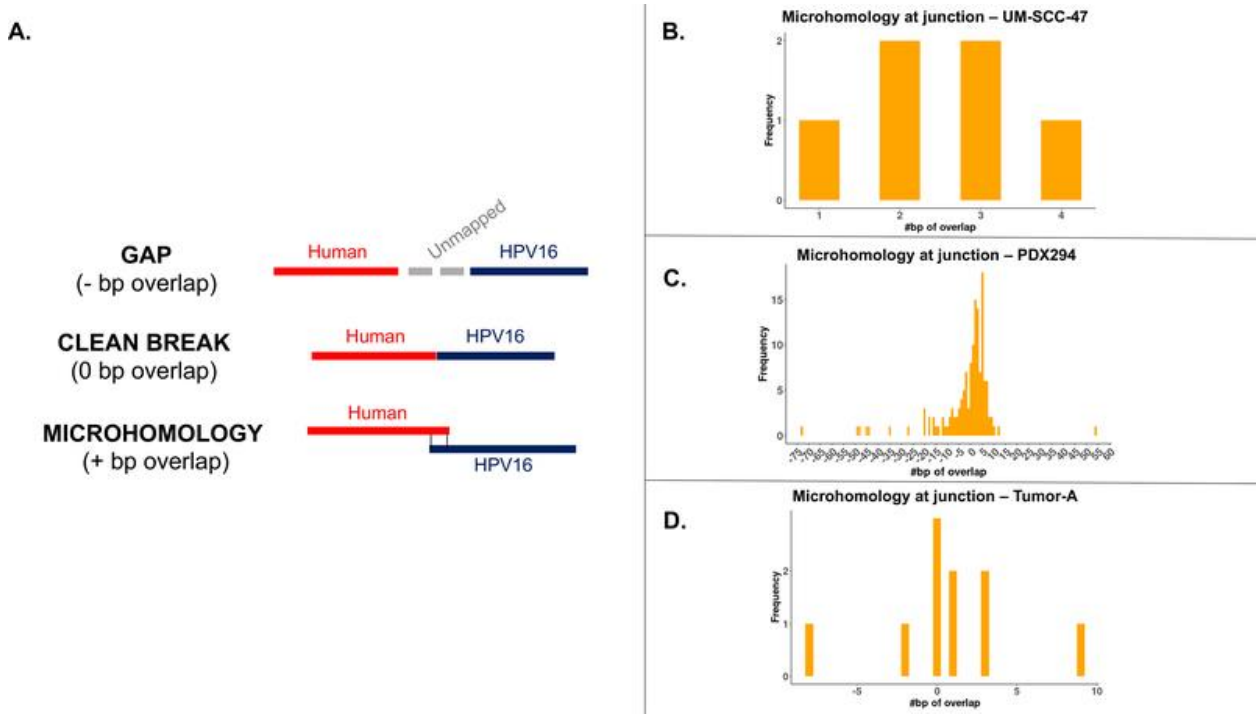


Figure 2.5 Microhomology at junction points. A. The three types of junction points. B. Level of microhomology (in bp) in UM-SCC-47. C. Level of microhomology (in bp) in PDX-294R. D. Level of microhomology (in bp) in TumorA. Junctions with a gap are shown as negative numbers.

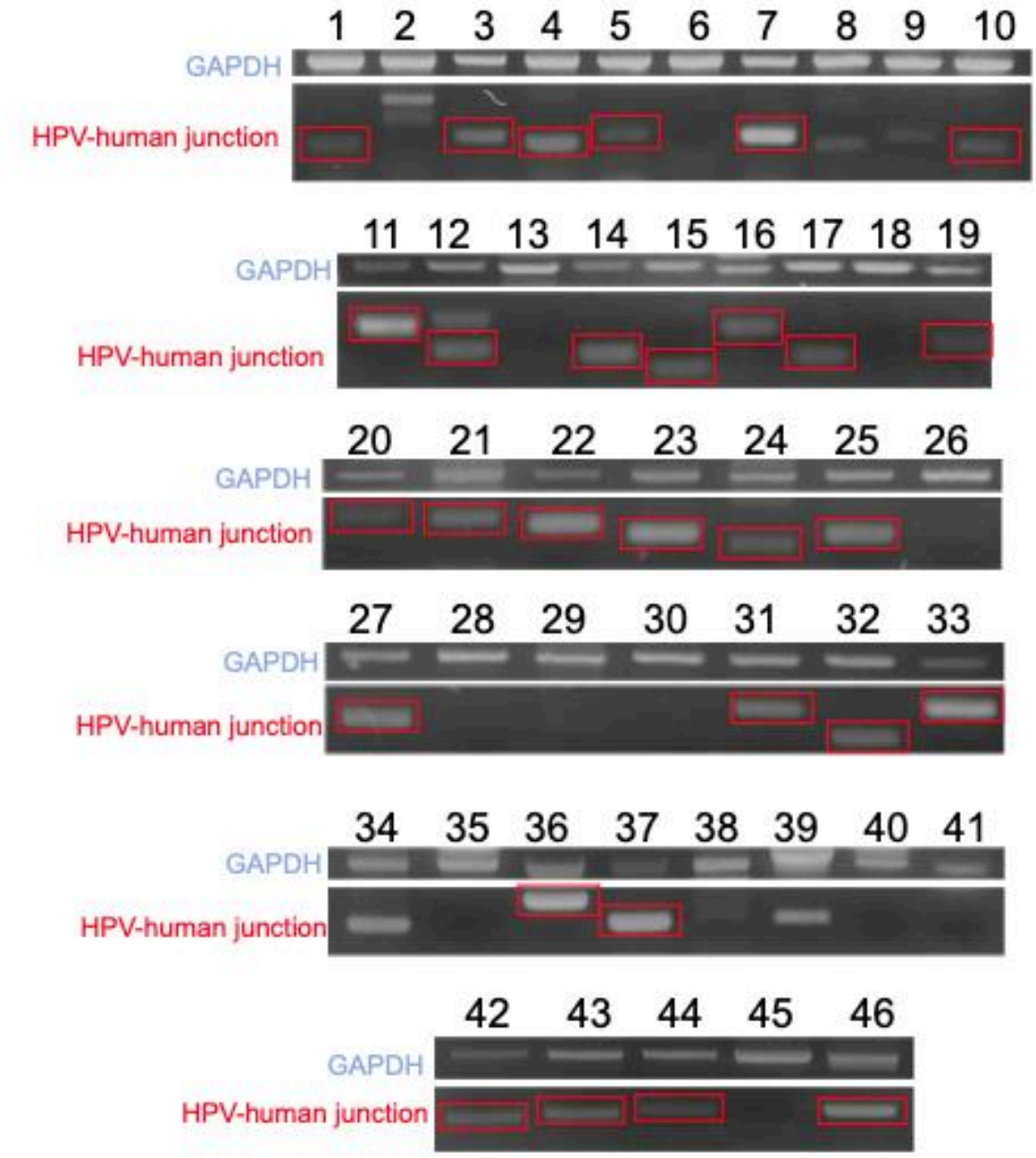


Figure 2.6 PCR validation gel electrophoresis. Top band of each row shows GAPDH (535 bp), bottom bands represent predicted HPV-human junctions (ranging from 70-250 bp). Red boxes demonstrate bands that appeared at the correct molecular weight and were validated by Sanger sequencing.

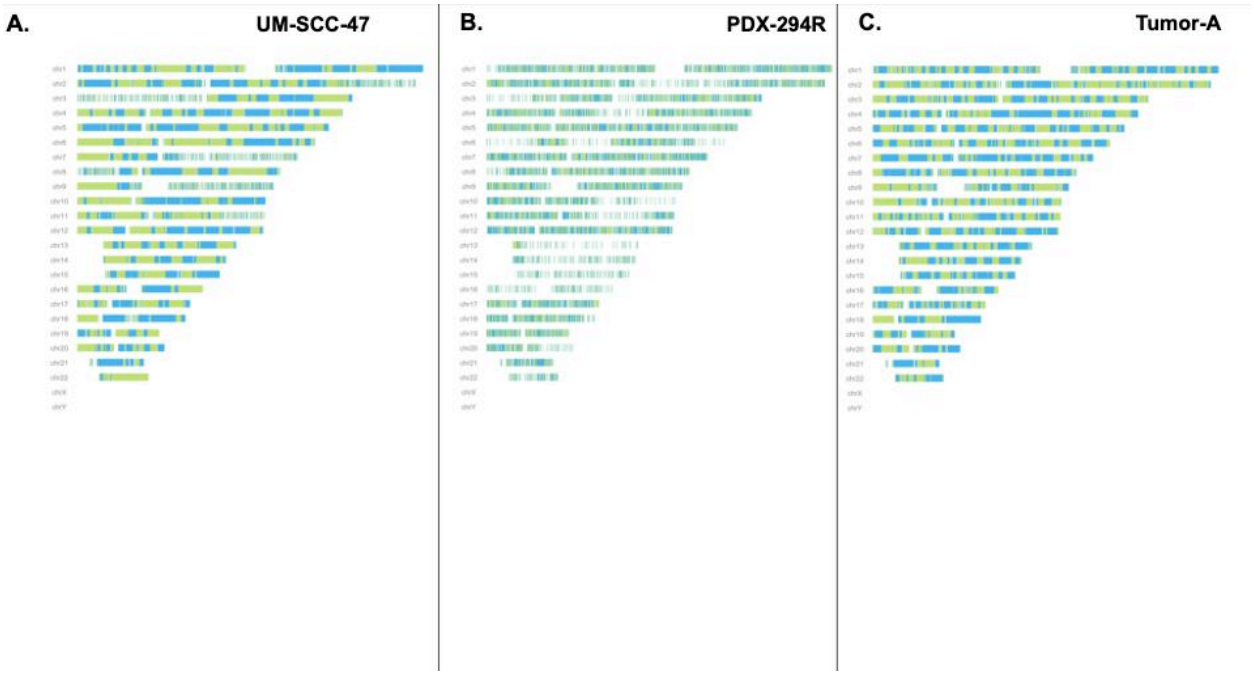


Figure 2.7 Linked read SNP phase plots for UM-SCC-47 (A) PDX-294R (B) and TumorA (C) genomes. Alternating colors represent different phase blocks, which are contiguous blocks of DNA from the same allele based on differential SNP phasing performed by LongRanger software.

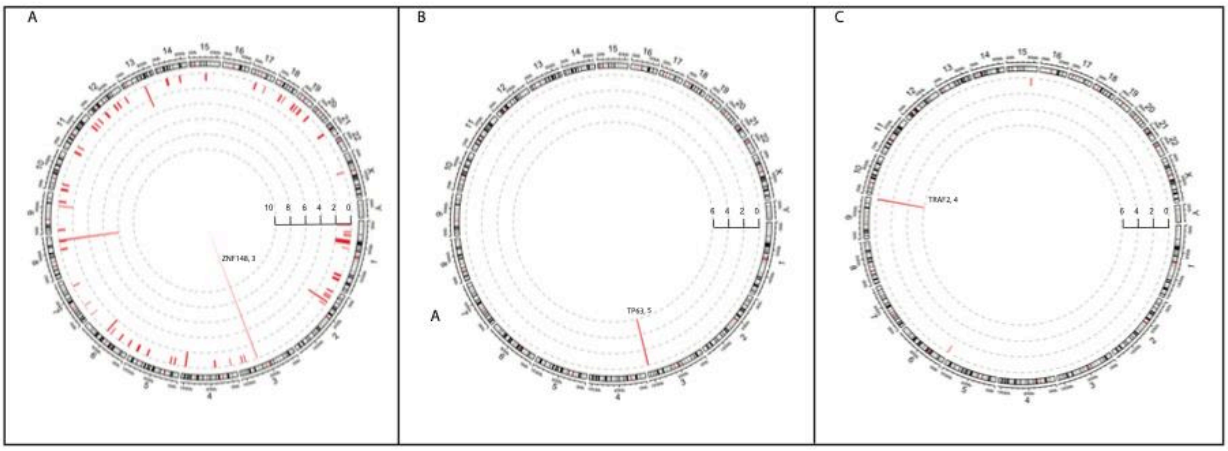


Figure 2.8 Distribution of integration sites in the human genome for PDX-204R (A), UM-SCC-47 (B) and TumorA (C). Each red bar denotes the integration sites within the region. Outliers were marked with genes that fell in and the corresponding count of integration sites.

Bibliography

1. Gao G, Wang J, Kasperbauer JL, et al. Whole genome sequencing reveals complexity in both HPV sequences present and HPV integrations in HPV-positive oropharyngeal squamous cell carcinomas. BMC Cancer. Apr 11 2019;19(1):352. doi:10.1186/s12885-019-5536-1

2. Nulton TJ, Olex AL, Dozmorov M, Morgan IM, Windle B. Analysis of The Cancer Genome Atlas sequencing data reveals novel properties of the human papillomavirus 16 genome in head and neck squamous cell carcinoma. Oncotarget. Mar 14 2017;8(11):17684-17699. doi:10.18632/oncotarget.15179

3. Parfenov M, Pedamallu CS, Gehlenborg N, et al. Characterization of HPV and host genome interactions in primary head and neck cancers. Proc Natl Acad Sci U S A. Oct 28 2014;111(43):15544-9. doi:10.1073/pnas.1416074111

4. Pinatti LM, Sinha HN, Brummel CV, et al. Association of human papillomavirus integration with better patient outcomes in oropharyngeal squamous cell carcinoma. Head Neck. Oct 19 2020;doi:10.1002/hed.26501

5. Tian R, Cui Z, He D, et al. Risk stratification of cervical lesions using capture sequencing and machine learning method based on HPV and human integrated genomic profiles. Carcinogenesis. Oct 16 2019;40(10):1220-1228. doi:10.1093/carcin/bgz094

6. McBride AA, Warburton A. The role of integration in oncogenic progression of HPV-associated cancers. PLoS Pathog. Apr 2017;13(4):e1006211. doi:10.1371/journal.ppat.1006211

7. Bodelon C, Untereiner ME, Machiela MJ, Vinokurova S, Wentzensen N. Genomic characterization of viral integration sites in HPV-related cancers. Int J Cancer. Nov 01 2016;139(9):2001-11. doi:10.1002/ijc.30243

8. Akagi K, Li J, Broutian TR, et al. Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. Genome Res. Feb 2014;24(2):185-99. doi:10.1101/gr.164806.113

9. Pinatti LM, Walline HM, Carey TE. Human Papillomavirus Genome Integration and Head and Neck Cancer. J Dent Res. Jun 2018;97(6):691-700. doi:10.1177/0022034517744213

10. Luft F, Klaes R, Nees M, et al. Detection of integrated papillomavirus sequences by ligation-mediated PCR (DIPS-PCR) and molecular characterization in cervical cancer cells. Int J Cancer. Apr 01 2001;92(1):9-17.

11. Klaes R, Woerner SM, Ridder R, et al. Detection of high-risk cervical intraepithelial neoplasia and cervical cancer by amplification of transcripts derived from integrated papillomavirus oncogenes. Cancer Res. Dec 15 1999;59(24):6132-6.

12. Wang Q, Jia P, Zhao Z. VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. Plos One. 2013;8(5):e64465. doi:10.1371/journal.pone.0064465

13. Wang Q, Jia P, Zhao Z. VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. Genome Med. 2015;7(1):2. doi:10.1186/s13073-015-0126-6

14. Chen Y, Yao H, Thompson EJ, Tannir NM, Weinstein JN, Su X. VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. Bioinformatics. Jan 15 2013;29(2):266-7. doi:10.1093/bioinformatics/bts665

15. Holmes A, Lameiras S, Jeannot E, et al. Mechanistic signatures of HPV insertions in cervical carcinomas. NPJ Genom Med. 2016;1:16004. doi:10.1038/npjgenmed.2016.4

16. Montgomery ND, Parker JS, Eberhard DA, et al. Identification of Human Papillomavirus Infection in Cancer Tissue by Targeted Next-generation Sequencing. Appl Immunohistochem Mol Morphol. Aug 2016;24(7):490-5. doi:10.1097/PAI.0000000000000215

17. Morel A, Neuzillet C, Wack M, et al. Mechanistic Signatures of Human Papillomavirus Insertions in Anal Squamous Cell Carcinomas. Cancers (Basel). Nov 22 2019;11(12)doi:10.3390/cancers11121846

18. Nkili-Meyong AA, Moussavou-Boundzanga P, Labouba I, 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. Feb 6 2019;9(1):1504. doi:10.1038/s41598-018-37871-2

19. Heft Neal ME, Bhangale AD, Birkeland AC, et al. Prognostic Significance of Oxidation Pathway Mutations in Recurrent Laryngeal Squamous Cell Carcinoma. Cancers (Basel). Oct 22 2020;12(11)doi:10.3390/cancers12113081

20. NIAID. Papillomavirus Episteme. Bioinformatics and Computational Biosciences Branch. 2020. https://pave.niaid.nih.gov/

21. Van Doorslaer K, Li Z, Xirasagar S, et al. The Papillomavirus Episteme: a major update to the papillomavirus sequence database. Nucleic Acids Res. Jan 4 2017;45(D1):D499-D506. doi:10.1093/nar/gkw879

22. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. Jul 15 2009;25(14):1754-60. doi:10.1093/bioinformatics/btp324

23. Institute B. Picard toolkit. Broad Institute GitHub Repository. 2019;

24. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. Sep 2010;20(9):1297-303. doi:10.1101/gr.107524.110

25. Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics. Mar 1 2014;30(5):614-20. doi:10.1093/bioinformatics/btt593

26. Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome Res. Sep 1999;9(9):868-77. doi:10.1101/gr.9.9.868

27. Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing. 2019;

28. Van Rossum G, Drake F.L. Python 3 Reference Manual: Python Documentation Manual Part 2. CreateSpace Independent Publishing Platform. 2009;

29. Khanal S, Shumway BS, Zahin M, 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. Jul 13 2018;9(54):30419-30433. doi:10.18632/oncotarget.25754

30. Myers JE, Guidry JT, Scott ML, et al. Detecting episomal or integrated human papillomavirus 16 DNA using an exonuclease V-qPCR-based assay. Virology. Nov 2019;537:149-156. doi:10.1016/j.virol.2019.08.021

31. Olthof NC, Huebbers CU, Kolligs J, et al. Viral load, gene expression and mapping of viral integration sites in HPV16-associated HNSCC cell lines. Int J Cancer. Mar 1 2015;136(5):E207-18. doi:10.1002/ijc.29112

32. Walline HM, Goudsmit CM, McHugh JB, et al. Integration of high-risk human papillomavirus into cellular cancer-related genes in head and neck cancer cell lines. Head Neck. May 2017;39(5):840-852. doi:10.1002/hed.24729

33. Hu Z, Zhu D, Wang W, 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. Feb 2015;47(2):158-63. doi:10.1038/ng.3178

34. Ferber MJ, Thorland EC, Brink AA, et al. Preferential integration of human papillomavirus type 18 near the c-myc locus in cervical carcinoma. Oncogene. Oct 16 2003;22(46):7233-42. doi:10.1038/sj.onc.1207006

35. Schmitz M, Driesch C, Jansen L, Runnebaum IB, Durst M. Non-random integration of the HPV genome in cervical cancer. Plos One. 2012;7(6):e39632. doi:10.1371/journal.pone.0039632 PONE-D-12-09523 [pii]

36. Walline HM, Komarck CM, McHugh JB, et al. Genomic Integration of High-Risk HPV Alters Gene Expression in Oropharyngeal Squamous Cell Carcinoma. Mol Cancer Res. Oct 2016;14(10):941-952. doi:10.1158/1541-7786.MCR-16-0105

37. Cancer Genome Atlas N. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. Jan 29 2015;517(7536):576-82. doi:10.1038/nature14129

38. Groves IJ, Coleman N. Human papillomavirus genome integration in squamous carcinogenesis: what have next-generation sequencing studies taught us? The Journal of pathology. May 2018;245(1):9-18. doi:10.1002/path.5058

39. Pannunzio NR, Li S, Watanabe G, Lieber MR. Non-homologous end joining often uses microhomology: implications for alternative end joining. DNA Repair (Amst). May 2014;17:74-80. doi:10.1016/j.dnarep.2014.02.006

40. Carvajal-Garcia J, Cho JE, Carvajal-Garcia P, et al. Mechanistic basis for microhomology identification and genome scarring by polymerase theta. Proc Natl Acad Sci U S A. Apr 14 2020;117(15):8476-8485. doi:10.1073/pnas.1921791117

41. Shuman AG, Gornick MC, Brummel C, Kent M, Spector-Bagdady K, Biddle E, et al. Patient and Provider Perspectives Regarding Enrollment in Head and Neck Cancer Research. Otolaryngology--head and neck surgery: official journal of American Academy of Otolaryngology-Head and Neck Surgery 2020;162:73–78.

42. Liu J, Pan S, Hsieh MH, Ng N, Sun F, Wang T, et al. Targeting Wnt-driven cancer through the inhibition of Porcupine by LGK974. Proceedings of the National Academy of Sciences of the United States of America 2013;110:20224–20229.

43. Poirson J, Biquand E, Straub M-L, Cassonnet P, Nominé Y, Jones L, et al. Mapping the interactome of HPV E6 and E7 oncoproteins with the ubiquitin-proteasome system. The FEBS journal 2017;284:3171–3201.

44. Thompson DA, Zacny V, Belinsky GS, Classon M, Jones DL, Schlegel R, et al. The HPV E7 oncoprotein inhibits tumor necrosis factor alpha-mediated apoptosis in normal human fibroblasts. Oncogene 2001;20:3629–3640.

45. Walline HM, Goudsmit CM, McHugh JB, Tang AL, Owen JH, Teh BT, et al. Integration of high-risk human papillomavirus into cellular cancer-related genes in head and neck cancer cell lines. Head & neck 2017;39:840–852.

Chapter 3 Heterogeneity and complexity of human papillomavirus integrations associated with distinct tumorigenic consequences

3.1 Introduction

Persistent infections with human papillomavirus (HPV) constitute a significant risk factor for the development of HPV-associated cancer. Annually in the United States, approximately 46,711 new cancer cases arise frequently associated with HPV according to the record from the Center for Disease Control and Prevention (CDC) [1]. HPV is responsible for nearly all cases of cervical cancers and a substantial number of cancers affecting the vagina, vulva, penis, anus, rectum, and head and neck. Notably, the occurrence of HPV-positive head and neck cancers, especially in the throat (oropharyngeal squamous cell carcinomas (OPSCCs), has significantly risen in the Western world in the last ten years [2]. Currently, up to 90% of OPSCCs are associated with HPV, exceeding the incidence of HPV-positive cervical cancers in the USA [3,4]. HPV possesses two oncogenes, E6 and E7, which interfere with cell cycle regulation. They achieve this by inhibiting the activity of crucial cell cycle regulators, TP53 and RB1. Additionally, HPV harbors the E2 gene, acting as a transcriptional repressor for E6 and E7.

HPV integrations were frequently occurred in cervical cancer and head and neck cancer. However, the exact mechanism and effects of this event was not fully understood. Several studies suggested HPV integration might induce additional oncogenic mechanisms to the

traditional viral oncogenesis E6/E7 pathways [5–8]. HPV integration events varied in complexity, with instances of single-copy segment integration, focal amplification in clustered regions with multiple fusions, and diverse patterns described in different studies [9–12]. HPV integration was categorized into two types in an early review: Type 1, involving the integration of a single genome into cellular DNA, and Type 2, featuring multiple tandem head-to-tail repeats of the HPV genome at a specific genomic locus [13]. Subsequent studies, such as Akagi et al. in 2018, introduced a "looping model" to describe focal duplication, with both HPV and human segments forming DNA concatemers around integration events [14]. Recent long-read sequencing efforts by Akagi et al. in 2023 and another study in 2022 focused on cervical cancer revealed additional complexities, including inter-chromosomal translocations and extrachromosomal circular DNA structures. The latter study classified four types of HPV integration (Type A-D) based on different structural variations and their association with E6 and E7 genes [9]. These investigations have significantly enhanced our comprehension of HPV integration structures, laying a foundation for subsequent explorations into the impacts of HPV integrations. Nevertheless, the classification of HPV integration types predominantly relied on descriptive approaches rather than establishing quantitative thresholds that could universally serve as standards for other research endeavors. This limitation arises from constraints related to sample size and the technology of resolving the complex structural analyses.

In this study, we integrated multi-omics datasets encompassing 291 FFPE HPV-positive OPSCC patients with targeted capture sequencing (TCS) data, 162 Head and neck squamous cell carcinomas (HNSCC) RNA-seq samples (129 of which were paired with TCS), and 16 HNSCC Oxford Nanopore long-read (LR) sequencing samples. We introduced a quantitative method for

defining two distinct types of HPV integrations (Type1 and Type2) based on their biological and statistical characteristics at the DNA level. To accurately portray the complexity of HPV integration, we developed a novel approach to comprehensively resolve local large HPV integration events associated with complex rearrangements at high resolution for LR sequencing data, surpassing the performance of existing LR whole-genome assemblers. Type2 HPV integrations, characterized by high copy numbers and clustered integration events, exhibited more complex profiles at the transcriptomic level.

3.2 Method

3.2.1 Long read based approach to resolve HPV integration events

Nanopore reads were aligned to reference catenated by hg38 and HPV genome using Minimap2[15]. Target chromosomes were identified from HPV integrations called from targeted capture sequencing. All the split reads that had one part aligned to human and other part aligned to HPV were selected as informative reads for the construction of two direction tree structure. All junctions (human-human, human-HPV, HPV-HPV) were called for each informative read. Each junction was stored in the tree structure as a node. The iteration began from searching the most abundant junction among all the reads and then extended the tree structure for 3'-5' and 5'-3' two directions. After one round, the remaining reads were repeated recursively storing their junctions until there was no informative read left. The approach walks the tree structure and prints out all the possible paths/assemblies and the read coverage at each junction. As loops were not permitted in a tree structure, redundant junctions within a read were recorded to resolve the

loops in the read. The read names were also stored in each node to potentially sort the different paths. Samtools [16] was also used in this pipeline for bam file implementations.

3.2.2 HPV fusion detection

We applied SearcHPV [17] on all targeted capture sequencing samples to call HPV-host integrations using hg38 catenated with the HPV genome as the reference. All types of HPV references were downloaded from PaVE [18] database (pave.niaid.nih.gov). We applied SurVirus to call HPV-host fusions on all RNA-seq data using hg38 catenated a transformed HPV genome[19] to avoid breaking HPV transcript by the origin of the circle. For long read DNA-seq data, minimap2 was used to align the reads to the same reference as TCS. We called HPV integrations and resolved the reads by our novel developed method described in Figure 3.1.

3.2.3 Normalization of copy number in targeted capture sequencing

The numbers of E6 and E2 reads at probe sites were validated by ddPCR with linearity of read depth and copy number, suggesting the applicability of normalizing the reads to calculate the copy number. Four genes (E2F2, MAP2K4, CD52, PPM1D) were found to have read depth with small variances in all the TCS samples normalized by the total number of reads in each sample. We chose MAP2K4 as the background gene to normalize the copy number for all TCS samples.

3.2.4 Comparison of HPV integrations called from Nanopore sequencing and Targeted capture sequencing

Because of the stolastic nature of sequencing alignments, many fusion callers adopted a small window from 10-50 bp to merge the junctions close to each other [17,20,21]. And for other methods, when benchmarking their results, deviation of virus integration locations could be more than 100bp. Some studies took a 200 bp window to define the same HPV integration events called from different methods [22]. To better compare HPV integrations from LR and TCS, we first clustered integrations within 50 kb as the same event. The choice of the 50kb bp window was based on previous studies [5,12]. For each event from two sequencing methods, if the closest two integrations were within 20 bp, we defined them as the common event.

3.2.5 Comparison of HPV integrations called from Targeted capture sequencing and RNA-seq

We set out to identify the common integrations across TCS and RNA-seq in an appropriate way. We also calculated the distance between nearest HPV integrations from TCS and RNA-seq (Figure 3.10), most integrations located within 50 kb. Considering the distinct sizes of cellular genes, if HPV integrations called from both TCS and RNA-seq fell into the same genes region, we defined them as the common events; if HPV integrations were intergenic then if the distance was within 50 kb, we assigned them as common events. Integrations fell into exact same locations across DNA and RNA level were also recorded.

3.3 Results

3.3.1 A novel approach based on long read sequencing resolved complex structures of HPV integrations with intratumoral heterogeneity

In the well-established OPSCC HPV16+ UM-SCC-47 cell line, numerous prior studies have consistently indicated clustered HPV integrations on chr3, near or within TP63, accompanied by extensive duplications, pointing towards complex fusion events [14,17,23–26]. In a previous investigation [17], we identified six high-confidence HPV integrations through TCS and conducted local assembly for UM-SCC-47 using Nanopore sequencing data with a published assembler wtbdg2 [27]. The resulting 60kb scaffold comprised a 15kb, twice-amplified human segment and a 7.5kb, twice-amplified HPV segment. However, the 10X linked plot suggested an overall duplication structure length of approximately 130kb [17].

From our TCS data and WGS reported by others [14], the copy number at HPV integrations of UM-SCC-47 varies from 10-50 copies, while the overall copy number at the HPV integration event region was uniformly presented as 50 copies (Figure 3.7). The variations at each HPV integration implied the complex arrangements might have multiple types of segments amplified for different times. To comprehensively resolve such HPV integration events with focal genomic amplifications, we developed a novel approach based on tree structure. This method demonstrated superior performance compared to existing whole-genome long-read assemblers based on de Bruijn graph algorithms in elucidating the local structure within large duplication regions. In this approach, we preserved the junction information from each read instead of merging edges strategy used by the de-bruijn graph based methods when dealing with

duplication structures (Figure 3.1 A-B). Our approach was applied on a LR-TCS paired cohort with 7 HNSCC cell lines, 1 PDX model and 8 HNSCC frozen tissue samples. One cell line was HPV18 positive, while the others were HPV16 positive. Figure 3.1C illustrates UM-SCC-47 and UM-SCC-104, highlighting two distinct types of HPV integration events.

In UM-SCC-47, the key structure duplicated 34 times containing one human segment starting and two HPV segments connected by the HPV genome origin. The human-HPV junctions flanking this structure were "E2 to TP63 intron10" and "TP63 exon14 to E2". The two HPV segments spanned nearly the entire HPV genome, with a 406bp gap at the E2 gene. Variant 1 of the key structure, with 12 copies, exhibited slight offsets at the human-HPV and HPV-HPV junctions at the HPV origin. Another structure, variant 2, encompassed variant 1 followed by an additional human segment from TP63 intron10 to intron11, partially amplified twice from the two HPV segments to the second human segment. Variant 3 initiated from a human-HPV junction, "E2 to TP63 intron10," connecting to a human segment, followed by a human-human junction indicating a deletion. The human segment extended from TP63 intron10 to intron11 and underwent 14 duplications (Figure 3.1C). The mapping plot in Figure 3.1C delineates the connections among these junctions within this intricate duplication of HPV integration events, highlighting intratumoral heterogeneity within this event. A portion of this structure was previously captured in an assembly generated by wtdbg2, indicating amplification twice (Lisa M. Pinatti et al., 2021). Another study attempted to elucidate the same model through whole-genome sequencing (WGS) and employed targeted PCR, Sanger sequencing, and chromosome walking to establish connectivities (Akagi et al., 2014). In their map, they identified the key structure amplified 20 times and variant 3 amplified twice.

Conversely, in the UM-SCC-104 cell line model, an entirely distinct structure revealed only one copy of the integrated HPV segments within the host gene SLC47A2 intron9. The integrated HPV broke at the E2 region, covering nearly the entire HPV genome with an 11bp gap and a 65bp overlap at the origin of the HPV reference (Figure 3.1C). This structure was corroborated by TCS data and Nanopore assembly from wtbdg2 (Figure 3.7-3.8). By resolving these two representative models, our novel approach demonstrated high resolution and more comprehensive performance compared to existing methods. The establishment of this method facilitated further confirmation of HPV integration events with complex structural variations.

3.3.2 HPV integration events can be classified into two types based on their association with complex rearrangement

In multiple prior studies, the count of integration clusters served as a parameter for grouping HPV integration events in HNSCC and cervical cancers [9,12,13]. To achieve a more comprehensive identification of different types of HPV integration, we introduced two additional parameters to assess the complexity of rearrangements at HPV integrations: the normalized local copy number and the maximum overall copy number at the HPV integration (Figure 3.2A). Utilizing our core cohort comprising 291 HNSCC FFPE HPV+ patients with TCS data, we explored the relationship among these three parameters and observed that a higher number of integrations clustered correlated with an increase in duplications. Notably, this association did not strictly follow linearity ($R^2 = 0.39$, $P = 3.41\times e^{-153}$) due to some HPV integration events exhibiting exceptionally high copy numbers (Figure 3.2B). By comparing the distribution of the

expected overall copy number at the integrations (green in Figure 3.2B) to the observed overall copy number (blue in Figure 3.2B), we noticed that the observed copy number displayed a steeper slope when associated with the number of clustered HPV integrations. This finding suggested that multiple HPV integrations might induce additional host genomic amplifications in the surrounding regions.

We subsequently determined the threshold to define the types of HPV integrations based on their biological and statistical characteristics. We phased the long reads for a PDX model, PDX-294R, using 10X linked read sequencing data. The result indicated that HPV integration occurred in only one chromatid as a heterozygous event (Figure 3.9). Previous studies have also adopted this characteristic of HPV to resolve the structure of HPV integrations (Akagi et al., 2023; Liu et al., 2023). Under such characteristics, if only one copy of HPV integrated into the human reference without duplications, the local copy number would be one, while the maximum overall copy number would be two. Employing these two cut-offs, we identified the threshold for the number of HPV integrations in a cluster within 50 kb (Figure 3.2C). Finally, we classified Type2 HPV integration events if the local copy number at the host-HPV breakpoint was greater than 1, the maximum overall copy number was greater than 2, and the number of HPV integrations in a cluster exceeded 2; otherwise, the HPV integration events were assigned as Type1. Integrations clustered within 50 kb were considered as one HPV integration event.

To validate our defined approach, we applied the same threshold in the LR-TCS paired cohort. HPV integrations were called from LR using the approach described in Figure 3.1. Due to distinct coverage between these two sequencing technologies, TCS exhibited higher sensitivity

than LR, as previously validated by PCR sequencing in UM-SCC-47 and PDX-294R [17]. Supportive rates were calculated based on common HPV integration events (see the method section) shared by these two sequencing technologies. The supportive rate from LR to TCS for Type2 events was 85%, while for Type1 events, it was 74%. The definitions of Type1 and Type2 events from LR and TCS demonstrated significant associations (P = 0.0002, Figure 3.2E-F).

3.3.3 Characteristics of Type1 and Type2 HPV integrations

The classification of HPV integration types was applied to both the core cohort and the LR-TCS paired cohort, as described in Figures 1 and 2E-F. The core cohort consisted of 251 HNSCC FFPE HPV+ patients, including 158 patients with normal/tumor paired samples, 15 patients with only tumor samples, and 78 patients with multiple recurrent samples (normal, local, recurrent, or metastatic sites). Six patients were HPV18+, four patients were HPV33+, and the remaining patients were all HPV16+. In total, 6870 HPV integrations were identified from 196 samples (core cohort: 6558 integrations from 181 samples in 179 patients; cell line and tissue cohort: 312 integrations from 15 samples). Of these, 10% (702/6870) of HPV integrations were defined as Type2, while the remaining integrations were classified as Type1. Approximately 42% (76/181) of samples contained at least one Type2 HPV integration. Type2 integrations were found to be distributed across the entire HPV genome and from chromosome 1 to X of the human genome. Several hotspots were observed, including EOMES, TP63, HEMGN, CASC11, MYC, LINC00484, TRMO, TRAF2, DTX4, PVT1, KLF4, and PTCSC2. Notably, multiple Type2 HPV integrations with high copy numbers were located on chromosomes 8 and 9 (Figure 3.3A). The

distribution on the HPV genome indicated that Type2 breakpoints were significantly enriched at the E1 and E2 genes, while fewer were located in E6, E7, URR, and L2 (Figure 3.3B).

3.3.4 Type1/Type2 HPV integrations have different impacts on transcriptomic level

In 162 HNSCC RNA-seq samples, 129 samples were paired with TCS data, and among them, 77 samples exhibited at least one HPV integration at either the DNA or RNA level. The method of comparing HPV integrations in DNA and RNA was detailed in the method section. Among these samples, 77 HPV integration in DNA level paired with 78 integrations in 17 patients. As depicted in the middle panel of Figure 3.4A, the number of HPV integrations at the DNA level could: (1) be more than that at the RNA level, as observed in EOMES; (2) be fewer than that at the RNA level, as in TP63; (3) be equal to that at the RNA level, as seen in most cases of Type1 events.

Another pattern emerged when comparing the breakpoints on the HPV genome for DNA and RNA levels. Multiple HPV breakpoints from RNA (bottom panel, Figure 3.4A) aligned with canonical HPV splicing sites (orange lines), a phenomenon not observed in DNA. Based on these findings, a chimeric transcripts model (Figure 3.4B) was proposed to illustrate different fusion patterns for various HPV integration events. If HPV is inserted into the intron of the human gene, these breakpoints might be spliced out during transcription. Breakpoints falling into the exon of the human gene could be preserved if the transcription process remains unblocked. For the HPV genome, if the entire genome is inserted, HPV might be transcribed as canonical isoforms, generating additional fusions only present in RNA levels. Non-canonical fusions on the HPV

genome might also be preserved and consistent at the DNA level if the inserted breakpoint is not spliced out or only part of the HPV gene is integrated and transcribed. To further explore the relationship between the types of HPV integration and transcription patterns, we compared the number of transcribed fusions of Type2 and Type1 with the number of fusions only in DNA (Figure C). Type2 predominated in transcribed fusions, while Type1 had more integrations only in DNA and not in RNA. A significant association was identified between Type1 and Type2 and whether HPV integration was transcribed ($P = 2.85 \times e^{-84}$). We then examined the number of integrations of Type1 and Type2 events that aligned with canonical splicing sites of HPV genes or non-canonical breakpoints on the HPV genome. Interestingly, Type2 events presented significantly more non-canonical breakpoints, while Type1 aligned more with canonical splicing sites ($P = 0.01$). Our results suggest that Type2 events are more likely to be transcribed as chimeric fusions and may be associated with more complex chimeric transcript patterns, resembling a mixture of models illustrated in Figure 3.4B.

To assess the influence of HPV integration on cellular gene expression, we investigated the distribution of gene expression levels surrounding Type 1 and Type 2 integrations within the range of 0-50 kb to 250-300 kb (Figure 3.5 A,D). Z-scores were normalized relative to HPV-negative samples. Type 2 integrations exhibited higher expression levels, evident as heavier tails compared to Type 1 integrations within the 0-200 kb range. However, this effect diminished as the distance increased beyond 200 kb. Specifically, oncogenes within 0-150 kb of Type 2 integration structures displayed significantly higher z-scores than genes up to 150 kb from Type 1 integration structures (binomial test, $P < 0.001$, plot not displayed). Furthermore, samples with

HPV integrations demonstrated elevated expression levels in nine genes listed in Figure 3.5B. The impact of HPV integration on oncogene expression levels was also distinguishable from other genes (Figure 3.5C, binomial test, $P < 0.05$).

3.3.5 Heterogeneity and clonal evolution could be induced by HPV integration events in recurrent patients

In the cohort of 78 patients with recurrent HNSCC+, we conducted an investigation on the patterns of HPV integration in 10 patients where HPV integrated into cellular genes (Figure 3.6A). The observed heterogeneity in HPV integration profiles revealed distinct aspects: (1) Variability in the copy number of HPV integrations within the same gene across different recurrent sites, exemplified by SCNN1A, TNFRSF1A, LTBR, and PLEKHG6 in SOP-075LR1-3. (2) Identification of multiple unique integrations in different samples from the same patients. We extended this analysis to explore HPV-HPV junctions, specifically examining two paired HPV+ HNSCC cell line models, UPCI:90 and UPCI:152. UPCI:90, originating from the base of the tongue, served as the primary tumor, while UPCI:152 represented a recurrent node situated at the hypopharynx approximately 1 year later. Both samples exhibited two major HPV-HPV junctions with high copy numbers, with UPCI:90 displaying higher local copy numbers for both junctions. Additionally, five unique HPV-HPV junctions were exclusively observed in UPCI:90, albeit with low copy numbers (Figure 3.6B). To gain a deeper understanding of the structure of Type2 integration events in this model, we employed our novel approach outlined in Figure 3.1 to resolve the HPV integration structures for these two models. A particularly complex event on Chromosome 9 is illustrated in Figure 3.6C. This structure amplified human and HPV segments

and notably flanked numerous inversions of human structural variations. While the key structure was shared by both models, there were variations in copy number (UPCI:90: 114X; UPCI:152: 37X). The host-HPV junctions were situated at "HEMGN intron1 to URR" and "E1 to intergenic upstream of HEMGN," while the HPV-HPV junctions fell within the URR regions. Two variant structures were shared by both models, with larger duplications observed in UPCI:90. Host-HPV junctions for variant 2 were "E6 to PTCSC2 intron1" and "intergenic upstream of PTCSC2 intron1 to URR"; for variant 3, they were "TRMO intron2 and E1." A unique structure, variant 1, was exclusively presented in UPCI:90 as a variant of the key structure. Intriguingly, all three genes involved in this event were situated on the negative strand of the human genome. Considering that HPV genes are on the positive strand, the inversion structures suggested the potential for this complex rearrangement to be transcribed. Our results indicate that the heterogeneity of integration structures can be summarized by variations in (1) copy number at HPV integrations; (2) independent HPV-HPV, HPV-host, and host-host junctions; (3) distinct characteristics of rearrangement breakpoints. Furthermore, by combining findings from TCS (Figure 3.6B) and LR (Figure 3.6C), we demonstrated that UPCI:90 displayed an overall more complex rearrangement with various unique junctions and much higher local copy numbers, suggesting a potential clonal selection process from UPCI:90 to UPCI:152.

3.4 Discussion

We introduced an innovative approach for deciphering complex HPV integrations with extensive rearrangements in nanopore sequencing data. Through benchmarking on UM-SCC-47, we demonstrated the capability of our method to generate more comprehensive and high-resolution

structures compared to existing methodologies. Prior investigations involving long-read sequencing on HPV integration often focused on resolving single reads or utilized in-house scripts [9,28,29]. Our method addresses a technological gap in the specific tools available for HPV integration research using long-read sequencing.

We established the criteria for Type1 and Type2 HPV integration events based on statistical and biological features in our extensive cohort with higher power, providing a potential standard cutoff for future studies in the field. We delineated the genomic locations of Type2 events on both HPV and the human genome. Notably, we identified several integration hotspots, including TP63 and MYC, which have been frequently reported in previous research [12,17,30,31]. Integration sites such as HEMGN, TRMO, TRAF2, CASC11, and KLF4 were also detected in various studies [12,17,32–35]. EOMES exhibited differential expression in HPV-positive cancers in multiple investigations [36,37]. Specifically, CASC11 and PVT1 were identified as potential biomarkers in cervical cancer [38,39], while KLF4 played a role in the life cycle of HPV31 late viral gene expression [40].

In our investigation, we unveiled distinctive transcriptomic profiles associated with Type1 and Type2 HPV integration events. Type2 events displayed a higher likelihood of transcription, potentially resulting in the generation of complex chimeric transcripts featuring non-canonical fusion sites. The occurrence of such non-canonical fusions aligns with observations from prior research [12]. The influence of HPV integration on neighboring cellular genes varied between Type2 and Type1 events. Oncogenes in proximity to integrations demonstrated elevated expression levels in Type2 events compared to Type1 events. This impact of HPV integrations

on nearby genes diminished as the distance increased, notably becoming less pronounced beyond 200 kb. A similar analysis conducted in a previous study [12] involving 194 genes with breakpoints highlighted a gradual decline in the impact of HPV integration when the distance reached 100-150 kb.

Recent studies have shed light on the intratumoral heterogeneity of HPV integration and potential clonal evolution through investigations of primary tumors [9,29]. Nevertheless, there is a scarcity of reports on the heterogeneity of HPV integration in recurrent patients. Our examination revealed heterogeneity in integration structures, encompassing variations in local HPV copy numbers, distribution of specific HPV integration sites within structures, and distinctive characteristics of rearrangement breakpoints (human-HPV, HPV-HPV, human-human). Building on these observations, we hypothesized that HPV integration events undergo clonal selection. To test this hypothesis, we compared the integration profiles of matched primary and recurrent tumors, focusing on one recurrent HPV+ model pair, UPCI:90 and UPCI:152, using long-read whole-genome sequencing. Our findings provide some of the initial evidence indicating the heterogeneity and clonal selection of HPV integration events during pathogenesis, further associating them with distinct transcriptomic profiles. Based on this data, we propose a new working model for disease pathogenesis, wherein the heterogeneity and clonal evolution of HPV integration events play a pivotal role in driving the disease process.

In a recent cervical cancer study, it was discovered that multiple HPV integrations are associated with poorer survival outcomes compared to single HPV integrations [9]. In our ongoing research, we plan to conduct survival analyses on Type1 and Type2 events, exploring the potential of

considering HPV integration types as prognostic biomarkers. Validation of our fusion sites pattern model will require comprehensive assembly of isoforms of chimeric transcripts. To gain deeper insights into the mechanism of clonal selection in recurrent patients, a comparison between preserved and lost integrations, along with their impact on gene expression might be insightful.

In summary, our study has filled a technological gap by resolving local complex HPV integration events from long-read sequencing, establishing a standard for analyzing different types of HPV integrations. We have demonstrated that Type2 HPV integration may drive distinct biology consequences, and the heterogeneity of HPV integration could serve as a critical driver of pathogenic progression. The classification of HPV integration has the potential to become a prognostic biomarker, aiding in the precise grouping of patients for tailored treatments.

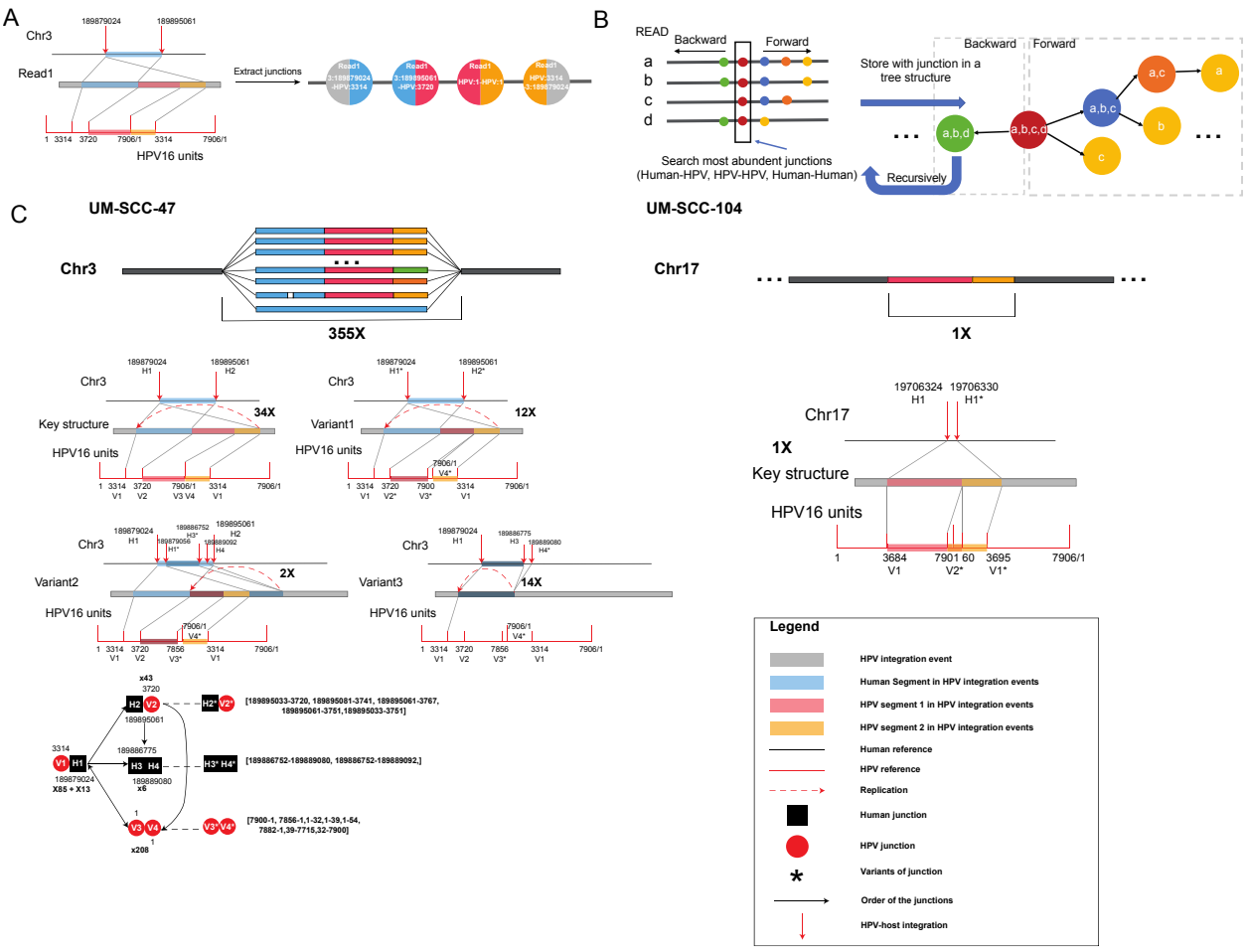


Figure 3.1 Type2 and Type1 representative models resolved by our novel approach. A. Schematic plot of storing a read in the nodes. A read from UM-SCC-47 was shown as an example. The four junctions were called for this read and stored as four nodes. B. The pipeline of our novel approach. All the informative reads were searching for the most abundant junctions and then stored to a two direction tree structure. This process repeated until not read left. C. UM-SCC-47 as an example of Type2; UM-SCC-104 as an example of Type1. These HPV integration eventswere resolved from Nanopore sequencing using the novel approach. Here we only presented the structure with more than 3 reads coverin the nodes. The number of most abundant junctions were used as the copy number to summarize the whole structure. In the names of junctions, “H” noted “human”; “V” noted “Virus”; numbers indicated different junctions. If two junctions were within 10bp, the same number will be assigned but “*” was added to indicate a variant form. Such cases of variant form junctions were displayed in square brackets. Detailed legends were shown in the legend panel.

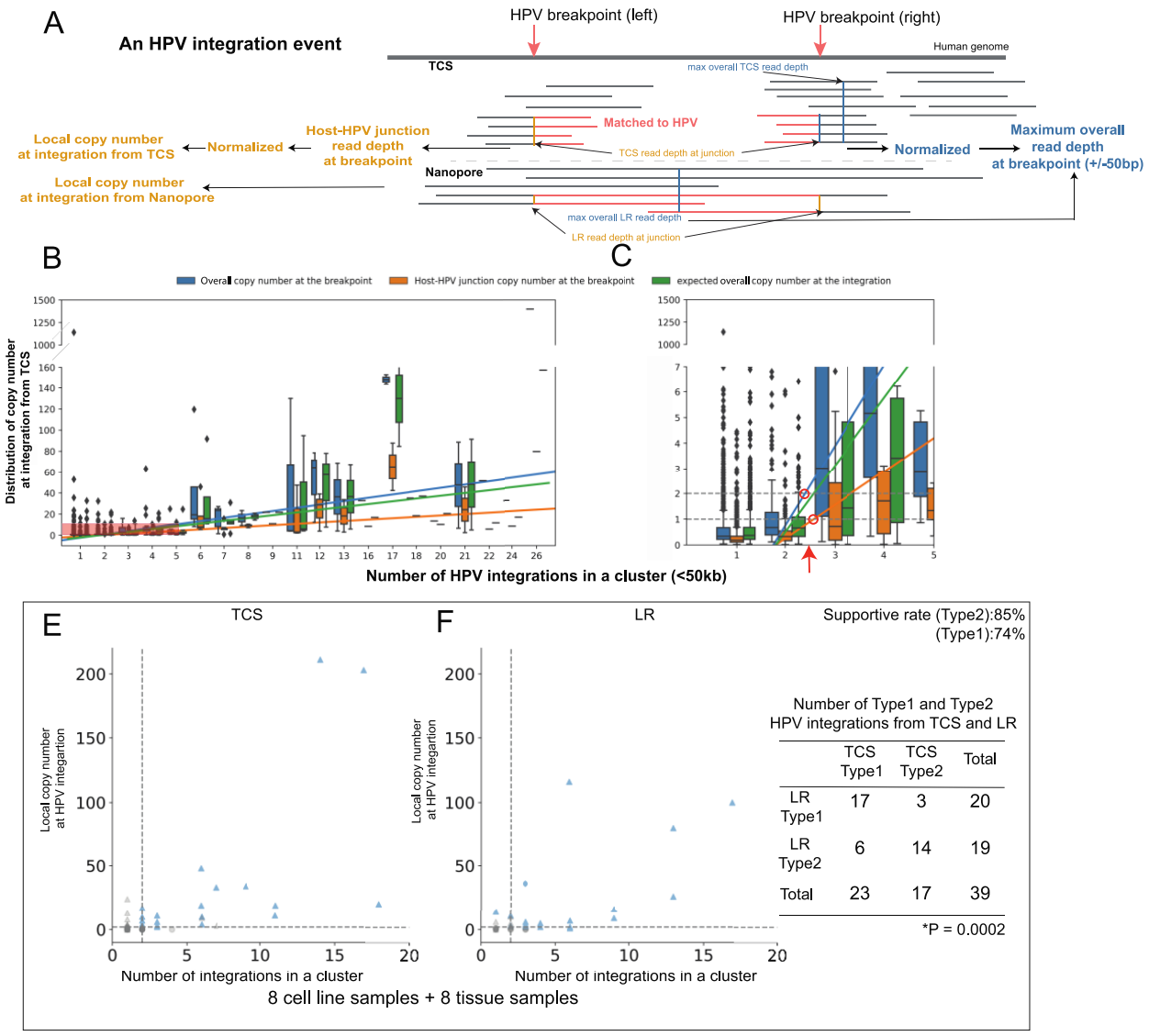


Figure 3.2 Definition of Type1 and Type2 HPV integration events. A.Calculation of local copy number and overall copy number at HPV integration. Split reads at human-HPV breakpoints were counted to calculate the local copy number at the HPV integrations. The maximum read depth within 50 bp at the breakpoints were counted for maximum overall copy number. For TCS data, read depths were normalized based onbackground genes (See method section for details). As a single molecular technology, for nanopore data, the normalization step was not performed. B-C. Statistical summary for three parameters of HPV integration events. X-axis: number of HPV integration clustered within 50kb. Y-axis: distribution of copy number at integrations from TCS. The overall copy number and local copy num ber were in blue and orange, respectively. The expected overall copy number was in green and calculated as the double of the local copy number based on the hypothesis that HPV only integrated into one chromatid as a heterozygous event. Linear regression models were fitted as references after removing one outlier suggested by leverage (See Figure 3.8). Part of B was zoomed in as C. The two dashed horizontal lines indicated the normal copy number if only one copy of HPV integrated into the host genome without duplications. The red arrow denoted the intersection of observed copy numbers and the normal copy numbers as the number of HPV integrations clustered increased. E-F. Type2 HPV integrations defined in TCS and LR data. Type2 were colored in blue and defined by the number of integrations in the cluster greater than 2; local copy number greater than 1; maximum overall copy number greater than 2. Triangle denoted HPV integration events with the maximum overall copy number greater than 2. The supportive rate of definition for types of HPV integrations using different technology

were listed in the plot. Chi-square test was performed on different types defined by two technologies showing significant association between the results from LR and TCS.

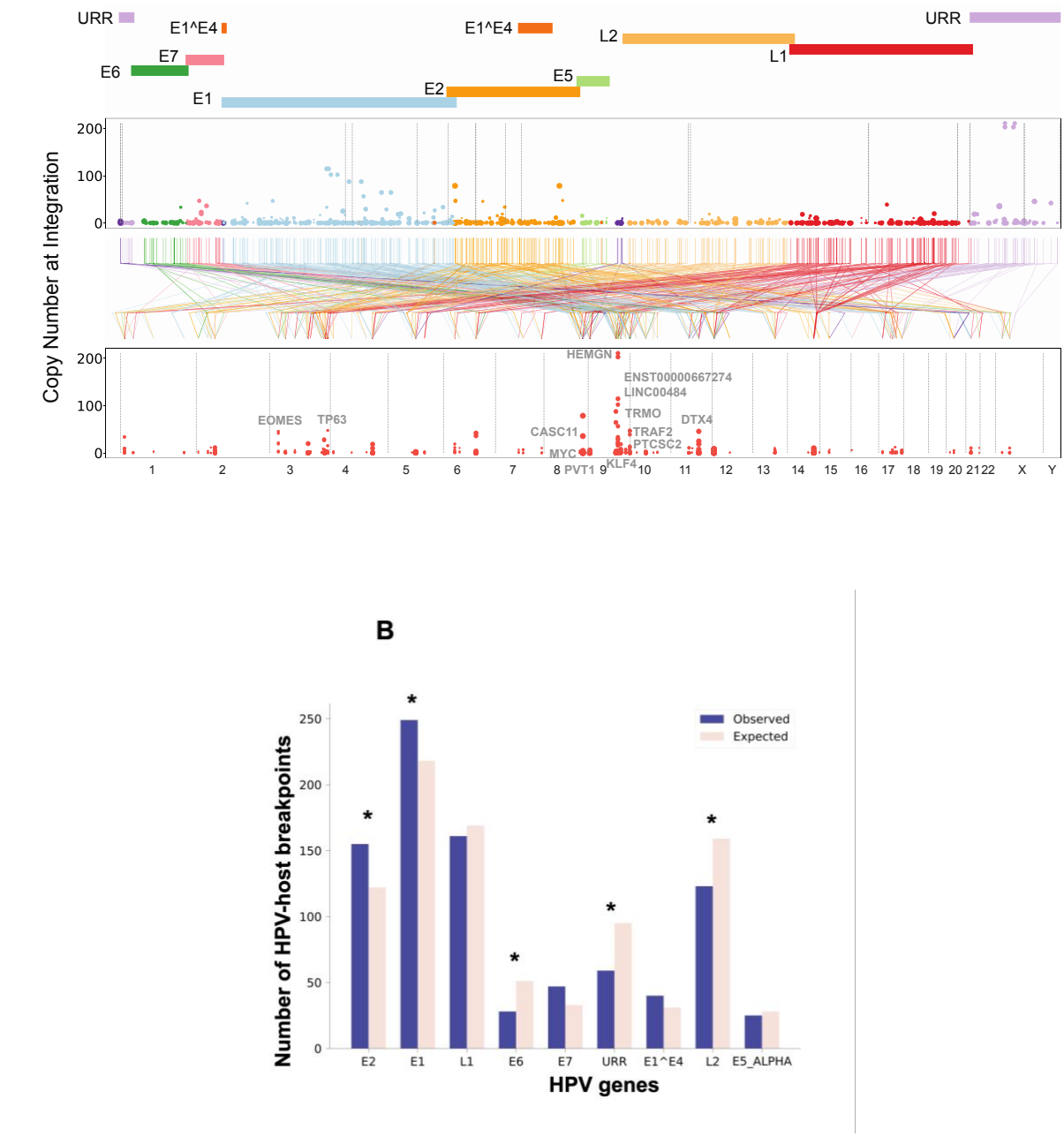


Figure 3.3 HPV breakpoints for Type2 events on human genome and HPV genome. A. Bottom panel: Type2 HPV integration events in human genome; Top panel: Type 2 integration events in HPV genome. Breakpoints on the human and HPV genome were linked and colored by the HPV gene the breakpoints fell into. The size of the dot denoted the number of integrations in this event. Note that since HPV integration events were defined based on location of breakpoints on the human genome, one HPV integration event might have multiple breakpoints on the HPV genome. Hotspot genes were marked. B. Distribution of HPV breakpoints on HPV genome. Chi-square test was performed. "*" indicated the significant difference.

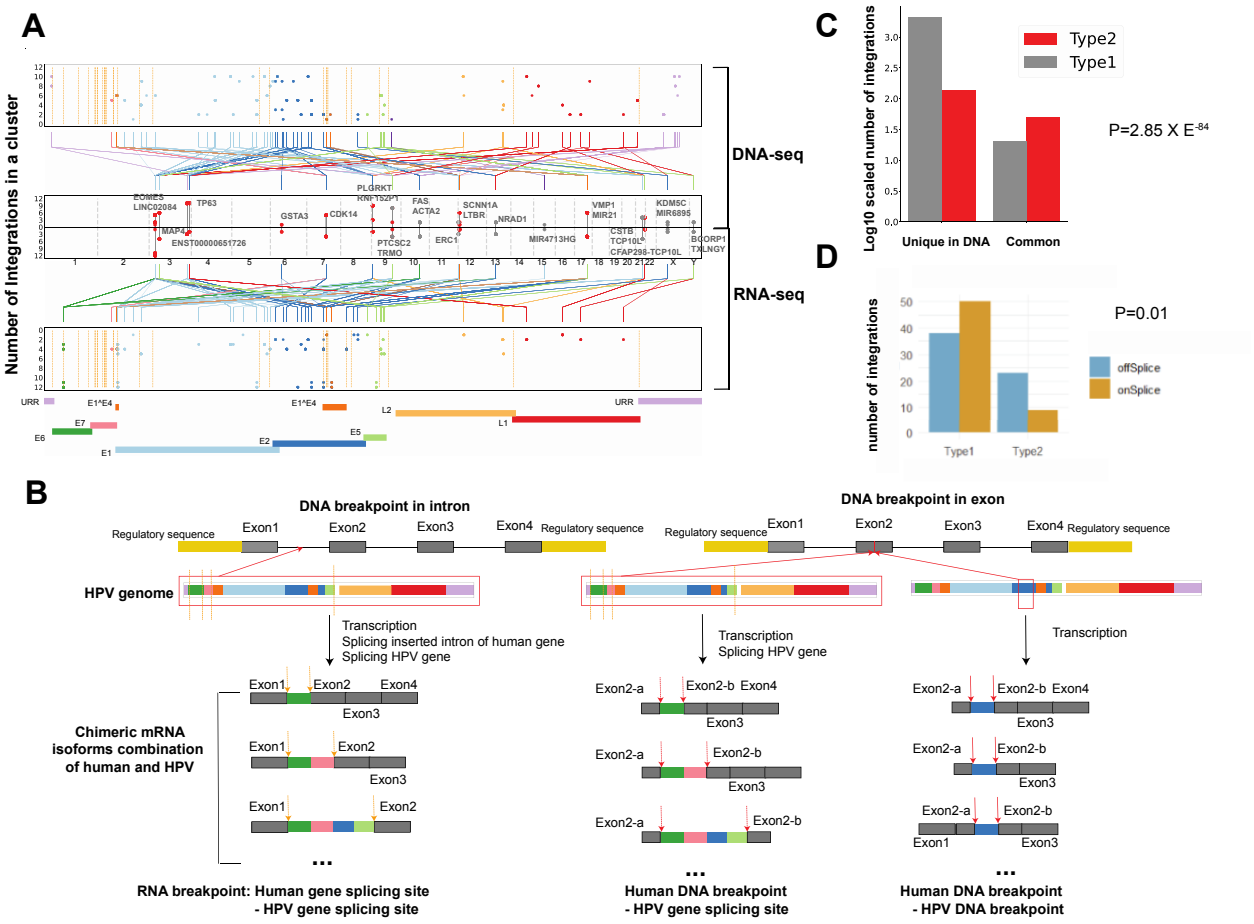


Figure 3.4 Type1/Type2 HPV integrations in DNA and RNA paired samples. A. Common HPV-host fusions between TCS and RNA-seq. Top panel: HPV breakpoints on HPV genome of TCS; Bottom panel: HPV breakpoints on HPV genome of RNA-seq. Each breakpoint was presented as a dot colored by HPV genes. Middle panel: common HPV fusions on human genome from TCS and RNA-seq. Red: Type2; Grey: Type1. Each paired HPV fusion event was connected. Y-axis indicated the number of HPV integrations in the event (clustered within 50k bp). Breakpoints on the human and HPV genome were linked and colored by HPV genes. Orange dashed lines on top and bottom panel were canonical splicing sites of HPV transcripts (https://pave.niaid.nih.gov/ (Van Doorslaer et al. 2017)). B. Models of chimeric transcripts. chimeric transcripts were generated by HPV integration events that were transcribed. If HPV integrated into the intron of the gene, after splicing, the original breakpoints in DNA level were lost. We observed HPV fusions the same as the splicing sites. If HPV integrated into exon regions, chimeric transcripts might preserve the fusions in DNA level. The same logic also applied to HPV integrated parts and resulted in three different patterns of human-HPV fusions in transcriptomic level: human gene splicing site - HPV gene splicing site; human DNA breakpoint - HPV gene splicing site; human DNA breakpoint - HPV DNA breakpoint. C. Number of transcribed Type1 and Type2 HPV integration events. If a HPV integration event was identified in both DNA and RNA level, we regarded it as a transcribed event. If we only identified it in DNA level, then it was not transcribed as chimeric transcripts. Different types of integrations were significantly associated with whether they were transcribed. Chi-square test: P = 2.85 X e-84. D. Number of Type1 and Type2 integrations with different patterns of human-HPV fusions. OnSplice: HPV fusion aligned with canonical HPV gene splicing sites; OffSplice: HPV fusion not aligned to HPV gene splicing sites.

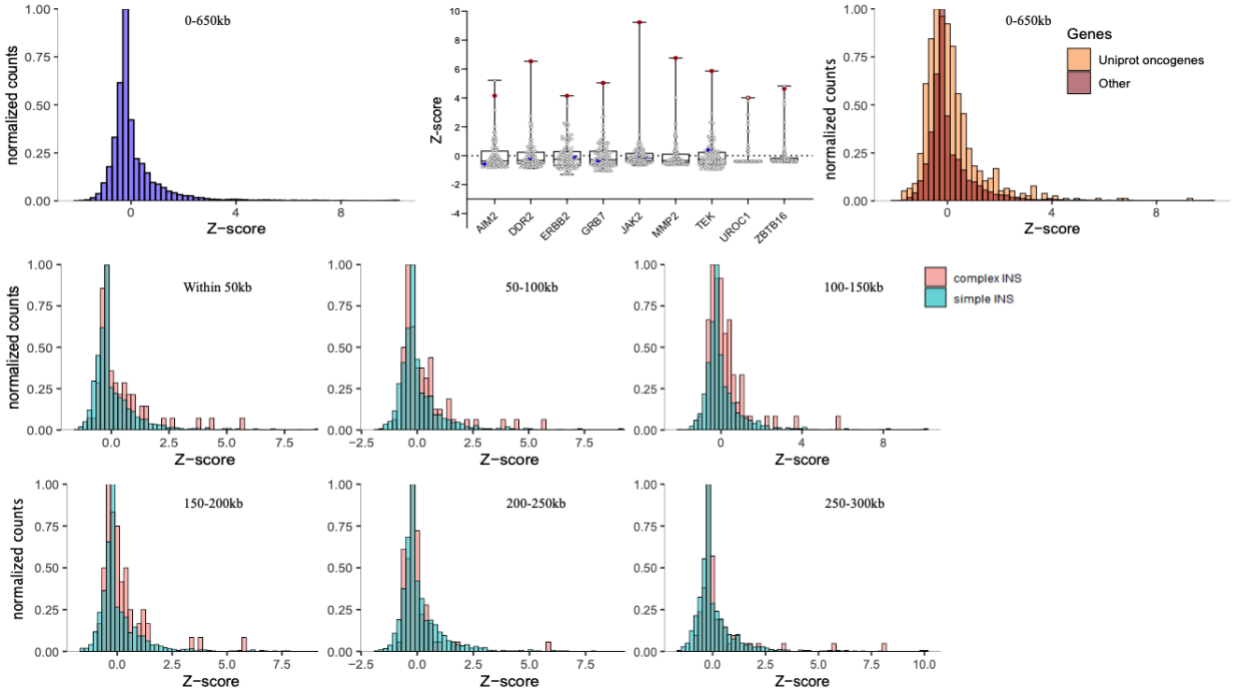


Figure 3.5 RNA expression levels of Type2 and Type1 HPV integrations. A. Distribution of normalized RNA expression levels of all HPV+ HNSCC samples surrounding HPV integrations within 650kb, controlled by the HPV- HNSCC samples. B. Genes with outlier expression levels at integration sites. Red dot: samples with Type2 integration sites. Blue dot: samples with Type1 integration sites. C. Distribution of normalized RNA expression levels for Uniprot oncogenes https://www.uniprot.org/ and other genes. D. RNA expression level of oncogenes surrounding Type1 and Type2 integration events within 50 kb to 300 kb.

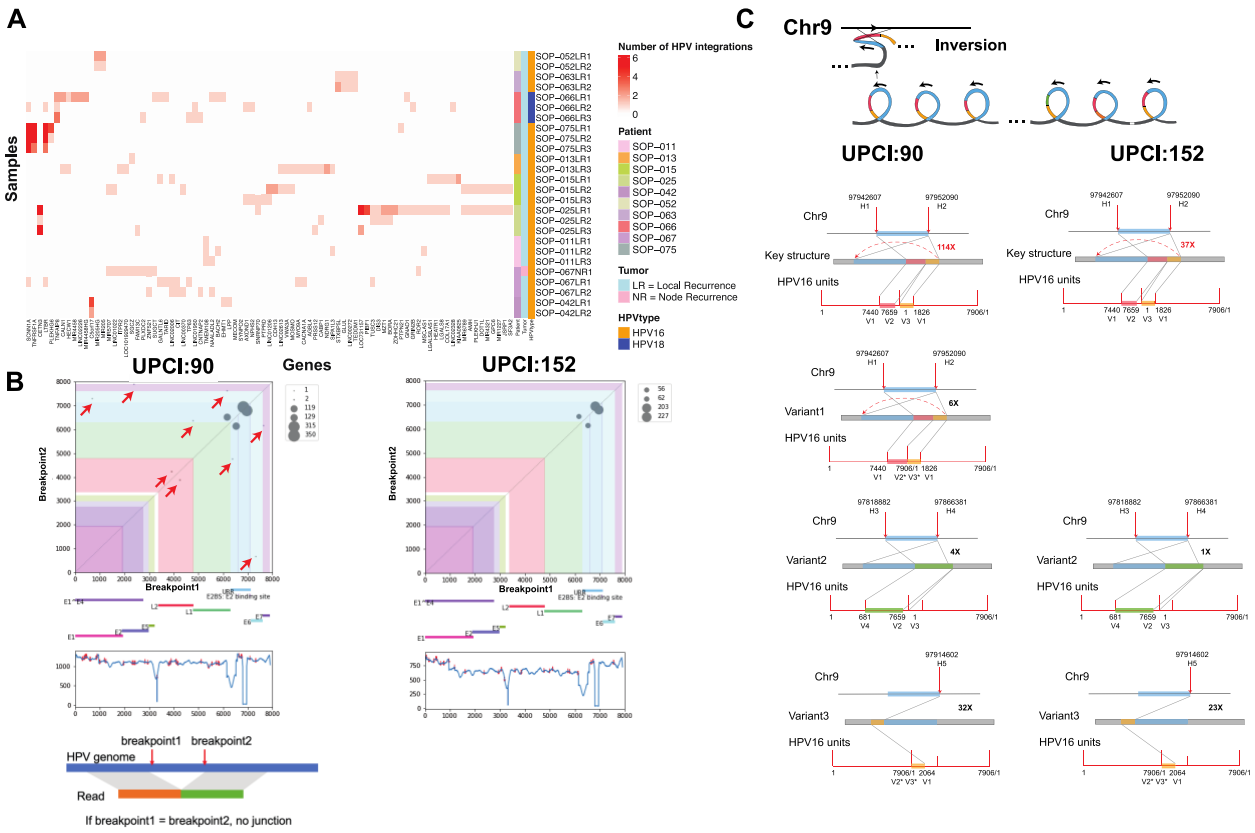


Figure 3.6 Heterogeneity of HPV integration events in recurrent patients. A. heatmap of the number of HPV integrations fell into genes in recurrent samples. X-axis: cellular gene names. Y-axis: recurrent samples. LR: local recurrent node; NR: node recurrent node. B. HPV-HPV junctions in UPCI:90 and UPCI:152 called from TCS. HPV-HPV junctions were displayed by the two breakpoints in each axis as depicted in the schematic plot. The size of dots indicated the normalized copy number at HPV-HPV junctions. Red arrow addressed small unique HPV-HPV junctions that only existed in UPCI:90. C. Structure of one Type2 HPV event on chromosome 9 in UPCI:90 and UPCI:152 resolved from LR. Only structures covered by more than 3 reads were displayed.

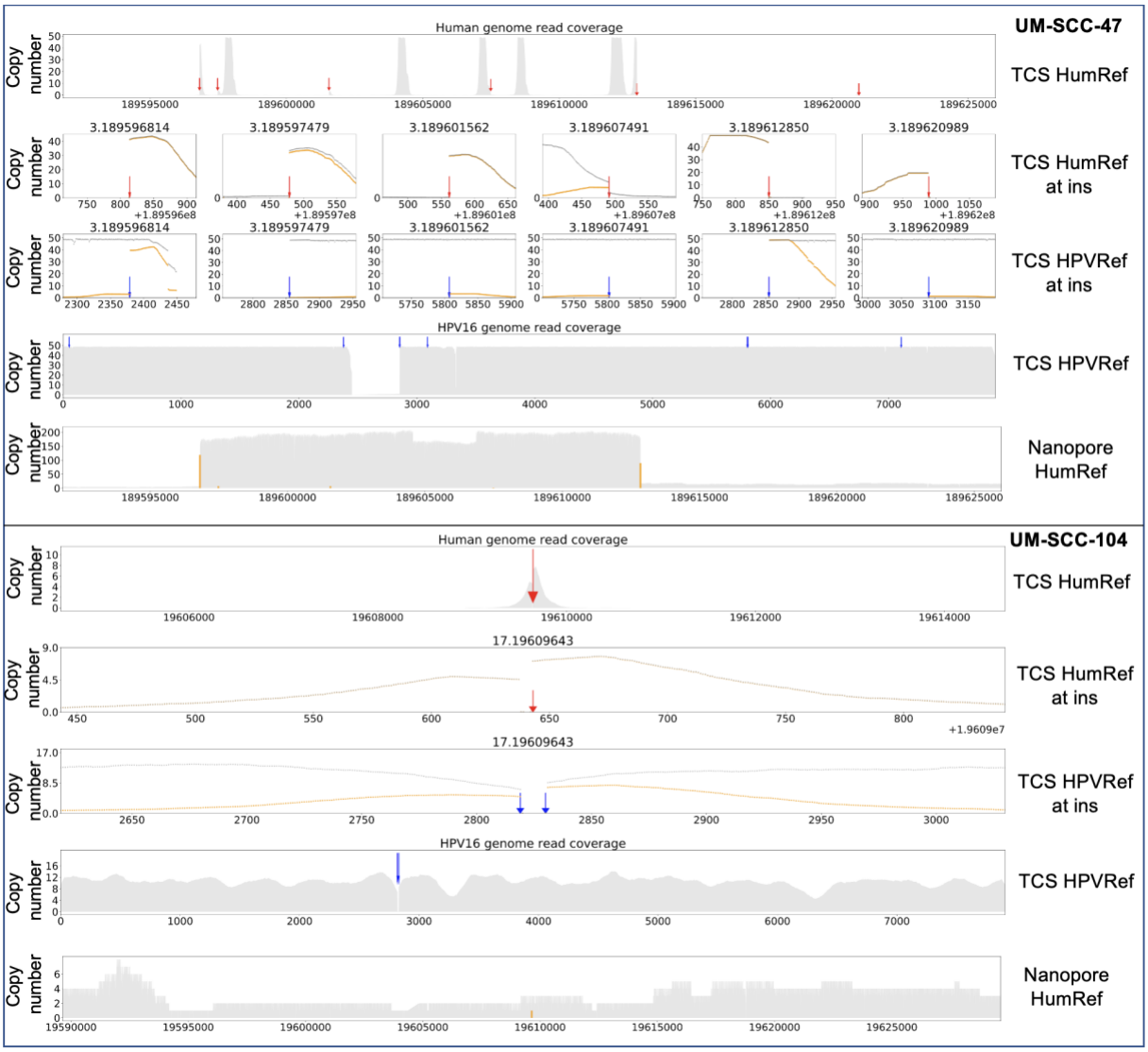


Figure 3.7 HPV integrations in UM-SCC-47 and UM-SCC-104 from TCS and LR. Red arrow: HPV breakpoints on human genome; Blue arrow: HPV breakpoints on HPV genome; Grey line and shade: Overall copy number (normalized read depth for TCS); Orange line: Copy number for informative reads (normalized read depth for TCS).

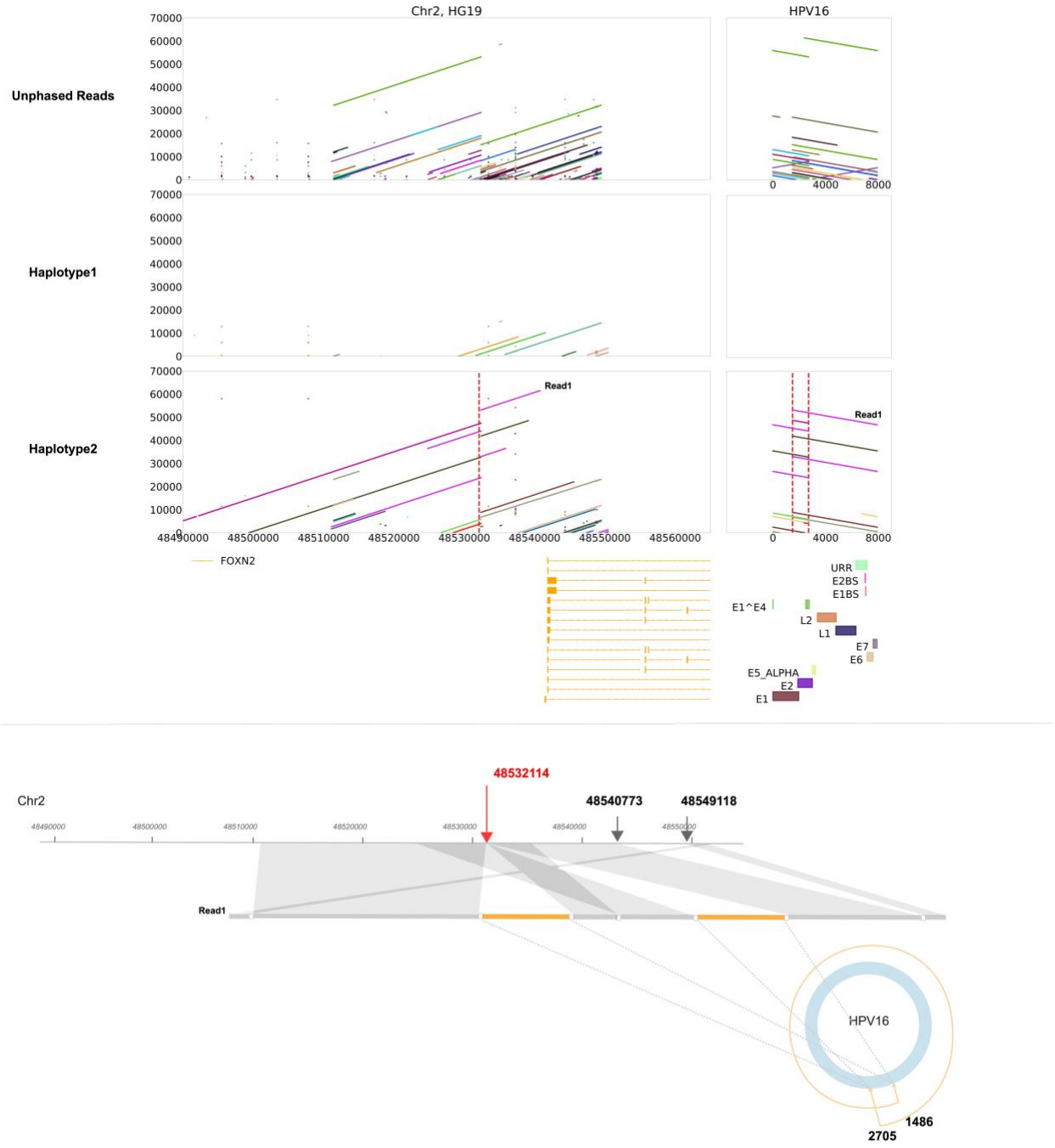
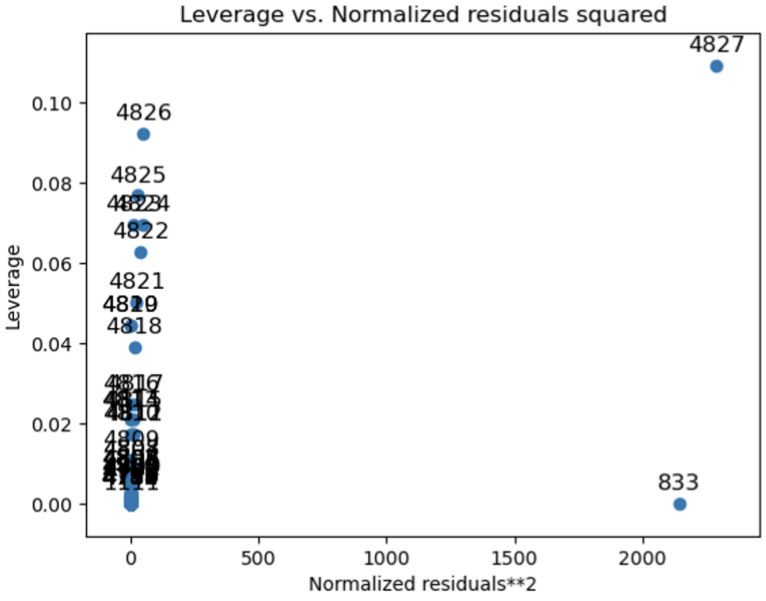


Figure 3.8 Phased LR reads for a Type2 event in *FXON2* of PDX-294R. Haplotype2 indicated multiple copies of HPV integrations. Haplotype1 had no HPV reads. The longest read in the region was displayed as a schematic plot to show the partial structure of this event. This event was aligned against the HG19 and HPV16 reference genome to keep consistent with 10X linked data.



OLS Regression Results

Dep. Variable:	SPCov	R-squared:	0.134
Model:	OLS	Adj. R-squared:	0.134
Method:	Least Squares	F-statistic:	747.8
Date:	Tue, 12 Sep 2023	Prob (F-statistic):	3.41e-153
Time:	01:32:53	Log-Likelihood:	-22314.
No. Observations:	4828	AIC:	4.463e+04
Df Residuals:	4826	BIC:	4.464e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-8.7851	0.516	-17.013	0.000	-9.797	-7.773
NumIns	8.9326	0.327	27.346	0.000	8.292	9.573

Omnibus:	14426.104	Durbin-Watson:	1.553
Prob(Omnibus):	0.000	Jarque-Bera (JB):	832983800.060
Skew:	42.867	Prob(JB):	0.00
Kurtosis:	2036.080	Cond. No.	2.88

Figure 3.9 Diagnostic of linear regression model for local copy number and number of HPV integration events. The outliers with highest leverage were filtered for the model fitting in Figure 3.2B.

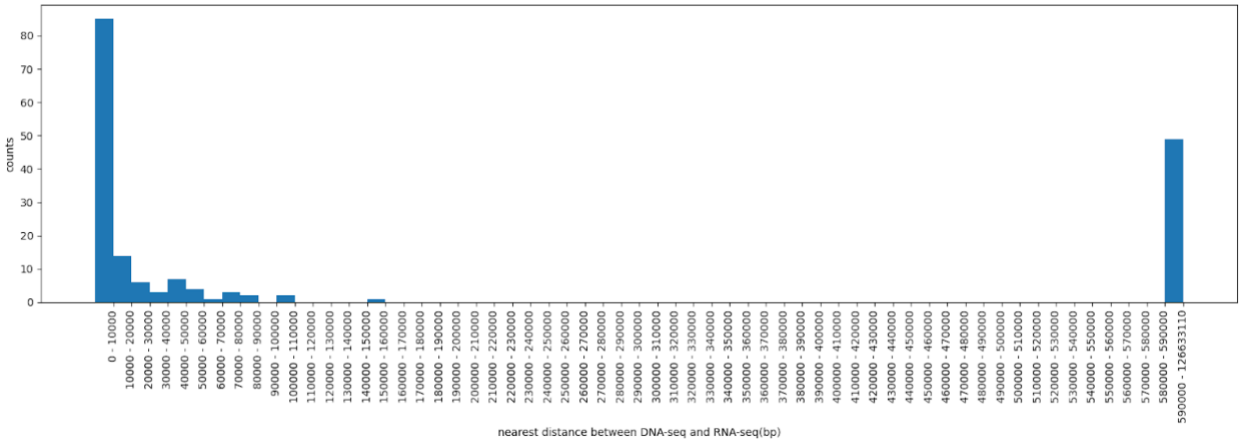


Figure 3.10 Distribution of nearest distance between TCS and RNA-seq

Bibliography

1. CDC. Centers for disease control and prevention. Centers for Disease Control and Prevention. Published January 5, 2024. Accessed January 8, 2024. https://www.cdc.gov/

2. Leemans CR, Braakhuis BJM, Brakenhoff RH. The molecular biology of head and neck cancer. Nat Rev Cancer. 2011;11(1):9-22.

3. Faraji F, Zaidi M, Fakhry C, Gaykalova DA. Molecular mechanisms of human papillomavirus-related carcinogenesis in head and neck cancer. Microbes Infect. 2017;19(9-10):464-475.

4. Pinatti LM, Walline HM, Carey TE. Human Papillomavirus Genome Integration and Head and Neck Cancer. J Dent Res. 2018;97(6):691-700.

5. Hu Z, Zhu D, Wang W, 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):158-163.

6. Oyervides-Muñoz MA, Pérez-Maya AA, Rodríguez-Gutiérrez HF, et al. Understanding the HPV integration and its progression to cervical cancer. Infect Genet Evol. 2018;61:134-144.

7. Olthof NC, Speel EJM, Kolligs J, 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):e88718.

8. Parfenov M, Pedamallu CS, Gehlenborg N, 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):15544-15549.

9. Zhou L, Qiu Q, Zhou Q, et al. Long-read sequencing unveils high-resolution HPV integration and its oncogenic progression in cervical cancer. Nat Commun. 2022;13(1):2563.

10. Wang C, Bai R, Liu Y, et al. Multi-region sequencing depicts intratumor heterogeneity and clonal evolution in cervical cancer. Med Oncol. 2023;40(2):78.

11. Warburton A, Markowitz TE, Katz JP, Pipas JM, McBride AA. Recurrent integration of human papillomavirus genomes at transcriptional regulatory hubs. NPJ Genom Med. 2021;6(1):101.

12. Symer DE, Akagi K, Geiger HM, et al. Diverse tumorigenic consequences of human papillomavirus integration in primary oropharyngeal cancers. Genome Res. 2022;32(1):55-70.

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

14. Akagi K, Li J, Broutian TR, et al. Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. Genome Res. 2014;24(2):185-199.

15. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094-3100.

16. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-2079.

17. Pinatti LM, Gu W, Wang Y, et al. SearcHPV: A novel approach to identify and assemble human papillomavirus-host genomic integration events in cancer. Cancer. 2021;127(19):3531-3540.

18. Van Doorslaer K, Li Z, Xirasagar S, et al. The Papillomavirus Episteme: a major update to the papillomavirus sequence database. Nucleic Acids Res. 2017;45(D1):D499-D506.

19. Yan B, Liu X, Zhang S, et al. DisV-HPV16, versatile and powerful software to detect HPV in RNA sequencing data. BMC Infect Dis. 2019;19(1):479.

20. Wang Q, Jia P, Zhao Z. VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. Genome Med. 2015;7(1):2.

21. Nguyen NPD, Deshpande V, Luebeck J, Mischel PS, Bafna V. ViFi: accurate detection of viral integration and mRNA fusion reveals indiscriminate and unregulated transcription in proximal genomic regions in cervical cancer. Nucleic Acids Res. 2018;46(7):3309-3325.

22. Tennakoon C, Sung WK. BATVI: Fast, sensitive and accurate detection of virus integrations. BMC Bioinformatics. 2017;18(Suppl 3):71.

23. Walline HM, Goudsmit CM, McHugh JB, 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):840-852.

24. Khanal S, Shumway BS, Zahin M, 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):30419-30433.

25. Myers JE, Guidry JT, Scott ML, et al. Detecting episomal or integrated human papillomavirus 16 DNA using an exonuclease V-qPCR-based assay. Virology. 2019;537:149-156.

26. Olthof NC, Huebbers CU, Kolligs J, et al. Viral load, gene expression and mapping of viral integration sites in HPV16-associated HNSCC cell lines. Int J Cancer. 2015;136(5):E207-E218.

27. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods. 2020;17(2):155-158.

28. Liu M, Han Z, Zhi Y, et al. Long-read sequencing reveals oncogenic mechanism of HPV-human fusion transcripts in cervical cancer. Transl Res. 2023;253:80-94.

29. Akagi K, Symer DE, Mahmoud M, et al. Intratumoral Heterogeneity and Clonal Evolution Induced by HPV Integration. Cancer Discov. 2023;13(4):910-927.

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

31. Rusan M, Li YY, Hammerman PS. Genomic landscape of human papillomavirus-associated cancers. Clin Cancer Res. 2015;21(9):2009-2019.

32. Ragin CCR, Reshmi SC, Gollin SM. Mapping and analysis of HPV16 integration sites in a head and neck cancer cell line. Int J Cancer. 2004;110(5):701-709.

33. Rodriguez I, Rossi NM, Keskus A, et al. Insights into the Mechanisms and Structure of Breakage-Fusion-Bridge Cycles in Cervical Cancer using Long-Read Sequencing. medRxiv. Published online August 22, 2023. doi:10.1101/2023.08.21.23294276

34. Holzhauser S. Effect of Ionising Radiation on HPV-Positive and HPV-Negative Oropharyngeal Cancer Cell Lines. phd. Cardiff University; 2018. Accessed January 8, 2024. https://orca.cardiff.ac.uk/id/eprint/120510/

35. Mainguené J, Vacher S, Kamal M, et al. Human papilloma virus integration sites and genomic signatures in head and neck squamous cell carcinoma. Mol Oncol. 2022;16(16):3001-3016.

36. Baedyananda F, Chaiwongkot A, Varadarajan S, Bhattarakosol P. HPV16 E1 dysregulated cellular genes involved in cell proliferation and host DNA damage: A possible role in cervical carcinogenesis. PLoS One. 2021;16(12):e0260841.

37. Tosi A, Parisatto B, Menegaldo A, et al. The immune microenvironment of HPV-positive and HPV-negative oropharyngeal squamous cell carcinoma: a multiparametric quantitative and

spatial analysis unveils a rationale to target treatment-naïve tumors with immune checkpoint inhibitors. J Exp Clin Cancer Res. 2022;41(1):279.

38. Hsu W, Liu L, Chen X, Zhang Y, Zhu W. LncRNA CASC11 promotes the cervical cancer progression by activating Wnt/beta-catenin signaling pathway. Biol Res. 2019;52(1):33.

39. Wang X, Wang G, Zhang L, Cong J, Hou J, Liu C. LncRNA PVT1 promotes the growth of HPV positive and negative cervical squamous cell carcinoma by inhibiting TGF-β1. Cancer Cell Int. 2018;18:70.

40. Moody C. Mechanisms by which HPV Induces a Replication Competent Environment in Differentiating Keratinocytes. Viruses. 2017;9(9). doi:10.3390/v9090261

Chapter 4 Analysis of Human Papilloma Virus Content and Integration in Mucoepidermoid Carcinoma

This chapter was published in 2022 in Viruses (PMID: 36366450). The author of this dissertation served as the first author of this paper. The main text and supplementary figures of this paper was presented below. Other supplementary materials could be referred to the published journal.

4.1 Introduction

Mucoepidermoid carcinomas (MEC) are the most common malignancies of the salivary glands comprising between 30–40% of all salivary gland cancers [1,2,3]. While MECs commonly arise in the parotid gland, they can occasionally form in other head and neck sites including the submandibular and sublingual glands, as well as the minor salivary glands of the oropharynx, oral cavity, and sinonasal cavities. Disease-specific survival is variable for patients with MEC and is dependent on factors such as histologic grade, tumor location, tumor stage, nodal status, patient age, margin status, and perineural invasion [2,3,4]. Importantly, however, it remains extremely challenging to differentiate between aggressive and non-aggressive MEC, which in the future may be improved by better understanding the molecular composition of this disease. In fact, a series of highly recurrent genetic alterations in MEC that lead to a Chr(11;19) (q14–21;

p12–13) rearrangement and induce the formation of a *CRTC1-MAML2* fusion gene is one of the most widely studied alterations in this disease [5,6,7,8,9]. At present however, the prognostic significance of the *CRTC1-MAML2* gene fusion in MEC is unclear, further supporting the need to better define molecular drivers of the disease [6,9].

Given both the anatomic distribution of MEC primary sites and the well-established role of high-risk human papillomavirus (HPV) as drivers of certain head and neck squamous cell carcinoma (HNSCC), it is interesting to speculate whether HPV may be associated with MEC as another potential molecular driver. Accordingly, given the multiple studies showing a strong prognostic role of HPV in HNSCC [10,11] as well as the emerging role of HPV ctDNA in HNSCC disease monitoring [12,13,14], it is important to evaluate whether HPV content could have a similar molecular role in MEC. Unfortunately, however, there has been controversy in the literature about the role of high-risk HPV in MEC. Indeed, a 2011 study by Brunner et al. demonstrated that 2/6 (33.3%) of MEC cases analyzed contained high-risk HPV DNA by in situ hybridization as well as diffuse p16 overexpression [15]. In contrast, a study by Isayeva et al. used nested RT-PCR on RNA extracted from 98 MEC samples to show a much higher HPV positivity rate of 35/98 (36%) (where 23% of tumors contained HPV16, 6% contained HPV18, and 7% contained both HPV16 and HPV18) [16]. These authors presented orthogonal data using several HPV detection approaches including in situ hybridization, HPV16/18 E6 immunohistochemistry, and additional PCR-based validations to support the presence of HPV in their cohort. In strong contrast to this data, Bishop et al. used RNAscope-based ISH analysis with the HPV HR [17] probe set to demonstrate that a cohort of 71 MEC cases were all HPV negative[17]. The authors concluded that HPV does not appear to have any etiologic role in MEC carcinogenesis,

independent of *CRTC1-MAML2* fusion status, though they acknowledged that their data conflicted with Isayeva et al.

Given discrepant data around the prevalence of HPV in MEC, we sought to leverage our recently published MEC transcriptome data and advanced bioinformatics techniques to clarify the prevalence and status of HPV in MEC.

4.2 Materials and Methods

4.2.1 Clinical specimens and annotation of viral genomes

A retrospective cohort of patients with MEC was previously identified from the University of Michigan pathology archive using an Institutional Review Board (IRB)-approved protocol for next generation sequencing of DNA and RNA (HUM00080561). However, patients were not consented for deposit of data in public databases. The cohort was previously typed for *CRTC1/3-MAML2* gene fusion status by RT-qPCR [18]. As previously noted, clinical, histologic, and outcomes data were collected from medical records and death was documented from electronic medical record notes and the Social Security Death Index [18]. Total RNA was previously submitted to the University of Michigan DNA sequencing core for library preparation and sequencing using the Illumina TruSeq Stranded Total RNA library prep kit. This data were previously summarized [19]. Here, to study viral content in greater detail, we leveraged the HPViewer pipeline using default settings to characterize HPV read counts in the cohort [20]. A previously defined threshold of > 5 reads was required to call a sample HPV positive.

4.2.2 Immunohistochemistry

Immunohistochemical staining was performed on the DAKO Autostainer (Agilent, Carpinteria, CA, USA) using Envision+ and diaminobenzadine (DAB) as the chromogen. De-paraffinized sections were labeled with the mouse p16 Ab-1 (DCS-50.1/47) (Neomarker, MS-218-P) and a mouse Ab (Thermofisher, Wyman Street, Waltham, MA, catalog #31430) was used as secondary antibody. Microwave epitope retrieval as specified was used prior to staining for all antibodies. Appropriate negative (no primary antibody) and positive controls (as listed) were stained in parallel with each set of slides studied. p16 immunostained slides were analyzed as previously described by our team [21].

4.2.3 HPV16 capture-based targeted DNA sequencing and analysis

Formalin-fixed paraffin-embedded (FFPE) blocks were obtained for the single patient with HPV16+ MEC (as described below) and five additional patients without HPV reads by HPViewer, selected at random, for confirmatory NGS-based tumor analysis. Regions with >60% tumor content, as identified by our head and neck pathologist (J.B.M.), were identified for DNA isolation with the Qiagen Allprep DNA/RNA FFPE kit (Qiagen, Hilden, Germany). Using the DNA Thruplex kit for library preparation (Takara Biosciences), targeted capture sequencing on DNA that passed our quality control standards was performed by the University of Michigan Advanced Genomics Core as previously described [22]. We employed a custom-designed probe

panel from Nextera that included high density probes covering the HPV16 genome as well as probes for targeting several common cancer-related genes [23]. Following library preparation and capture, the samples were sequenced on an Illumina NOVASeq6000 using a 300-cycle run and FastQ files were archived. HPV integrations were called by SearcHPV [24], an HPV integration caller that we recently developed for the detection of HPV-human integration loci from targeted capture DNA sequencing data. Downstream analysis was performed with R 3.6.1 and Python.

4.2.4 RNA-seq data analysis

Quality of the sequencing reads was evaluated using FastQC v.0.11.5. The quality reports did not reveal any adapter contamination; therefore, it was not considered necessary to perform quality trimming. The reads were mapped to the hg19 reference genome following a two-step alignment workflow of STAR v2.5.3a. Next, samtools v1.2 was used to extract uniquely mapped reads and Cufflinks v2.2.1 was used to generate the FPKM data. The *--max-bundle-frags* parameter of cufflinks was adjusted from its default value of 1,000,000 to 100,000,000 to allow us to compute FPKM at loci with high depth of coverage. Additionally, we applied SurVirus [25] and SearcHPV [24] to identify potential HPV-host fusions on the one HPV+ MEC sample (MEC1).

4.2.5 HPV oncogene expression analysis

The first step in viral oncogene expression analysis was to build a reference genome of human and viral sequences. For this purpose, we used a modified version of the HPV16 genome, and its corresponding annotation file as described in [26]. The human genome sequence was obtained from

the 1000 Genomes Project (Phase II). RSEM v1.3.3 was then used to build reference files of these human and modified HPV sequences. The same pipeline was also used to estimate gene expression levels from the HPV positive MEC (MEC1) RNA sequenced sample.

4.3 Results

We recently performed comprehensive transcriptome sequencing on 48 FFPE MEC tumors with a majority of tumors arising in the parotid. To now characterize the HPV content of tumors within this cohort, we analyzed the data using the HPViewer algorithm [20]. This analysis nominated only one of the forty-eight tumors as potentially HPV positive, with high HPV16 read counts (Figure 4.1A, Supplemental Table S1). We then evaluated p16 protein expression, a marker known to correlate with HPV status in oropharyngeal HNSCC, and found that MEC1 showed diffuse positive cytoplasmic and nuclear staining of p16 by immunohistochemistry, while none of the five selected HPV-negative MEC samples stained positive (Figure 4.1B), suggesting that p16 may also function as a marker of HPV status in MEC.

Clinically, MEC1 was a locally recurrent MEC of the anterior ethmoid sinuses in a 51-year-old male who underwent anterior sub-cranial resection with pathology showing high-grade MEC without perineural invasion and with negative margins. The patient had initially presented three years prior with a high grade MEC of the anterior ethmoid sinuses treated with subtotal resection followed by adjuvant radiation. He is now alive with no evidence of recurrent disease over 19 years out from salvage surgery. Our previous RT-PCR molecular sub-typing analysis of his tumor demonstrated the presence of a *CRTC1-MAML2* fusion [18].

To further validate our HPV annotation of this cohort, we performed PCR and Sanger sequencing of genomic DNA from MEC1, which confirmed the presence of HPV16 DNA (Figure 4.4) in this tumor. Accordingly, 0/5 of selected cases without HPV reads in our RNAseq data were also confirmed to lack HPV16 DNA by this method (data not shown). We then performed targeted capture NGS with a custom high-density HPV16 capture panel on the tumor DNA, which demonstrated high HPV16 read counts from MEC1, but not MEC23, which had no RNA-seq support for any HPV and served as a negative control (Figure 4.2). Analysis of host control genes in both of the targeted libraries confirmed that both were successfully sequenced to >500X depth (Supplemental Table S2).

To test for sites of HPV integration in MEC1, we used our recently described SearcHPV pipeline[24] to perform HPV-host integration analysis and identified 22 insertion sites in the host genome from targeted capture sequencing data (Figure 4.3A). Breakpoint sequence analysis of the integration sites indicated that most (21/22) HPV-host junctions in MEC1 have some degree of microhomology (Figure 4.3B). Further gene level analysis showed that 13 HPV integrations occurred in known genes, with an in-line insertion into the *TMEM163*, *HIP1,* and *SIRT1* genes and reverse orientation insertions in the remaining ten integrations. (Figure 4.3C, Supplemental Table S3). Seven of these genes were expressed at a lower level than the median of all MECs analyzed; five genes were expressed higher than the median; and one gene, RP11-354K1.1, was not expressed in any of the MECs (Figure 4.3D). Finally, expressed HPV-host integration transcripts were not identified from RNA-seq for MEC1 by two different callers, SearcHPV [25] and SurVirus [26].

4.4 Discussion

Our primary objective in the present study was to utilize our advanced bioinformatics pipelines to evaluate for the presence and physical state of transcriptionally active HPV in MECs of various major and minor salivary gland subsites. We found an exceedingly low prevalence of HPV (1/48 tumors, 2.1%) in our MEC cohort, with the single positive case of recurrent MEC of the anterior ethmoid sinus harboring transcriptionally active HPV16+ DNA with multiple complex integration events into various cancer-related genes. Concurrently, this tumor showed upregulation of p16 by IHC and altered expression of host genes affected by viral integration events. Our data raises several important points for the discussion of MEC tumor biology and the development of clinically useful, predictive, and prognostic biomarkers for this disease.

The etiologic role and prognostic implications of HPV in head and neck malignancies besides squamous cell carcinoma of the oropharynx remains a contentious and much-debated topic in our field [27]. Validation of HPV as a causative driver of other head and neck malignancies would have vast and exciting implications for treatment selection and prognostication. The data on HPV in major and minor salivary gland malignancies is inconclusive and limited by inconsistent HPV detection methods, small patient cohorts, and heterogeneous tumor histologies and subsites [15,16,17]. For example, a 2009 study by Vageli et al. utilized HPV L1 consensus PCR and RT-PCR to analyze HPV status in nine parotid gland tumors, including pleomorphic adenomas, Warthin's tumor, and acinic cell carcinoma [28]. The authors reported the presence of HPV16 or HPV18 DNA in seven of nine (77.8%) tumors and posited that high-risk HPV may be an etiologic agent in various salivary gland neoplasms. This preliminary data subsequently inspired later studies by Brunner et al., Isaveya et al., and Bishop et al. on the potential role of HPV in MEC [15,16,17].

Our findings of rare HPV positivity in MEC are consistent with Bishop et al. [17]. The authors of that study analyzed 92 MECs of various subsites with the objective of determining the prevalence of transcriptionally active HPV and its co-occurrence with *CRTC1-MAML2* translocation. They utilized HPV E6/E7 RNA ISH to assay HPV status but failed to identify transcriptionally active virus in any of their samples, independent of *CRTC1-MAML2* fusion status. Overexpression of p16 was not evaluated in their cohort. In a much smaller sample of six minor salivary gland MECs, Brunner et al. used HPV16/18 DNA ISH and p16 IHC to show the presence of HPV16/18 DNA in two (33.3%) MECs of the oral cavity [15]. Interestingly, strong and diffuse nuclear and cytoplasmic p16 staining on IHC was seen in both of these tumors but also in three additional MECs negative for HPV16/18 DNA by ISH. Due to the small number of tumors analyzed, the 33.3% HPV positivity rate may represent sampling bias rather than a true prevalence of transcriptionally active virus in MECs.

Our data conflicts most notably with that of Isayeva et al. in which the authors found a strikingly high prevalence of HPV16/18 DNA in 49/98 (50%) of their MEC cohort [16]. The authors used several complementary methods of HPV detection, including nested RT-PCR, HPV16/18 E6/E7 immunofluorescence, and HPV L1 consensus PCR to support their conclusion that high-risk HPV is convincingly implicated in the pathogenesis of MEC. However, their quoted HPV prevalence rate was based on nested RT-PCR only and they reported moderate discordance between HPV16/18 E6/E7 detection via immunofluorescence and their other detection methods. Further, the authors found no statistical correlation between HPV16/18 DNA detection via nested RT-PCR and p16 overexpression, tumor subsite, and tumor grade. Ultimately, the validity

of their conclusions remains uncertain due to failure to replicate this prevalence rate in multiple independent studies and lack of correlative p16 overexpression indicating biologically relevant HPV infection in MEC [15,17,29]. An advantage of our study over previous ones is the use of contemporary, high-throughput, sophisticated, and highly sensitive bioinformatics pipelines for the detection of HPV DNA and viral transcription [24,25,26].

Herein, we conclude that transcriptionally active HPV is a rare occurrence in MEC, independent of tumor subsite. Previous studies have failed to show a higher prevalence of HPV in MECs of anatomic subsites within lymphoid tissue of Waldeyer's ring. Further, there seems to be no predilection for HPV positivity in mucosal subsites or minor versus major salivary glands. Our single HPV-positive tumor was a recurrent MEC of the anterior ethmoid sinuses. It is interesting to speculate that, although an infrequent event overall, MECs of sinonasal subsites may be more prone to harbor HPV as a molecular driver. However, we cannot definitively reach that conclusion with our data.

Histologically, MECs are characterized by a variable, triphasic pattern of mucinous, intermediate, and epidermoid cells with histologic grade dependent on degree of nuclear atypia, mitoses, necrosis, perineural invasion, and cystic components [30]. Previous authors have not found any recurring histologic features characteristic of HPV-positive MECs that may differentiate these tumors from their non-HPV associated counterparts [16]. Further, transcriptionally active HPV does not seem to localize to mucinous, intermediate, or epidermoid components preferentially. Similarly, our single HPV-positive MEC did not harbor any distinguishing histologic features that may be of routine clinical utility (Figure 4.1). Thus, we conclude that the

presence of transcriptionally active HPV does not confer a distinguishable histologic pattern of MECs nor reliably impact tumor grade.

When MECs harbor transcriptionally active HPV however, the virus may significantly alter host gene expression by complex viral integration mechanisms. We showed that HPV integrated into 13 host genes, including *PIK3AP1, SIRT1*, *ARAP2*, *TMEM161B-AS1*, and *EPS15L1* as well as 9 non-genic regions. Interestingly, the PI3K pathway has been implicated in MECs, and alterations were identified in 52% of high-grade cases in a targeted sequencing analysis of one cohort, which was consistent with RNAseq analysis of 8 high-grade tumors showing a re-programming of PI3K signaling effectors [31]. Previous mechanistic data has shown that *PIK3AP1* expression drives AKT phosphorylation in gastric and thyroid cancer models, suggesting an oncogenic role of this gene in other cancers [32,33]. Notably, however, transcriptional analysis of MEC1 suggested that the integration events showed no significant patterns on the impact of the expression of these genes. Thus, we expect that the causal mechanism by which HPV contributes to pathogenesis in the HPV+ tumor identified in our cohort is through elevated E6 and E7 oncogene expression leading to p16 overexpression.

Despite this notion, viral integration into *SIRT1* is also of potential pathogenetic interest to MEC. *SIRT1* has a complex and multi-faceted role in cancer that includes the regulation of *TP53* as well as responses to DNA damage, metabolism stress, and inflammation [34]. *SIRT1* is a NAD^+-dependent Class III histone deacetylase that has been shown to antagonize cellular senescence [35] and is also an established negative regulator of *CRTC2*, and possibly *CRTC1* as well [36], suggesting that this integration event could enhance *CRTC1-MAML2* expression in this

tumor. Likewise, in the context of mutant p53, *SIRT1* acts as a tumor suppressor [37]. Given the established role of HPV16_E6 in repressing p53 activity, it is possible that *SIRT1* could act as a tumor suppressor in the context of HPV16+ MEC1, consistent with the relatively low expression of *SIRT1* observed in this sample. To our knowledge, *ARAP2*, *TMEM161B-AS1*, and *EPS15L1* have not previously been indicated as playing a role in MEC pathogenesis. As future genetic analyses of MEC tumors are published, it will be interesting to see if any of these genes disrupted by HPV integration are also altered by other genetic mechanisms, as the observation of multiple mechanisms of genetic disruption would support a critical role for the genes.

In conclusion, using our integrative sequencing analysis, we demonstrate that transcriptionally active HPV is a rare occurrence in MEC. However, when present, HPV can have a substantial role in altering the host genome, including through the direct integration and disruption of host genes. Our data suggest that alternative drivers other than HPV are much more frequently responsible for the pathogenesis of MEC, including *CRTC1/3-MAML2* fusion negative cases. From a long-term perspective, while the prevalence of HPV-related cancers is increasing around the world, our data suggests that HPV-related MECs are likely to account for only a minor subset of this disease. Our data is consistent with recently published series in concluding that no particular MEC subsite is more prone to harbor transcriptionally active HPV. Indeed, while our data resolve fundamental questions about the role of HPV in this disease, it also supports a need for future research to help identify additional genetic drivers of MEC. Due to the rarity of transcriptionally active HPV in MEC, we cannot advocate for routine p16 or HPV DNA ISH testing in clinical practice.

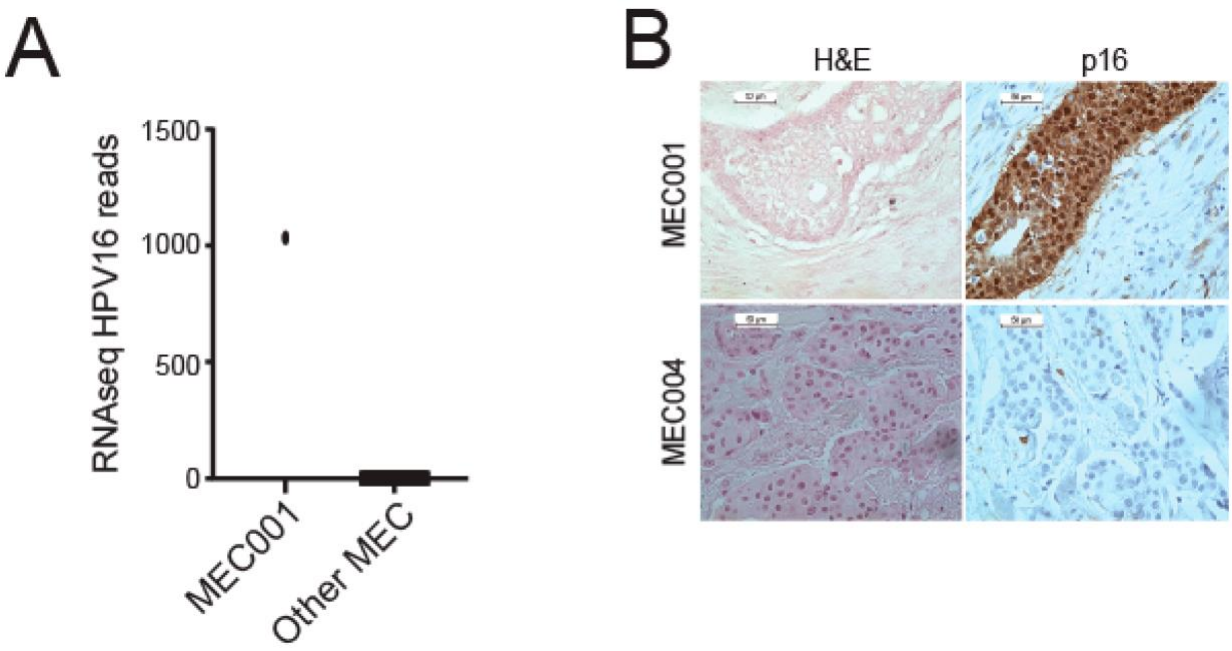


Figure 4.1 Analysis of HPV type distribution in our MEC cohort. A. Number of RNA-seq HPV16 reads for MEC1 and other MECs. B. p16 immunohistochemistry performed on sections from the HPV16+ tumor (MEC1) and a representative HPV negative tumor (MEC4).

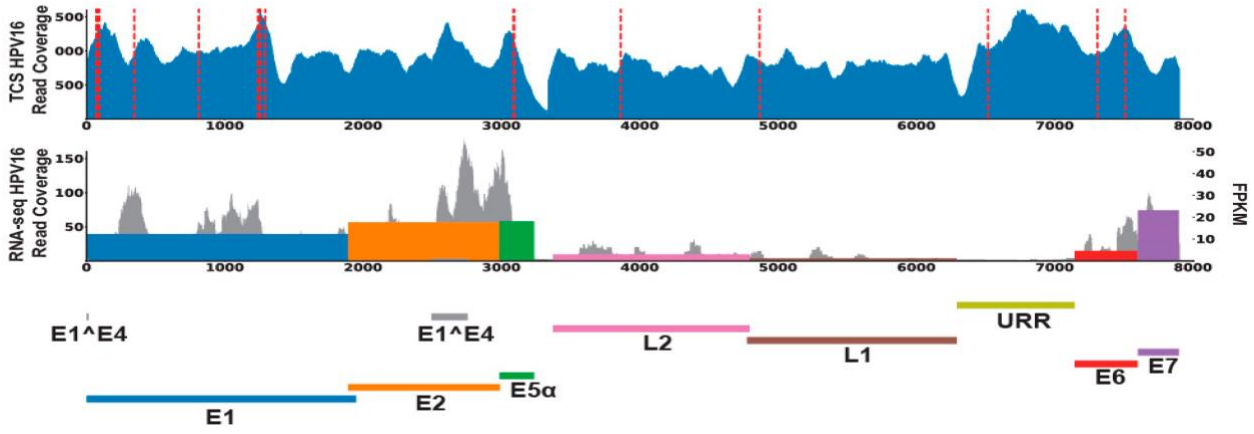


Figure 4.2 HPV16 DNA and RNA content in MEC1. The read coverages of HPV for MEC1 were filled in blue (first lane) for targeted capture sequencing data; grey for RNA-seq data (second lane). Red dashed lines denoted the HPV-host integrations called from targeted capture sequencing data. HPV genes expression levels (FPKM) were marked as colored bars in the second lane.

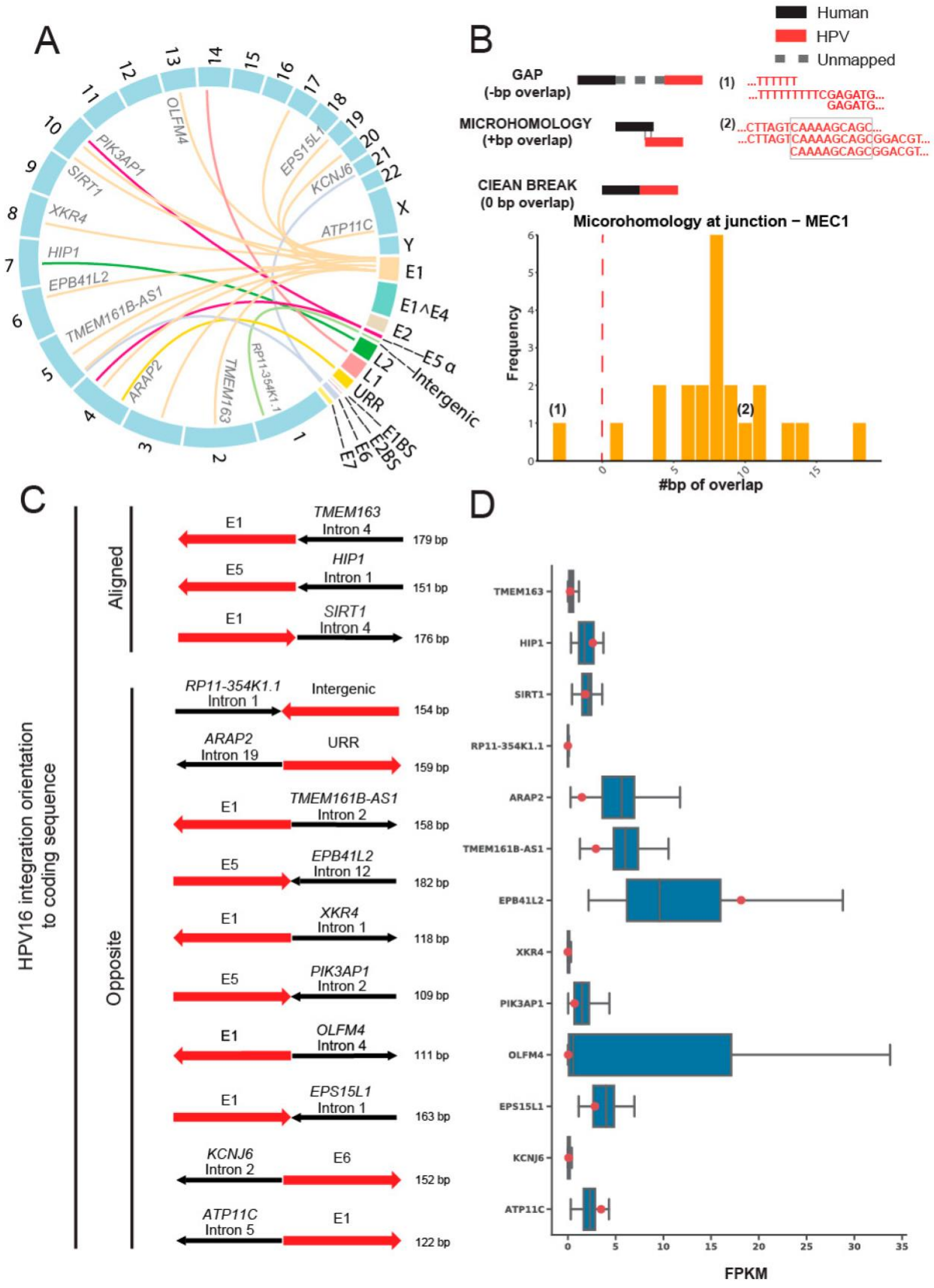

A
B
C
D
Human
HPV
Unmapped
GAP
(-bp overlap)
MICROHOMOLOGY
(+bp overlap)
CIEAN BREAK
(0 bp overlap)
Micorohomology at junction – MEC1
Frequency
#bp of overlap
HPV16 integration orientation
to coding sequence
Aligned
Opposite
FPKM

Figure 4.3 HPV16 integration site analysis in the host genome of MEC1. A. Link plot of the HPV-host integrations in MEC1. Lines were colored by HPV genes. Host genes that integrations fell into were marked. B. Microhomology at HPV-host junctions in MEC1. The microhomology was defined as the overlapped base pairs between human and HPV segments at the junctions. Overlapped bases referred to positive scores of microhomology, e.g., example (1); gaps referred to negative scores, e.g., example (2), and clean ends referred to zeros. C. HPV integration orientation to coding sequence. The human segments were colored in black and HPV segments were colored in red. Arrows of human segments indicated the human gene orientation: right, positive strands; left, negative strands. Arrows of HPV segments pointed the HPV gene orientation in contrast to the corresponding human gene. D. RNA expression levels of genes that HPV integrations fell into. Blue box plots denoted the FPKM of all 48 MECs for the 13 genes that HPV integrations fell into. Red points showed the FPKM for MEC1. Note that FPKM at RP11-354K1.1 were zero for all MECs.

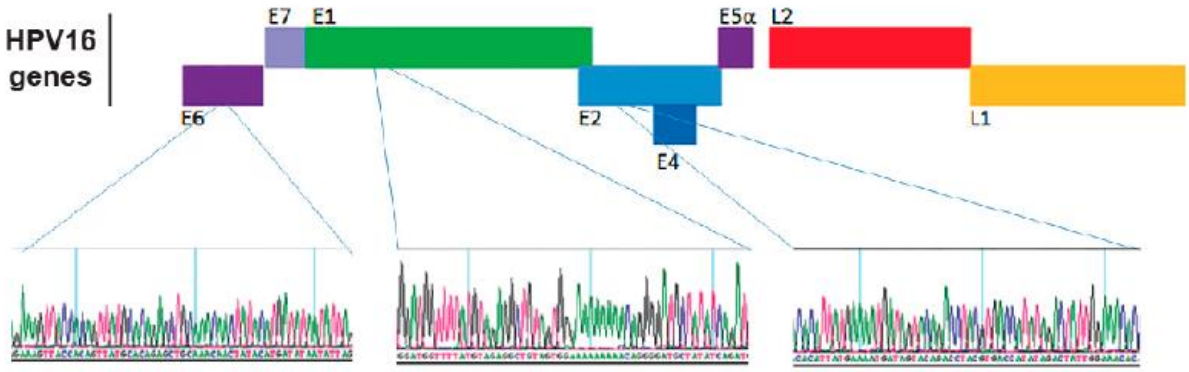


Figure 4.4 HPV16 PCR and Sanger sequencing analysis. We independently isolated genomic DNA from the FFPE block for MEC1 and performed PCR and Sanger sequencing on three independent regions of the HPV16 genome. Representative fragments of the Sanger traces validating the presence of HPV16 DNA are shown.

Bibliography

1. Dombrowski N.D., Wolter N.E., Irace A.L., Cunningham M.J., Mack J.W., Marcus K.J., Vargas S.O., Perez-Atayde A.R., Robson C.D., Rahbar R. Mucoepidermoid carcinoma of the head and neck in children. Int. J. Pediatr. Otorhinolaryngol. 2019;120:93–99. doi: 10.1016/j.ijporl.2019.02.020.

2. Granic M., Suton P., Mueller D., Cvrljevic I., Luksic I. Prognostic factors in head and neck mucoepidermoid carcinoma: Experience at a single institution based on 64 consecutive patients over a 28-year period. Int. J. Oral Maxillofac. Surg. 2018;47:283–288. doi: 10.1016/j.ijom.2017.09.005.

3. Nance M.A., Seethala R.R., Wang Y., Chiosea S., Myers E.N., Johnson J.T., Lai S.Y. Treatment and survival outcomes based on histologic grading in patients with head and neck mucoepidermoid carcinoma. Cancer. 2008;113:2082–2089. doi: 10.1002/cncr.23825.

4. McHugh C.H., Roberts D.B., El-Naggar A.K., Hanna E.Y., Garden A.S., Kies M.S., Weber R.S., Kupferman M.E. Prognostic factors in mucoepidermoid carcinoma of the salivary glands. Cancer. 2012;118:3928–3936. doi: 10.1002/cncr.26697.

5. Kang H., Tan M., Bishop J.A., Jones S., Sausen M., Ha P.K., Agrawal N. Whole-Exome Sequencing of Salivary Gland Mucoepidermoid Carcinoma. Clin. Cancer Res. 2017;23:283–288. doi: 10.1158/1078-0432.CCR-16-0720.

6. Okumura Y., Miyabe S., Nakayama T., Fujiyoshi Y., Hattori H., Shimozato K., Inagaki H. Impact of CRTC1/3-MAML2 fusions on histological classification and prognosis of mucoepidermoid carcinoma. Histopathology. 2011;59:90–97. doi: 10.1111/j.1365-2559.2011.03890.x.

7. Tonon G., Modi S., Wu L., Kubo A., Coxon A.B., Komiya T., O'Neil K., Stover K., El-Naggar A., Griffin J.D., et al. t(11;19)(q21;p13) translocation in mucoepidermoid carcinoma creates a novel fusion product that disrupts a Notch signaling pathway. Nat. Genet. 2003;33:208–213. doi: 10.1038/ng1083.

8. Yan K., Yesensky J., Hasina R., Agrawal N. Genomics of mucoepidermoid and adenoid cystic carcinomas. Laryngoscope. 2018;3:56–61. doi: 10.1002/lio2.139.

9. Nakayama T., Miyabe S., Okabe M., Sakuma H., Ijichi K., Hasegawa Y., Nagatsuka H., Shimozato K., Inagaki H. Clinicopathological significance of the CRTC3–MAML2 fusion transcript in mucoepidermoid carcinoma. Mod. Pathol. 2009;22:1575–1581. doi: 10.1038/modpathol.2009.126.

10. Rautava J., Kuuskoski J., Syrjänen K., Grenman R., Syrjänen S. HPV genotypes and their prognostic significance in head and neck squamous cell carcinomas. J. Clin. Virol. 2012;53:116–120. doi: 10.1016/j.jcv.2011.11.005.

11. Sedaghat A.R., Zhang Z., Begum S., Palermo R., Best S., Ulmer K.M., Levine M., Zinreich E., Messing B.P., Gold D., et al. Prognostic significance of human papillomavirus in oropharyngeal squamous cell carcinomas. Laryngoscope. 2009;119:1542–1549. doi: 10.1002/lary.20533.

12. Haring C.T., Bhambhani C., Brummel C., Jewell B., Bellile E., Neal M.E.H., Sandford E., Spengler R.M., Bhangale A., Spector M.E., et al. Human papilloma virus circulating tumor DNA assay predicts treatment response in recurrent/metastatic head and neck squamous cell carcinoma. Oncotarget. 2021;12:1214–1229. doi: 10.18632/oncotarget.27992.

13. Haring C.T., Brummel C., Bhambhani C., Jewell B., Neal M.H., Bhangale A., Casper K., Malloy K., McLean S., Shuman A., et al. Implementation of human papillomavirus circulating tumor DNA to identify recurrence during treatment de-escalation. Oral Oncol. 2021;121:105332. doi: 10.1016/j.oraloncology.2021.105332.

14. Chera B.S., Kumar S., Shen C., Amdur R., Dagan R., Green R., Goldman E., Weiss J., Grilley-Olson J., Patel S., et al. Plasma Circulating Tumor HPV DNA for the Surveillance of Cancer Recurrence in HPV-Associated Oropharyngeal Cancer. J. Clin. Oncol. 2020;38:1050–1058. doi: 10.1200/JCO.19.02444.

15. Brunner M., Koperek O., Wrba F., Erovic B.M., Heiduschka G., Schoppper C., Thurnher D. HPV infection and p16 expression in carcinomas of the minor salivary glands. Eur. Arch. Otorhinolaryngol. 2012;269:2265–2269. doi: 10.1007/s00405-011-1894-2.

16. Isayeva T., Said-Al-Naief N., Ren Z., Li R., Gnepp D., Brandwein-Gensler M. Salivary mucoepidermoid carcinoma: Demonstration of transcriptionally active human papillomavirus 16/18. Head Neck Pathol. 2013;7:135–148. doi: 10.1007/s12105-012-0411-2.

17. Bishop J.A., Yonescu R., Batista D., Yemelyanova A., Ha P.K., Westra W.H. Mucoepidermoid Carcinoma Does Not Harbor Transcriptionally Active High Risk Human Papillomavirus Even in the Absence of the MAML2 Translocation. Head Neck Pathol. 2014;8:298–302. doi: 10.1007/s12105-014-0541-9.

18. Birkeland A.C., Foltin S.K., Michmerhuizen N.L., Hoesli R.C., Rosko A.J., Byrd S., Yanik M., Nor J.E., Bradford C.R., Prince M.E., et al. Correlation of Crtc1/3-Maml2 fusion

status, grade and survival in mucoepidermoid carcinoma. Oral Oncol. 2017;68:5–8. doi: 10.1016/j.oraloncology.2017.02.025.

19. Heft Neal M.E., Gensterblum-Miller E., Bhangale A.D., Kulkarni A., Zhai J., Smith J., Brummel C., Foltin S.K., Thomas D., Jiang H., et al. Integrative sequencing discovers an ATF1-motif enriched molecular signature that differentiates hyalinizing clear cell carcinoma from mucoepidemoid carcinoma. Oral Oncol. 2021;117:105270. doi: 10.1016/j.oraloncology.2021.105270.

20. Hao Y., Yang L., Neto A.G., Amin M., Kelly D., Brown S., Branski R., Pei Z. HPViewer: Sensitive and specific genotyping of human papillomavirus in metagenomic DNA. Bioinformatics. 2018;34:1986–1995. doi: 10.1093/bioinformatics/bty037.

21. Merz L.E., Afriyie O., Jiagge E., Adjei E., Foltin S.K., Ludwig M.L., McHugh J.B., Brenner J.C., Merajver S.D. Clinical characteristics, HIV status, and molecular biomarkers in squamous cell carcinoma of the conjunctiva in Ghana. Health Sci. Rep. 2019;2:e108. doi: 10.1002/hsr2.108.

22. Cao Y., Haring C.T., Brummel C., Bhambhani C., Aryal M., Lee C., Neal M.H., Bhangale A., Gu W., Casper K., et al. Early HPV ctDNA Kinetics and Imaging Biomarkers Predict Therapeutic Response in p16+ Oropharyngeal Squamous Cell Carcinoma. Clin. Cancer Res. 2022;28:350–359. doi: 10.1158/1078-0432.CCR-21-2338.

23. Heft Neal M.E., Bhangale A.D., Birkeland A.C., McHugh J.B., Shuman A.G., Rosko A.J., Swiecicki P.L., Spector M.E., Brenner J.C. Prognostic significance of oxidation pathway mutations in recurrent laryngeal squamous cell carcinoma. Cancers. 2020;12:3081. doi: 10.3390/cancers12113081.

24. Bs L.M.P., Gu W., Wang Y., Bs A.E., Ms A.D.B., Ba C.V.B., Carey T.E., Mills R.E., Brenner J.C. SearcHPV: A novel approach to identify and assemble human papillomavirus–host genomic integration events in cancer. Cancer. 2021;127:3531–3540. doi: 10.1002/cncr.33691.

25. Rajaby R., Zhou Y., Meng Y., Zeng X., Li G., Wu P., Sung W.-K. SurVirus: A repeat-aware virus integration caller. Nucleic Acids Res. 2021;49:e33. doi: 10.1093/nar/gkaa1237.

26. Yan B., Liu X., Zhang S., Yu S., Tong F., Xie H., Song L., Zhang Y., Wei L. DisV-HPV16, versatile and powerful software to detect HPV in RNA sequencing data. BMC Infect. Dis. 2019;19:479. doi: 10.1186/s12879-019-4123-z.

27. Fakhry C., Lacchetti C., Rooper L.M., Jordan R.C., Rischin D., Sturgis E.M., Bell D., Lingen M.W., Harichand-Herdt S., Thibo J., et al. Human Papillomavirus Testing in Head and Neck Carcinomas: ASCO Clinical Practice Guideline Endorsement of the College of American Pathologists Guideline. J. Clin. Oncol. 2018;36:3152–3161. doi: 10.1200/JCO.18.00684.

28. Vageli D., Sourvinos G., Ioannou M., Koukoulis G.K., Spandidos D.A. High-risk human papillomavirus (HPV) in parotid lesions. Int. J. Biol. Mrk. 2007;22:239–244.

29. Jour G., West K., Ghali V., Shank D., Ephrem G., Wenig B.M. Differential Expression of p16INK4A and Cyclin D1 in Benign and Malignant Salivary Gland Tumors: A Study of 44 Cases. Head Neck Pathol. 2013;7:224–231. doi: 10.1007/s12105-012-0417-9.

30. Brandwein M.S., Ivanov K., Wallace D.I., Hille J.J., Wang B., Fahmy A., Bodian C., Urken M.L., Gnepp D.R., Huvos A., et al. Mucoepidermoid carcinoma: A clinicopathologic study of 80 patients with special reference to histological grading. Am. J. Surg. Pathol. 2001;25:834–845. doi: 10.1097/00000478-200107000-00001.

31. Wang K., McDermott J.D., Schrock A.B., Elvin J.A., Gay L., Karam S.D., Raben D., Somerset H., Ali S.M., Ross J.S., et al. Comprehensive genomic profiling of salivary mucoepidermoid carcinomas reveals frequent BAP1, PIK3CA, and other actionable genomic alterations. Ann. Oncol. 2017;28:748–753. doi: 10.1093/annonc/mdw689.

32. Li J., Zhang Z., Hu J., Wan X., Huang W., Zhang H., Jiang N. MiR-1246 regulates the PI3K/AKT signaling pathway by targeting PIK3AP1 and inhibits thyroid cancer cell proliferation and tumor growth. Mol. Cell. Biochem. 2022;477:649–661. doi: 10.1007/s11010-021-04290-3.

33. Zhang F., Li K., Yao X., Wang H., Li W., Wu J., Li M., Zhou R., Xu L., Zhao L. A miR-567-PIK3AP1-PI3K/AKT-c-Myc feedback loop regulates tumour growth and chemoresistance in gastric cancer. eBioMedicine. 2019;44:311–321. doi: 10.1016/j.ebiom.2019.05.003.

34. Yang H., Bi Y., Xue L., Wang J., Lu Y., Zhang Z., Chen X., Chu Y., Yang R., Wang R., et al. Multifaceted Modulation of SIRT1 in Cancer and Inflammation. Crit. Rev. Oncog. 2015;20:49–64. doi: 10.1615/CritRevOncog.2014012374.

35. Webber L.P., Yujra V.Q., Vargas P.A., Martins M.D., Squarize C.H., Castilho R.M. Interference with the bromodomain epigenome readers drives p21 expression and tumor senescence. Cancer Lett. 2019;461:10–20. doi: 10.1016/j.canlet.2019.06.019.

36. Escoubas C.C., Silva-García C.G., Mair W.B. Deregulation of CRTCs in Aging and Age-Related Disease Risk. Trends Genet. 2017;33:303–321. doi: 10.1016/j.tig.2017.03.002.

37. Brooks C.L., Gu W. How does SIRT1 affect metabolism, senescence and cancer? Nat. Rev. Cancer. 2009;9:123–128. doi: 10.1038/nrc2562.

Chapter 5 Conclusion

In this dissertation, we developed two innovative Bioinformatics approaches in the field of tumor virus integrations in viral associated cancers. In Chapter 2, I developed a novel bioinformatics pipeline named "SearcHPV," which demonstrated superior accuracy and efficiency compared to existing pipelines, particularly on TCS data. Notably, our software's capability for local contig assembly around integration junction sites simplifies downstream confirmation experiments. This approach was used in Chapter 3 on large sample size and in Chapter 4 in a rare type of cancer. In Chapter 3, we presented a novel strategy to unravel intricate HPV integrations characterized by extensive rearrangements in nanopore sequencing data. We illustrated its proficiency in producing more comprehensive and high-resolution structures when compared to established methodologies on one cell line. These two methods filled the technological gaps in tools for HPV integration research offering enhanced capabilities for generating comprehensive and high-resolution structures for the downstream analysis on the mechanism of HPV integration in cancers.

By applying these new tools in Chapter 2-4, I demonstrated that HPV integration sites have been linked to structural variations in the human genome, supporting an additional genetic mechanism contributing to the frequent detection of HPV integration sites adjacent to host cancer-related genes. These structural variation events are attributed to rolling circle amplification at the

integration breakpoint, resulting in the generation of amplified genomic segments flanked by HPV segments. Our findings align with previous reports [1–4]. We identified that approximately 10% of integration events discovered in our study were associated with large-scale amplifications (Type2) and 42% of patients had such complex HPV integration events. Although the reasons for the association of structural variants with only some integration sites remain unclear, we presented that Type2 HPV integrations were associated with a more complex genomic consequences pattern than Type1, which might indicate the possibility of alternative integration mechanisms. Notably, we identified several HPV-HPV junctions associated with a large duplication segment, suggesting the involvement of HPV internal rearrangement in HPV integration events in both Chapter 2 and Chapter 3.

The observed enrichment of HPV integration events in cellular genes may result from various mechanisms. Integration could preferentially occur in regions of open chromatin during cell replication and keratinocyte differentiation. Other potential mechanisms include directed integration to specific host genes by homology or random integration, with clonal selection and expansion of events advantageous for oncogenesis, implicating non-homology-based DNA repair mechanisms. To elucidate differences in integration mechanisms, we assessed microhomology at HPV-human junction points. The majority of breakpoints exhibited some level of microhomology, with the most frequent levels of overlap being 0 and 3 bp, suggesting the involvement of non-homologous end joining (NHEJ) in repair at these sites. Additionally, some junction sites showed a gap of inserted sequence between the HPV and human genomes, possibly indicative of polymerase theta-mediated end joining (TMEJ), which frequently involves the insertion of 3–30 bp at the repair site. Such mechanisms were both observed in my study in

Chapter2 (HNSCC) and Chapter4 (MEC). Future analysis using our pipeline is expected to provide further insights into the specific roles of different DNA repair pathways in HPV-human fusion breakpoints.

Significantly, we identified several integration hotspots in Chapter 2 and Chapter 3. Among them TP63, MYC, HEMGN, TRMO, TRAF2, CASC11, and KLF4 were reported in prior research [1,5–7,1,7–11]. Novel hotspots included EOMES, CASC11 and PVT1. These genes have been identified as biomarkers in cervical cancer, displaying irregular expression levels in HPV+ cancers or playing a role in viral gene lifecycle [12,13,14,15,16] , supporting their potential value for further analysis.

Interestingly, in Chapter3, we examined a subcohort of patients with recurrent nodes. Previous investigations have illuminated the intratumoral heterogeneity of HPV integration and potential clonal evolution through the analysis of primary tumors [3,17]. However, there is a dearth of literature detailing the heterogeneity of HPV integration in recurrent patients. Our examination revealed heterogeneity in integration structures, encompassing variations in local HPV copy numbers, distribution of specific HPV integration sites within structures, and distinctive characteristics of rearrangement breakpoints (human-HPV, HPV-HPV, human-human). Building upon these observations, we postulated that HPV integration events undergo clonal selection. We focused on one recurrent HPV+ model pair, UPCI:90 and UPCI:152, utilizing long-read whole-genome sequencing. Our findings furnish initial evidence indicating the heterogeneity and clonal selection of HPV integration events during pathogenesis, which might play a pivotal role in driving the disease process.

Several future directions were unveiled from my work and might be addressed in ongoing research. In a recent investigation of cervical cancer, the presence of multiple HPV integrations is associated with inferior survival outcomes compared to cases with single HPV integrations [17]. We intend to perform survival analyses on Type1 and Type2 events, exploring the potential of categorizing HPV integration types as prognostic biomarkers. Our fusion site pattern model in Chapter3 necessitates the comprehensive assembly of isoforms of chimeric transcripts to validate. Moreover, to gain a profound understanding of the mechanism of clonal selection in recurrent patients, a comparative analysis between preserved and lost integrations, along with an assessment of their impact on gene expression, holds the promise of providing insightful findings. Specifically, in Chapter 4, our transcriptional analysis of a single case of HPV-positive MEC suggested that the integration events showed no significant patterns on the impact of the expression of these genes. Thus, we expect that the causal mechanism by which HPV contributes to pathogenesis in the HPV+ tumor identified in our cohort is through elevated E6 and E7 oncogene expression leading to p16 overexpression. However, the exact impact of HPV integration on cellular genes in MEC needs to be further investigated with evidence from more samples.

In conclusion, my dissertation presented two HPV integration pipelines that overcome the technology challenges in the field of viral-host integration analysis in both short read and long read aspects. We first demonstrated the complex HPV integration events quantitatively and defined them as “Type2” events. Our integrative analysis using multi-omics data showed that Type2 events may drive distinct tumorigenic characteristics, and the heterogeneity of HPV integration could serve as an essential driver of tumor progression. We broadened our methods

on MEC and found one patient that had 13 genes inserted by HPV and explored HPV as a rare driver of MEC. We found that the genetic mechanisms of host genome integration are similar from MEC to HNSCC.

Bibliography

1. Symer DE, Akagi K, Geiger HM, et al. Diverse tumorigenic consequences of human papillomavirus integration in primary oropharyngeal cancers. Genome Res. 2022;32(1):55-70.

2. Hu Z, Zhu D, Wang W, 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):158-163.

3. Akagi K, Symer DE, Mahmoud M, et al. Intratumoral Heterogeneity and Clonal Evolution Induced by HPV Integration. Cancer Discov. 2023;13(4):910-927.

4. Akagi K, Li J, Broutian TR, et al. Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. Genome Res. 2014;24(2):185-199.

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

6. Rusan M, Li YY, Hammerman PS. Genomic landscape of human papillomavirus-associated cancers. Clin Cancer Res. 2015;21(9):2009-2019.

7. Pinatti LM, Gu W, Wang Y, et al. SearcHPV: A novel approach to identify and assemble human papillomavirus-host genomic integration events in cancer. Cancer. 2021;127(19):3531-3540.

8. Ragin CCR, Reshmi SC, Gollin SM. Mapping and analysis of HPV16 integration sites in a head and neck cancer cell line. Int J Cancer. 2004;110(5):701-709.

9. Rodriguez I, Rossi NM, Keskus A, et al. Insights into the Mechanisms and Structure of Breakage-Fusion-Bridge Cycles in Cervical Cancer using Long-Read Sequencing. medRxiv. Published online August 22, 2023. doi:10.1101/2023.08.21.23294276

10. Holzhauser S. Effect of Ionising Radiation on HPV-Positive and HPV-Negative Oropharyngeal Cancer Cell Lines. phd. Cardiff University; 2018. Accessed January 8, 2024. https://orca.cardiff.ac.uk/id/eprint/120510/

11. Mainguené J, Vacher S, Kamal M, et al. Human papilloma virus integration sites and genomic signatures in head and neck squamous cell carcinoma. Mol Oncol. 2022;16(16):3001-3016.

12. Baedyananda F, Chaiwongkot A, Varadarajan S, Bhattarakosol P. HPV16 E1 dysregulated cellular genes involved in cell proliferation and host DNA damage: A possible role in cervical carcinogenesis. PLoS One. 2021;16(12):e0260841.

13. Tosi A, Parisatto B, Menegaldo A, et al. The immune microenvironment of HPV-positive and HPV-negative oropharyngeal squamous cell carcinoma: a multiparametric quantitative and spatial analysis unveils a rationale to target treatment-naïve tumors with immune checkpoint inhibitors. J Exp Clin Cancer Res. 2022;41(1):279.

14. Hsu W, Liu L, Chen X, Zhang Y, Zhu W. LncRNA CASC11 promotes the cervical cancer progression by activating Wnt/beta-catenin signaling pathway. Biol Res. 2019;52(1):33.

15. Wang X, Wang G, Zhang L, Cong J, Hou J, Liu C. LncRNA PVT1 promotes the growth of HPV positive and negative cervical squamous cell carcinoma by inhibiting TGF-β1. Cancer Cell Int. 2018;18:70.

16. Moody C. Mechanisms by which HPV Induces a Replication Competent Environment in Differentiating Keratinocytes. Viruses. 2017;9(9). doi:10.3390/v9090261

17. Zhou L, Qiu Q, Zhou Q, et al. Long-read sequencing unveils high-resolution HPV integration and its oncogenic progression in cervical cancer. Nat Commun. 2022;13(1):