scientific reports

Check for updates

OPEN

Case control study comparing the HPV genome in patients with oral cavity squamous cell carcinoma to normal patients using metagenomic shotgun sequencing

Ian Ganly[1]✉, Zhiheng Pei[2,3,8], Yuhan Hao[2,4,5], Yingfei Ma[3,6], Matthew Rosenthal[1], Zhenglin Wu[7,9], Jocelyn Migliacci[1], Bin Huang[7,10], Nora Katabi[11], Wenzhi Tseng[2], Stuart Brown[4], Yi-Wei Tang[7,12] & Liying Yang[2,3]✉

The aim of this study was to carry out a case control study comparing the HPV genome in patients with oral cavity squamous cell carcinoma (OC-SCC) to normal patients using metagenomic shotgun sequencing. We recruited 50 OC-SCC cases which were then matched with a control patient by age, gender, race, smoking status and alcohol status. DNA was extracted from oral wash samples from all patients and whole genome shotgun sequencing performed. The raw sequence data was cleaned, reads aligned with the human genome (GRCH38), nonhuman reads identified and then HPV genotypes identified using HPViewer. In the 50 patients with OC-SCC, the most common subsite was tongue in 26 (52%). All patients were treated with primary resection and neck dissection. All but 2 tumors were negative on p16 immunohistochemistry. There were no statistically significant differences between the cases and controls in terms of gender, age, race/ethnicity, alcohol drinking, and cigarette smoking. There was no statistically significant difference between the cancer samples and control samples in the nonhuman DNA reads (medians 4,228,072 vs. 5,719,715, P value = 0.324). HPV was detected in 5 cases (10%) of OC-SCC (genotypes 10, 16, 98) but only 1 tumor sample (genotype 16) yielded a high number of reads to suggest a role in the etiology of OC-SCC. HPV was detected in 4 control patients (genotypes 16, 22, 76, 200) but all had only 1–2 HPV reads per human genome. Genotypes of HPV are rarely found in patients with oral cancer.

The American Cancer Society estimates about 53,260 people will be diagnosed and 10,750 people will die of oral cancer and oropharyngeal cancer in the United States in 2020[1,2]. More than 95% of these cases will be squamous cell carcinoma (SCC)[1,2]. Although approximately 50% of these patients might be alive 5 years from now, about 20% will die from the disease within 1 year[2]. The 5-year survival of oral cancer has not significantly improved over the past several decades[3]. Although tobacco use is among the strongest risk factors for oral cancer, this factor does not completely explain the incidence of all oral cancer as the drastic decline in the prevalence of cigarette

[1]Head and Neck Service, Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, USA. [2]Department of Pathology, New York University School of Medicine, New York 10016, USA. [3]Department of Medicine, New York University School of Medicine, New York 10016, USA. [4]Applied Bioinformatics Laboratories, New York University School of Medicine, New York 10016, USA. [5]Center for Genomics and Systems Biology, Department of Biology, New York University, New York, NY 10016, USA. [6]Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518000, China. [7]Department of Laboratory Medicine, Memorial Sloan Kettering Cancer Center, New York, USA. [8]Department of Veterans Affairs, New York Harbor Healthcare System, New York, USA. [9]Department of Laboratory Medicine, The Eighth Affiliated Hospital of Sun Yat-Sen University, Shenzhen, China. [10]Department of Laboratory Medicine, The First Affiliated Hospital of Sun Yat-Sen University, Guangzhou, China. [11]Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, USA. [12]Medical Affairs, Cepheid, Danaher Diagnostic Platform, Shanghai, China. ✉email: ganlyi@mskcc.org; liying.yang@nyumc.org

smoking since 1975 (from ~ 40 to 20%) has caused only a moderate change in the incidence of oral cancer[4,5]. This indicates a paradigm shift in the cause of oral cancer and the need to search for other risk factors.

High risk genotypes of Human papilloma virus (HPV), genotypes 16, 18 and 33, now account for 60–80% of all oropharyngeal squamous cell cancers[6–11]. This has led to the hypothesis that some genotypes of HPV may also be responsible for the epidemiology change in the etiology of oral cancer as well. Although the prevalence of the high risk genotypes HPV 16, 18 and 33 varies greatly across multiple studies on oral cancer[12] it is now accepted that these high risk genotypes are unlikely to be responsible. With regards to other genotypes of HPV, over the past 10 years an increasing number of HPV types have being found in oral samples[13,14]. The currently available HPV detection kits only detect a limited number of HPV genotypes. Since there are over 200 different genotypes of HPV, we hypothesized these may be responsible for some cases of oral cancer. The aim of our study was therefore to carry out a case control study in 50 oral cancer patients and 50 matched control patients to detect all 200 genotypes of HPV using next generation sequencing.

Methods

To examine the hypothesis that oral HPV is associated with OC-SCC, we performed a case control study with mouthwash samples from 50 patients with OC-SCC and 50 subjects with no oral lesions. Total genomic DNA was extracted from cell pellets of the mouthwash samples. Subject recruitment, sample collection, data generation and analysis are detailed below.

Recruitment of human subjects for oral cavity squamous cell carcinoma cases and matched controls. A case–control study was approved by the Institutional Review Board of Memorial Sloan Kettering Cancer Center (IRB 15-256) and New York University School of Medicine (i15-00389). All methods were performed in accordance with the relevant guidelines and regulations. Informed consent was obtained from all patients in the study population. From Memorial Sloan Kettering Cancer Center (MSKCC), we recruited 50 oral cavity squamous cell carcinoma (OC-SCC) cases. These cases were then matched with a control patient by age, gender, race, smoking status and alcohol status. Overall, 100 subjects (50 cases and 50 controls) were enrolled in this study. The controls comprised patients with thyroid nodules (benign or malignant). These patients had complete head and neck examination including flexible laryngoscopy and were found to have no evidence or oral cancer. In patients with oral cancer, OC-SCC was confirmed by histological examination of biopsy specimens. Pathological grade and stage of OC-SCC were determined by histopathological examination at the time of surgical resection. Demographic and clinical information was collected for each patient.

Detection of HPV in OC-SCC tumor samples. High risk (HR) HPV infection was determined by Tissue HR HPV PCR and p16 immunohistochemistry on tumor tissue of OC-SCC patients. Ang et al. has reported that the expression of p16INK4a by immunohistochemistry correlated well (kappa = 0.80; 95% CI 0.73–0.87) with the presence of HPV DNA in tumors[15]. This is cheaper and easier to carry out than ISH and PCR and therefore immunostaining of tumor sections for p16INK4a is now used as an indirect marker for HPV status in clinical pathology laboratories around the world[16]. In prospective randomized trials on treatment of patients with HPV related oropharyngeal cancer, p16 immunohistochemistry is now used as the surrogate marker for HPV positivity in the USA. Rarely some p16 positive tumors may not be HPV related. The addition of HPV PCR to the detection methodology would increase specificity as described by Prigge et al.[17]. In our study, all pathology specimens were examined by a single pathologist specialized in head and neck pathology (NK). p16 immunohistochemistry was performed as follows: four-micrometer tumor sections were deparaffinized, and after heat-induced epitope retrieval, immunohistochemistry for p16INK4a was performed with the primary antibody dilution of 1:7 as per manufacturer's protocol (CINtec Histology Kit, catalog #9517, Roche mtm Laboratories AG, Heidelberg, Germany). Cases with nuclear and cytoplasmic immunolabeling in at least 70% of the tumor cells were considered positive for p16. In patients with available tissue, HR HPV PCR was done to confirm results of the p16 immunohistochemistry.

Detection of all 200 genotypes of HPV in mouthwash samples of OC-SCC patients and control patients using metagenomic shotgun sequencing (MSS). The detection of HPV in oral rinse samples and saliva samples using PCR has been reported as being a sensitive and specific method for the detection of HPV related oropharyngeal cancer[18–25]. This technique has been reported as a potential screening method for HPV oropharynx cancer[18–22] and also for the detection of persistent disease or recurrent disease following treatment in patients with HPV related oropharyngeal cancer[23–25]. We therefore used oral rinse specimens from patients with oral cavity cancer and patients with no head and neck cancer for the detection of HPV genotypes. The workflow for the collection and processing of samples and detection of HPV by DNA sequencing is shown in Fig. 1 and detailed as follows:

Mouthwash sample collection, processing, storage and DNA extraction. The participants rinsed their mouth vigorously with 10 ml sterile saline for 30 s and then mouthwash collected in a 50 cc falcon canical flask container. After centrifugation at 3120×*g* for 20 min, supernatants were decanted and then the cell pellets were transferred into a 2-ml Eppendorf tube and stored at − 80 °C freezer for further study. All oral rinse specimens were taken prior to surgical resection of the OC-SCC. These samples were de-identified and coded. Using the MoBio method, we successfully extracted DNA from all 100 oral wash samples. DNA yield was measured by the Nanodrop method.

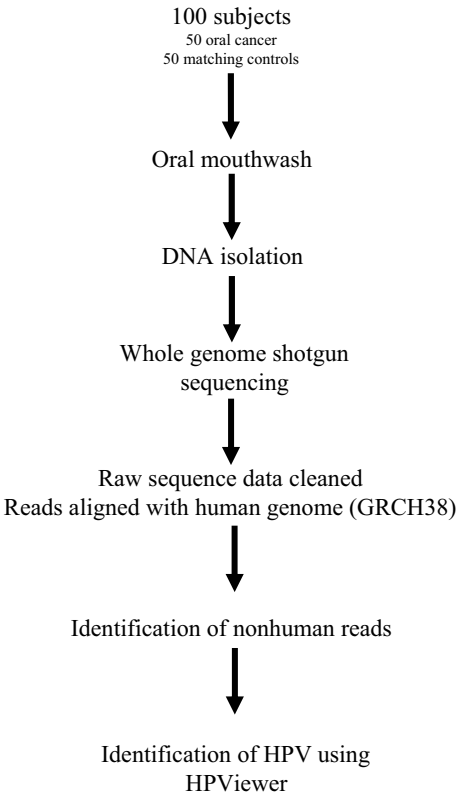

Figure 1. The workflow for the collection and processing of samples and detection of HPV by DNA sequencing.

Library preparation and samples sequencing. The DNA fragmentation and shotgun metagenomic library construction and sequencing was carried out at the BGI Americas Corp (Cambridge, MA) using Kapa kit and Illumina HISeq X Ten, with 100 samples pooled into 8 lanes.

Raw sequence data quality control. The Illumina sequencing process supplied raw sequence reads in fastq format and assigned a quality score called 'Phred scores' to describe the base accuracy. The raw sequence quality was reviewed using FASTQC software and any adapters and low-quality reads were removed using Trimmomatic. Low quality reads were defined as leading low quality or N bases, quality < 25; trailing low quality or N bases, quality < 25, scanning the read with a 4-base wide sliding window and cutting when the average quality per base drops below 25 and drops reads below 50 bases long. On average, each sample yielded 70,984,784 ± 7,145,632 raw reads, ranging 54,314,162 to 88,936,078. After trimming, 43,400,676 ± 5,586,559 reads left, ranging 32,127,260 to 58,483,416. The average percentage of clean to raw reads was 61.16% (Minimum: 47.12%; Median: 62.24%; Maximum: 72.69%) (Supplementary Fig. 1).

Identification of non-human DNA reads. All trimmed reads were aligned to the human genome (GRCh38). Those with ≥ 90% similarity were considered as human reads. On average, each sample yielded 36,184,844 ± 7,255,847 human reads, ranging 18,446,834 to 56,005,990. The sequencing depth was estimated by averaging the coverage across all 22 human autosomal chromosomes. There was no statistically significant difference between the cancer samples and control samples in the sequencing depth (medians 33,760,382 vs. 34,046,615, P value = 0.77, Mann Whitney test) (Supplementary Fig. 2).

Detecting and genotyping HPV in mouthwash samples obtained from patients with OC-SCC and healthy controls. We compared HPV prevalence and abundance between the 50 patients with OC-SCC and 50 subjects with no oral lesions. The nonhuman DNA reads were searched for HPV DNA using HPViewer (Supplementary Fig. 3). We have developed a pipeline to identify HPV reads generated by MSS based on the sequence similarities to the genome sequences of HPV prototypes and used it in a survey of HPV in healthy subjects[26]. Since then, we have made several improvements to allow more accurate detection and classification of HPV reads in human samples in a new software program HPViewer[27]. Briefly, we found that HPV shares not only massive amount of homologous sequences among different HPV types but also extensive simple repeats with human and some bacteria. The inter-type homologous sequences cause errors in HPV genotyping and the shared repeats between human and bacteria can be mistaken as HPV DNA. In HPViewer, these shared regions in the reference HPV genomes are masked to minimize these errors. We also replaced BLAST in the old pipeline with Bowtie2, an

Characteristics	OC-SCC (n = 50)	Normal controls	P value
Sex (%)			
Women	23 (46%)	23 (46%)	1[a]
Men	27 (54%)	27 (54%)	
Age (mean ± SD)	62.1 ± 11.9	61.6 ± 11.7	0.827[b]
Race (%)			
White	44 (88%)	44 (88%)	0.765[a]
Others	6 (12.0%)	6 (12%)	
Alcohol drinking (%)			
Never/quit	18 (36%)	17(34%)	0.364[a]
Moderate	27 (54%)	23 (46%)	
Heavy	5 (10%)	10 (20%)	
Smoking (%)			
Never	22 (44%)	24 (48%)	0.67032[a]
Quit	20 (40%)	21 (42%)	
Active	8 (16%)	5 (10%)	

Table 1. Patient characteristics of cases and controls. [a]Chi-square test. [b]*T* test, two tailed.

ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. We did not have a positive control group in the study. However, in the study to develop the software program HPViewer[27], we evaluated the metagenomic methods and HPViewer with six tumor tissue specimens and matched oral washes from patients with recurrent respiratory papillomatosis (RRP). HPViewer detected HPV reads in all tumor samples [288–4229 HPV (type 6 or 11) reads per tumor sample] while in the six matched oral wash samples, only two were positive for the concordant HPV types with only 2 HPV reads/sample. Because the 2 HPV reads in the oral wash samples shared the same SNPs with the HPV reads in the corresponding tumor samples, it was concluded that the oral HPV reads represent the HPV released from the tumors in the larynx. In various oral cavity sites in healthy people, HPV prevalence ranged 2.9 to 7.1% and type 16 occasionally detected[27]. Thus, the low level HPV reads detected in the oral washes samples most likely represent HPV released from the tumors or from infection site in the oral cavity.

Results

Patient characteristics of cases and controls. Patient demographics are shown in Table 1. There were no statistically significant differences between the cases and controls in terms of gender, age, race/ethnicity, alcohol drinking, and cigarette smoking. In the 50 patients with OC-SCC, the subsite was tongue in 26 (52%), floor of mouth in 8 (16%), lower gum in 7 (14%), upper gum in 4 (8%) (Table 2). All patients were treated with primary resection and neck dissection.

Tumor characteristics of OC-SCC. Pathology details are shown in Table 2. Of the 50 OC-SCC patients, 37 (74%) had pathological T1T2 tumor and 21 (42%) had a pathological positive neck. The majority of primary tumors were either well (20%) or moderately differentiated (64%). All but 3 tumors were negative on p16 immunohistochemistry. Of 26 samples tested by HPV PCR, 25 were negative for high risk HPV and only 1 was positive. The positive HPV PCR case was also positive on p16 imunohistochemistry.

Metagenomic HPV sequencing results of mouthwash samples of OC-SCC and control patients. The DNA yields ranged 63 ng/μl (range 11.1–127.8) for the cases and 54.3 ng/μl (range 12.6–106.5) for the controls.

Detection of nonhuman DNA reads. Nonhuman reads averaged 7,215,832 ± 6,165,570, ranging 601,022 to 27,624,778. On average, nonhuman reads accounted for 10.06 ± 8.36% of the total reads, ranging 1.01% to 37.37% (Fig. 2). There was no statistically significant difference between the cancer samples and control samples in the nonhuman DNA reads (medians 4,228,072 vs. 5,719,715, P value = 0.324, Mann Whitney test (Fig. 3).

Detecting and genotyping HPV in mouthwash samples obtained from patients with OC-SCC and controls. Using HPViewer, HPV was detected in five cases (10%) of OC-SCC and four controls (8%) (Table 3). The raw data for HPV reads is accessible at http://www.ncbi.nlm.nih.gov/bioproject/692713 using the BioProject ID PRJNA692713. In the 5 OC-SCC cases, only 1 tumor sample (sample 90) yielded a considerable number of HPV reads suggesting a role in the etiology of OC-SCC in this patient. This was genotype 16. This patient was also positive on tissue HPV by PCR and positive p16 immunohistochemistry. The location of the tumor was the anterior 2/3rds of the oral tongue. The other 4 tumor cases yielded only about 1–2 HPV reads per human genome. These were in serotypes HPV 10, 16 (2 cases) and 98. All 4 patients were negative on p16 tissue immunohistochemistry and tissue HPV-PCR. Of the 4 control patients, all had only 1–2 HPV reads per human genome. These were in genotypes HPV 16, 22, 76, and 200.

Characteristic	No (%)
Tumor subsite	
Tongue	26 (52)
Floor of mouth	8 (16)
Upper gum	4 (8)
Lower gum	7 (14)
Buccal	2 (4)
Retromolar trigone	2 (4)
Lip	1 (2)
Tumor size (mm)	
1–10	11 (22)
11–20	15 (30)
21–30	12 (24)
31–40	8 (16)
41–50	4 (8)
Pathology T stage	
T1	24 (48)
T2	14 (28)
T3	4 (8)
T4	8 (16)
Pathology N stage	
N0/Nx	29 (58)
N +	21 (42)
Tumor grade	
Well differentiated	10 (20)
Moderately differentiated	32 (64)
Poorly differentiated	8 (16)
Tissue p16 immunostain	
Positive	2 (4)
Negative	47 (94)
Not done	1 (2)
Tissue HR HPV PCR	
Positive	1 (2)
Negative	25 (50)
Not done	24 (48)

Table 2. Tumor characteristics of oral cancer patients.

Discussion

Over the past 20 years the epidemiology of oral cancer has been changing. Traditionally, oral cancer has been caused by smoking and heavy alcohol consumption[3]. There has been a steady decline in the use of cigarettes and alcohol in the population[4]. Despite this, the incidence of oral cancer has failed to decline. New studies show that there is an increasing number of patients who do not smoke or drink alcohol excessively but still develop oral cancer[5]. These patients tend to be younger with an increased frequency in females. The cause of oral cancer in these patients remains an enigma.

This change in epidemiology has occurred over the same time period as the change in epidemiology of oropharyngeal cancer. In oropharyngeal cancer it is now recognized that the agent causing cancer is high risk genotypes of the Human Papilloma virus, notably HPV 16, 18, and 33[12].

This has led several studies to be carried out on oral cancer patients to identify high risk HPV genotypes, either in saliva or in tissue tumor specimens. The results of these studies have been highly variable with some studies showing little association whereas others have reported a strong association[12]. However, it is now generally accepted that the high risk genotypes HPV16, 18, and 33 are unlikely to be responsible for the change in epidemiology or oral cancer. There are 200 different genotypes of HPV. The traditional HPV detection kits/methods cover only a limited number of high/low risk (HLR) HPV genotypes[28–32]. Most of these traditional HPV detection methods are PCR-based and detect 14 genotypes (HPV 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66, and 68). Since these detection kits do not cover all 200 different genotypes of HPV, it is possible that these detection methods may be failing to identify other genotypes of HPV which may be responsible for causing oral cancer.

The limited ability of current commercial HPV detection kits can be overcome by metagenomic shotgun sequencing (MSS). This is a non-selective approach that, in theory, permits the identification of all HPV sequences. Recently, MSS has been used to detect HPV in some human samples and to identify several novel

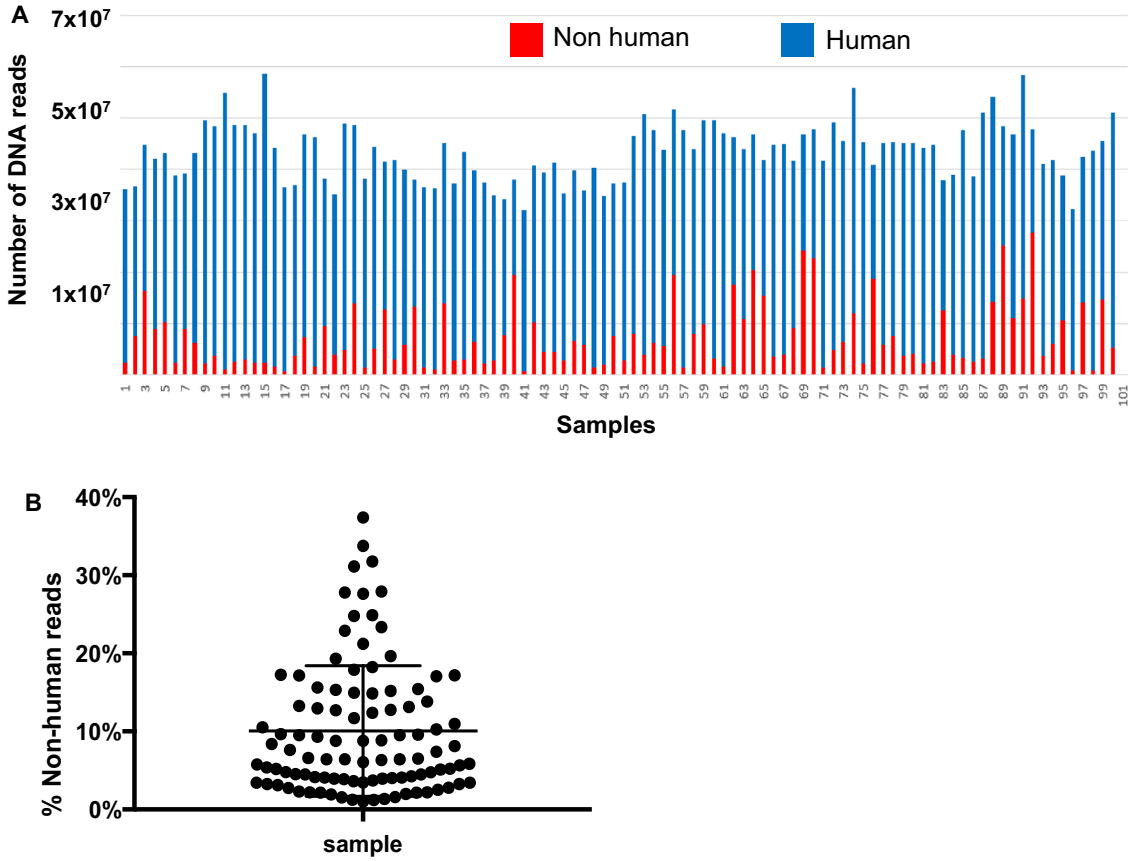

Figure 2. Distribution of nonhuman DNA reads. Number (**A**) and percentage (**B**) of nonhuman reads were calculated along with human reads and total reads. The Y axes show the percentage of nonhuman reads over the total reads. The X axes show sample distribution within the range. Mean reads ± SD is shown in Panel (**B**).

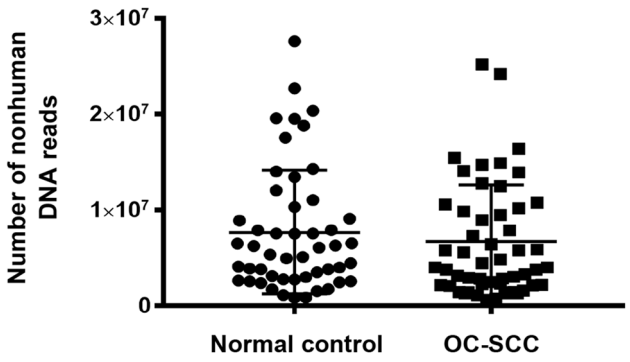

Figure 3. Comparison of OC-SCC and normal controls in nonhuman DNA reads.

HPV types[33–38]. In particular, a study that involved condyloma samples shown to be negative for HPV by traditional PCR revealed the ability of MSS to identify many putative HPV sequences[33]. Using MSS, we surveyed HPV distribution in various body sites of 103 healthy human subjects and found that the majority of the 109 HPV types detected could not be detected using the widely used commercial kits and do not belong to the HLR HPV types[26]. Interestingly, the HPV types detected have strong organ tropism and the oral HPV community is different from that of the vagina. These findings raised the possibility that the oral HPV types that are invisible to the traditional detection methods contribute to the etiology of human diseases, such as OC-SCC. The aim of our study was to carry out a case control study in 50 oral cancer patients and 50 matched control patients to detect all 200 genotypes of HPV by MSS.

In our study our cancer patients and control patients were well matched in terms of age, gender, and smoking and alcohol status. We used oral rinse samples from each patient and extracted DNA from cell pellets prior to sequencing. DNA extracted from oral rinse samples has been reported to be a sensitive and specific method for

Case	Phenotype	Tumor location	Raw HPV counts	Normalized HPV counts[a]	HPV type
T90	OC-SCC	Anterior 2/3 of tongue	42	52.93	16
T98	OC-SCC	Retromolar trigone extending to buccal mucosa and lower gingiva	2	1.34	16
T101	OC-SCC	Buccal mucosa	2	1.4	16
T78	OC-SCC	Lower gingiva	2	1.49	98
G31	OC-SCC	Lateral border of tongue	1	0.63	10
T71	Control	NA	2	1.52	22
T109	Control	NA	2	2.51	200
T116	Control	NA	2	1.39	76
T123	Control	NA	2	1.37	16

Table 3. HPV genotypes detected by MSS in oral cancer and control patients. [a]Number of HPV reads per human genome.

the identification of patients with oropharyngeal cancer[18–22] as well as the detection of persistent or recurrent cancer in patients who have completed treatment[23–25]. Our sequencing methodology using MSS was also sound with comparable reads in both tumor and control samples. Through our comprehensive detailed sequencing analysis we have shown that only 1 patient had a tumor sample (2%) that could be directly related to HPV. In this case, it was HPV16 which is the main genotype responsible for oropharyngeal cancer. Importantly, no other genotype of HPV was able to be detected at a high copy number indicating that these other genotypes are not responsible for oral cancer in these patients. These observations are highly relevant because they now provide evidence that HPV is rarely associated with oral cancer. There are 2 other studies which support our findings. A study by Bragelman using RNASeq to sequence the tumor mRNA in 7 patients with oral tongue cancer who did not smoke or drink alcohol failed to show any HPV sequences[39]. Another study by Li et al. examined the tumor transcriptomes of 20 patients with oral tongue cancer and failed to identify any HPV viral transcriptome[40]. A recent metanalysis by Sahovaler reported that in oral cavity locations, overall survival was not significantly associated with HPV positivity (hazard ratio [HR], 1.16; 95% CI 0.83–1.61; I2 = 71%)[41].

There is much interest in identifying factors responsible for oral cancer in these patients. It is possible other viruses such as herpes simplex, herpes zoster, Epstein Barr virus may be responsible though research on these common viruses have not shown any association to date[40]. Even more research is ongoing to identify bacteria which may be responsible. Studies on the oral microbiome have recently been published suggesting specific bacteria may be responsible[19,42,43]. Our own group recently identified that the periodontal pathogens *Fusobacterium, Prevotella, Alloprevotella* were enriched while commensal *Streptococcus* depleted in OC-SCC in nonsmoking patients with premalignant oral cavity lesions as well as oral cancer[19]. Clearly this is an area which requires much research effort to try to provide new insight into this devastating disease.

In conclusion, our study suggests no role in the aetiology of HPV in oral cavity cancer. However, our population of patients is fairly homogenous population (88% white ethnicity) and all from the USA. It is possible that there may be geographic or ethnic differences in the role of HPV across populations and we therefore cannot extrapolate from a single 50 patient study". Further research in different geographic and ethnic populations is needed. New research is also needed to explore other infectious agents such as bacteria or viruses that may be responsible for oral cancer. Although we saw few HPV reads in our metagenomic data, it is important to analyse the metagenomic sequences for other non human reads from other viruses, bacteria or even fungi. It is possible these other microbes may have an association with OSCC pathogenesis. To carry out such a comprehensive analysis requires complex bioinformatics as well as validation studies. These studies are currently underway.

Data availability

The raw data for HPV reads is accessible at http://www.ncbi.nlm.nih.gov/bioproject/692713 using the BioProject ID PRJNA692713. The accession codes for positive reads for HPV are given in Supplementary Table 1.

Received: 21 June 2020; Accepted: 27 January 2021
Published online: 16 February 2021

References

1. American Cancer Society. *Facts & Figures 2020* (American Cancer Society, Atlanta, 2020).
2. https://www.cancer.org/cancer/oral-cavity-and-oropharyngeal-cancer.html.
3. Neville, B. W. & Day, T. A. Oral cancer and precancerous lesions. *CA Cancer J. Clin.* **52**, 195–215 (2002).
4. http://www.cancercenter.com/oral-cancer/risk-factors/.
5. http://seer.cancer.gov/statfacts/html/oralcav.html, http://seer.cancer.gov/statfacts/html/oralcav.html.
6. Gillison, M. L. Human papillomavirus-related diseases: Oropharynx cancers and potential implications for adolescent HPV vaccination. *J. Adolesc. Health* **43**, S52-60 (2008).
7. Isayeva, T., Li, Y., Maswahu, D. & Brandwein-Gensler, M. Human papillomavirus in non-oropharyngeal head and neck cancers: A systematic literature review. *Head Neck Pathol.* **6**(Suppl 1), S104–S120 (2012).
8. Parkin, D. M. The global health burden of infection-associated cancers in the year 2002. *International J. Cancer* **118**, 3030–3044 (2006).

9. Munoz, N., Castellsague, X., de Gonzalez, A. B. & Gissmann, L. Chapter 1: HPV in the etiology of human cancer. *Vaccine* **24**(Suppl 3), 1–10 (2006).
10. Parkin, D. M. & Bray, F. Chapter 2: The burden of HPV-related cancers. *Vaccine* **24**(Suppl 3), 11–25 (2006).
11. Tang, K. W., Alaei-Mahabadi, B., Samuelsson, T., Lindh, M. & Larsson, E. The landscape of viral expression and host gene fusion and adaptation in human cancer. *Nat. Commun.* **4**, 2513 (2013).
12. Mehanna, H. *et al.* Prevalence of human papillomavirus in oropharyngeal and nonoropharyngeal head and neck cancer–systematic review and meta-analysis of trends by time and region. *Head Neck* **35**, 747–755 (2013).
13. Paolini, F. *et al.* Both mucosal and cutaneous papillomaviruses are in the oral cavity but only alpha genus seems to be associated with cancer. *J. Clin. Virol.* **56**, 72–76 (2013).
14. Lang Kuhs, K. A. *et al.* Prevalence of and risk factors for oral human papillomavirus among young women in Costa Rica. *J. Infect. Dis.* **208**, 1643–1652 (2013).
15. Ang, K. K. *et al.* Human papillomavirus and survival of patients with oropharyngeal cancer. *N. Engl. J. Med.* **363**, 24–35 (2010).
16. Lewis, J. S. Jr. p16 Immunohistochemistry as a standalone test for risk stratification in oropharyngeal squamous cell carcinoma. *Head Neck Pathol.* **6**(Suppl 1), S75-82 (2012).
17. Prigge, E., Arbyn, M., Doeberitz, M. & Reuschenbach, M. Diagnostic accuracy of p16INK4a immunohistochemistry in oropharyngeal squamous cell carcinomas: A systematic review and meta-analysis. *Int. J. Cancer* **140**, 1186–1198 (2017).
18. Yoshida, H. *et al.* Usefulness of human papillomavirus detection in oral rinse as a biomarker of oropharyngeal cancer. *Acta Otolaryngol.* https://doi.org/10.1080/00016489.2016.1274426 (2017).
19. Ganly, I. *et al.* Periodontal pathogens are a risk factor of oral cavity squamous cell carcinoma, independent of tobacco and alcohol and human papillomavirus. *Int. J. Cancer* **145**(3), 775–784. https://doi.org/10.1002/ijc.32152 (2019).
20. Tang, K. D. *et al.* Oral HPV16 prevalence in oral potentially malignant disorders and oral cavity cancers. *Biomolecules* **10**(2), E223. https://doi.org/10.3390/biom10020223 (2020).
21. Tang, K. D., Kenny, L., Frazer, I. H. & Punyadeera, C. High-risk human papillomavirus detection in oropharyngeal cancers: Comparison of saliva sampling methods. *Head Neck* **41**(5), 1484–1489. https://doi.org/10.1002/hed.25578 (2019).
22. Tang, K. D. *et al.* Unlocking the potential of saliva-based test to detect HPV-16-driven oropharyngeal cancer. *Cancers* **11**(4), E473. https://doi.org/10.3390/cancers11040473 (2019).
23. Rettig, E. M. *et al.* Prognostic implication of persistent human papillomavirus type 16 DNA detection in oral rinses for human papillomavirus-related oropharyngeal carcinoma. *JAMA Oncol.* **1**, 907–915. https://doi.org/10.1001/jamaoncol.2015.2524 (2015).
24. Agrawal, Y. *et al.* Oral human papillomavirus infection before and after treatment for human papillomavirus 16-positive and human papillomavirus 16-negative head and neck squamous cell carcinoma. *Clin. Cancer Res.* **14**, 7143–7150. https://doi.org/10.1158/1078-0432.CCR-08-0498 (2008).
25. Hanna, G. J. *et al.* Salivary HPV DNA informs locoregional disease status in advanced HPV-associated oropharyngeal cancer. *Oral Oncol.* **95**, 120–126. https://doi.org/10.1016/j.oraloncology.2019.06.019 (2019).
26. Ma, Y. Human papillomavirus community in healthy persons, defined by metagenomics analysis of human microbiome project shotgun sequencing data sets. *J. Virol.* **88**, 4786–4797 (2014).
27. Hao, Y. HPViewer: Sensitive and specific genotyping of human papillomavirus in metagenomic DNA. *Bioinformatics* **34**, 1986–1995 (2018).
28. Brink, A. A., Snijders, P. J. & Meijer, C. J. HPV detection methods. *Dis. Markers* **23**, 273–281 (2007).
29. Chen, W. *et al.* Evaluation of cobas 4800 high-risk HPV test as a tool in cervical cancer screening and cytology triage. *Zhonghua Zhong Liu Za Zhi* **34**, 543–548 (2012).
30. Sorbye, S. W., Arbyn, M., Fismen, S., Gutteberg, T. J. & Mortensen, E. S. HPV E6/E7 mRNA testing is more specific than cytology in post-colposcopy follow-up of women with negative cervical biopsy. *PLoS ONE* **6**, e26022 (2011).
31. Lee, G. Y. *et al.* Human papillomavirus (HPV) genotyping by HPV DNA chip in cervical cancer and precancerous lesions. *Int. J. Gynecol. Cancer* **15**, 81–87 (2005).
32. Abreu, A. L., Souza, R. P., Gimenes, F. & Consolaro, M. E. A review of methods for detect human Papillomavirus infection. *Virol J.* **9**, 262 (2012).
33. Johansson, H. *et al.* Metagenomic sequencing of "HPV-negative" condylomas detects novel putative HPV types. *Virology* **440**, 1–7 (2013).
34. Mokili, J. L. *et al.* Identification of a novel human papillomavirus by metagenomic analysis of samples from patients with febrile respiratory illness. *PLoS ONE* **8**, e58404 (2013).
35. Kocjan, B. J., Steyer, A., Sagadin, M., Hosnjak, L. & Poljak, M. Novel human papillomavirus type 174 from a cutaneous squamous cell carcinoma. *Genome Announc.* **1**, 1 (2013).
36. Foulongne, V. *et al.* Human skin microbiota: High diversity of DNA viruses identified on the human skin by high throughput sequencing. *PLoS ONE* **7**, e38499 (2012).
37. Li, L. *et al.* Identification of a novel human gammapapillomavirus species. *J. Gen. Virol.* **90**, 2413–2417 (2009).
38. Ekstrom, J., Bzhalava, D., Svenback, D., Forslund, O. & Dillner, J. High throughput sequencing reveals diversity of human papillomaviruses in cutaneous lesions. *Int. J. Cancer* **129**, 2643–2650 (2011).
39. Brägelmann, J. *et al.* Oral cavity tumors in younger patients show a poor prognosis and do not contain viral RNA. *Oral Oncol.* **49**(6), 525–533 (2013).
40. Li, R. *et al.* Clinical, genomic, and metagenomic characterization of oral tongue squamous cell carcinoma in patients who do not smoke. *Head Neck.* **37**(11), 1642–1649 (2015).
41. Sahovaler, A. *et al.* Survival outcomes in human papillomavirus-associated nonoropharyngeal squamous cell carcinomas: A systematic review and meta-analysis. *JAMA Otolaryngol. Head Neck Surg.* https://doi.org/10.1001/jamaoto.2020.3382 (2020).
42. Lee, W. H. *et al.* Bacterial alterations in salivary microbiota and their association in oral cancer. *Sci. Rep.* **7**, 16540. https://doi.org/10.1038/s41598-017-16418-x (2017).
43. Hsiao, J. R. *et al.* The interplay between oral microbiome, lifestyle factors and genetic polymorphisms in the risk of oral squamous cell carcinoma. *Carcinogenesis* **39**, 778–787. https://doi.org/10.1093/carcin/bgy053 (2018).

Author contributions

I.G., Z.P. and L.Y. wrote the manuscript text and created the figures. Y.H., Y.M., B.H., W.T., S.B. carried out the bioinformatics analyses. M.R., I.G., J.M. collected all samples. J.M. carried out statistical analyses. N.K. analysed all pathology of tumor samples and did all immunohistochemistry. Z.W. carried out DNA extraction of samples. Y.-W.T. reviewed and edited the manuscript. All authors reviewed and edited the manuscript.

Funding

This work was supported in part by grants from the National Institute of Dental and Craniofacial Research (R21DE025352 to ZP, LY, IG) and NIH/NCI Cancer Center Support Grant P30 CA008748 (MSKCC). Z.P. is staff physician at the Department of Veterans Affairs New York Harbor Healthcare System. The content is the

sole responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, the U.S. Department of Veterans Affairs or the United States Government.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-83197-x.

Correspondence and requests for materials should be addressed to I.G. or L.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.