Molecular Oncology

FEBS PRESS science publishing by scientists

Detection of human papillomavirus in cases of head and neck squamous cell carcinoma by RNA-seq and VirTect

Atlas Khan[1], Qian Liu[2], Xuelian Chen[3,4], Andres Stucky[3,4], Parish P. Sedghizadeh[3,4], Daniel Adelpour[3,4], Xi Zhang[3,4,5], Kai Wang[2] and Jiang F. Zhong[3,4]

1 Division of Nephrology, Department of Medicine, College of Physicians and Surgeons, Columbia University, New York, NY, USA
2 Raymond G. Perelman Center for Cellular and Molecular Therapeutics, The Children's Hospital of Philadelphia, Philadelphia, PA, USA
3 Division of Periodontology, Diagnostic Sciences and Dental Hygiene, Herman Ostrow School of Dentistry, University of Southern California, Los Angeles, CA, USA
4 Division of Biomedical Sciences, Herman Ostrow School of Dentistry, University of Southern California, Los Angeles, CA, USA
5 Department of Hematology, Xinqiao Hospital, Army Medical University, Chongqing, China

Keywords
carcinogenesis; head and neck squamous cell carcinoma; human papillomavirus; RNA-seq; viral detection; VirTect

Correspondence
X. Zhang and J. F. Zhong, Division of Periodontology, Diagnostic Sciences and Dental Hygiene, Herman Ostrow School of Dentistry, University of Southern California, Los Angeles, CA 90089, USA
Tel: (213) 740-0085
E-mails: zhangxxi@sina.com (XZ); jzhong@usc.edu (JFZ)
and
K. Wang, Raymond G. Perelman Center for Cellular and Molecular Therapeutics, The Children's Hospital of Philadelphia, 3501 Civic Center Boulevard, 5th Floor CTRB, Philadelphia, PA 19104, USA
Tel: (267) 425-9573
E-mail: wangk@email.chop.edu (KW)

Atlas Khan and Qian Liu are the co-first authors

(Received 26 September 2018, revised 13 December 2018, accepted 20 December 2018, available online 23 February 2019)

doi:10.1002/1878-0261.12435

Next-generation sequencing provides an opportunity to detect viral species from RNA-seq data of human tissues, but existing computational approaches do not perform optimally on clinical samples. We developed a bioinformatic method called VirTect for detecting viruses in neoplastic human tissues using RNA-seq data. Here, we used VirTect to analyze RNA-seq data from 363 head and neck squamous cell carcinoma (HNSCC) patients and identified 22 human papillomavirus (HPV)-induced HNSCCs. These predictions were validated by manual review of pathology reports on histopathologic specimens. VirTect showed better performance in recall and accuracy compared to the two existing prediction methods, VirusFinder and VirusSeq, in identifying viral sequences from RNA-seq data. The majority of HPV carcinogenesis studies thus far have been performed on cervical cancer and generalized to HNSCC. Our results suggest that carcinogenesis of HPV-induced HNSCC and other cases of HNSCC involve different genes, so understanding the underlying molecular mechanisms will have a significant impact on therapeutic approaches and outcomes. In summary, RNA-seq together with VirTect can be an effective solution for the detection of viruses from tumor samples and can facilitate the clinicopathologic characterization of various types of cancers with broad applications for oncology.

Abbreviations
DE, differential expression; HNSCC, head and neck squamous cell carcinoma; HPV, human papillomavirus; NGS, next-generation sequencing; Rb, retinoblastoma; TCGA, The Cancer Genome Atlas.

1. Introduction

The advent of next-generation sequencing (NGS) provides an opportunity to accurately and comprehensively detect and identify viral pathogens in clinical samples (Chiu, 2013; Dunne *et al.*, 2012). Researchers have developed some computational tools for the discovery and identification of viruses in NGS data from human tissues (Chen *et al.*, 2013; Isakov *et al.*, 2011; Kostic *et al.*, 2011) with subtractive analyses including the detection of viral integration sites (Wang *et al.*, 2015). For example, in a study conducted by Isakov *et al.* (2011), three steps were utilized for virus identification: (I) align short reads against a human reference genome using TopHat (Trapnell *et al.*, 2009), (II) subtract nonhuman sequences from human sequences, and then (III) categorize viruses based on nucleic acid databases. Similarly, Kostic *et al.* (2011) considered the reads which were not mapped to the human genome after subtractive filtration to be 'candidate nonhuman pathogen-derived reads'. Another set of researchers also used a subtractive approach with a threshold of the minimum number of reads, which were mapped to virus genomes in their defined virus database (Chen *et al.*, 2013). Using the aforementioned methods, it is possible to encounter some noise from sequencing/alignment or poly(A) sequences which may also have high coverage and a high number of reads aligned to viral sequences; however, they do not actually represent a viral sequence. Thus, to address this issue, there is an urgent need to develop more robust computational tools which can correctly detect and identify viruses from human tissues, and discriminate viral sequences from sequencing artifacts or sequencing/alignment noise including insignificantly aligned reads or poly(A) sequences.

Here, we describe a new method called VirTect for detecting viruses in RNA-seq data from human clinical samples. Our approach to virus detection uses two filters to discriminate real viral sequences from noise and artifacts, thereby reducing false-positive rates: a threshold for the coverage of mapped reads for virus genomes to reduce the effect of insignificant alignment, and a threshold of the length of continuous mapped regions to capture significant expressed transcripts for any pathogen genome in our defined virus database. After these filtrations, it is expected that a nonhuman sequence (which is not able to be mapped to a human reference genome) would very likely be a viral sequence, and based on these nonhuman sequences, VirTect could improve the accuracy of detecting pathogenic viruses from RNA-seq data.

The utility of VirTect then is demonstrated by performing detailed analysis of RNA-seq data from cancer patients. With the awareness of chemical carcinogens such as tobacco, certain types of cancer have steadily decreased. However, cancers induced by viral infection have increased significantly (Relman, 1999). Recent studies indicate that approximately 12% of human cancers are viral in etiology (Zur Hausen, 2009). For example, human papillomavirus (HPV) comprises a group of more than 200 related viruses, and some HPV subtypes were found to be associated with specific types of cancers such as cervical, vaginal, vulvar, anal, penile, and head and neck squamous cell carcinomas (HNSCC) (Arbyn *et al.*, 2011; Bosch *et al.*, 1995; Leyden *et al.*, 2005). The oncogenic HPV subtypes (e.g., subtypes 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, and 68) are sexually transmitted viruses (Weinstock *et al.*, 2004). Importantly, the proportion of HNSCCs that are HPV-positive increased from 18% in 1973 to 32% in 2005 in the United States, representing an unprecedented and dramatic epidemiologic spike (Taberna *et al.*, 2017). However, the mechanism of carcinogenesis in HPV-induced HNSCC is not well studied compared to that of HPV-induced cervical cancers. In the present work, we performed detailed analysis of RNA-seq data from 363 HNSCC patients using VirTect and identified 20 HPV16-positive and 2 HPV33-positive patients. The viral and host (patient) gene expression profiles of the 20 HPV16-positive RNA-seq data further suggest that the E2 and E7 genes of HPV, but not E6, play a major role in HPV-induced HNSCC. The molecular mechanism of HPV oncogenesis in HNSCC differs from what has been reported in cervical carcinogenesis (Leung *et al.*, 2017). The viral/host gene expression profiles of HPV-induced HNSCC elucidate HPV oncogenesis in HNSCC. In doing so, this new bioinformatic tool provides actionable targets for developing new diagnostic assays and therapeutics for HNSCC. VirTect is available at https://github.com/WGLab/VirTect.

2. Materials and methods

2.1. Datasets

We downloaded 363 HNSCC samples from the Cancer Genome Atlas (TCGA) (https://cancergenome.nih.gov/) and used them to evaluate and test VirTect.

2.2. Quality control and trimming

We performed quality control (QC) using FASTQ (Andrews, 2010) and then used CUTADAPT (Martin, 2011) to trim paired-end reads of all samples to ensure that the quality was appropriate for further investigation. We did the following filtrations using CUTADAPT version 1.2.1:

I Removed adaptor contamination, which may affect virus detection.

II Trimmed the reads when the average quality score in a sliding window fell below a phred score of 20.
III Discarded reads shorter than 40 bp.

The CUTADAPT command line was as follows:
cutadapt -a adapter1 -A adapter2 -m 40 -q 20,20 –trim-n -o sample_1_trimmed.fastq.gz -p sample_2_trimmed.fastq.gz sample_1.fastq.gz sample_2.fastq.gz

where *adapter1* is a forward adapter, *adapter2* is a reverse adapter, *-q* is for quality, and *-m* is used for the minimum length of reads (i.e., reads < 40 bp long were trimmed in our analysis). After trimming of all samples, we used VirTect for virus detection.

2.3. Virus detection method (VirTect)

VirTect contains several steps, as illustrated in Fig. 1:

Step 1: Paired-end (PE) reads were aligned to a human genome reference to subtract the nonhuman sequences from the whole-genome sequence of human samples using TopHat (Trapnell *et al.*, 2009).

Step 2: In the second step, the bam file of the nonhuman reads was converted to FASTQ format using BEDTools (Quinlan and Hall, 2010).

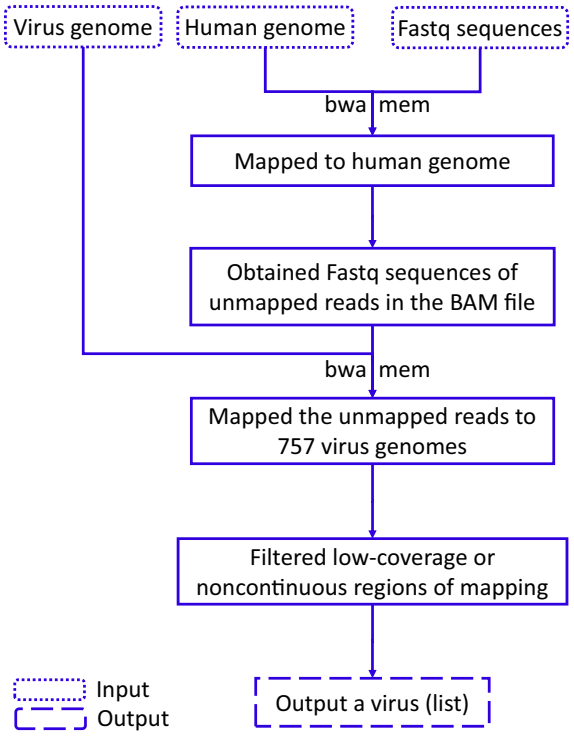

Fig. 1. Schematic overview of the VirTect pipeline for virus detection from NGS data.

Step 3: The nonhuman sequences were mapped to the 757 different viruses currently in the database using BWA-MEM (Li and Durbin, 2009).

Step 4: Filtrations were performed to remove noise/artifacts or poly(A) sequences, which may have high coverage with thousands of reads mapped to virus genomes but may not represent real viral sequences. Two types of filtrations are used for this purpose:

I Cutoff of the coverage of nonhuman sequence where reads were aligned to virus genome but not to human genome. A default setting is 5×; that is, the coverage will be > 5×.
II Cutoff for continuously mapped regions. Our threshold is 100 by default but this is a user-predefined parameter. Since we used 48-bp sequences, this worked well for the data analyzed here; however, when reads are longer (more than 100 bp or larger), this threshold should be larger to fit users' data.

Step 5: Finally, a list of viruses were generated from the samples that passed the filtrations, and this list can then be subjected to further investigation.

2.4. Availability of data and materials

Our implementation of VirTect is available as a software tool at https://github.com/WGLab/VirTect.

VirTect was implemented in Python programming language and has been tested on Linux platforms. It depends on third-party publicly available tools, including SAMtools (Li *et al.*, 2009), BEDTools, Bowtie 2, TopHat, and BWA.

3. Results

An overview of VirTect is shown in Fig. 1. NGS data in FASTQ format provided the input, which was then mapped to a human reference genome using TopHat (Trapnell *et al.*, 2009). After the subtraction of human sequences, nonhuman sequences were aligned using BWA-MEM against the nonhuman sequences in our defined virus database (currently 757 different viruses) to identify viruses. Finally, VirTect performed filtrations to discriminate true viral sequences from noise or artifacts and reported identified virus(es).

3.1. Detection of HPV in HNSCC patients with VirTect

To test the accuracy of VirTect, we analyzed 363 HNSCC samples acquired from the TCGA database and detected HPV transcripts in 22 cases, of which 20

contained HPV16 transcripts and 2 had HPV33 transcripts. We examined pathology reports as a validation of our predictions. Histopathology and clinical assays agreed with our RNA-seq analysis results (Table 1). The histopathology of H&E slides from each HPV+ case was rigorously examined by a pathologist to confirm the morphology of viral infection. Histopathology findings (detailed in Fig. 2) confirmed morphologic features consistent with HPV infection and confirmed the HNSCC phenotype in each case. A random subset of 30 negative cases were also examined, and the results suggested that samples without detectable HPV genes were indeed HPV− pathologically.

To compare VirTect with existing tools, we ran VirTect, VirusSeq, and VirusFinder on a RNA-seq data set, in which pathology reports are available to indicate whether a patient has HPV16 infection. We used the metrics of precision (i.e., predicted HPV+ cases confirmed with pathology/total predicted HPV+ cases), recall (i.e., predicted HPV+ cases confirmed with pathology/all pathological HPV+ cases), and accuracy (i.e., predicted cases confirmed with pathology/total cases) of virus detection for comparison (Table 2). The results demonstrated that all the three tools had a precision of 1.00. However, VirusSeq had much lower recall (0.087); VirusFinder had a recall of 0.913. That is, VirusSeq and VirusFinder could not identify all samples where HPV16 virus was found in the pathology reports. In contrast, VirTect demonstrated 100% consistency with the pathology reports. We recognize that this is a relatively small sample size due to limited availability of pathology reports; nevertheless, this analysis confirmed that VirTect compares favorably to competing approaches on clinical samples.

HPV16 transcripts encoding key viral oncoproteins (i.e., E7, E4, E5_a, and E1^E4) were detected in most of the HPV-positive samples. The HPV gene expression in an HPV16+ case, as visualized in IGV, is shown in Fig. 3A. In this sample, the HPV16 oncoprotein E7 was expressed with high coverage and continuously

Table 1. Clinical confirmation of HPV16 infection in HNSCC cases.

Patient number (TCGA case number)	Sample location	Pathological evidence	P16 assay confirmation	Morphology confirmation	Sex
1 (abbcc1d8-ab74-4459-bf28-cc627bef440e)	Oropharynx/tonsil	Oropharynx/tonsil	P16 positive (cytoplasm and nuclei)	Yes	Female
2 (e37f479f-ca05-45e4-ac6d-a974bed6e7f8)	Left tonsil	Left tonsil	NA	Yes	Male
3 (5ac57aee-4be1-4a29-a53f-343f5a3d2e86)	Pharynx, oropharynx	Pharynx, oropharynx	NA	Yes	Male
4 (cac30b32-03ef-4ecb-88d8-d5577ab6b2a6)	Right base of tongue	Right base of tongue	NA	Yes	Male
5 (59a7d695-cb0c-47b8-ba9d-ca52ca2cfa7d)	Right tonsil	Right tonsil	P16 positive	Yes	Male
6 (cd032ddb-55f2-4c77-8dcf-e4e630f7de6f)	Right tonsil	Right tonsil	NA	Yes	Male
7 (05f01280-bf77-4682-a7a8-20dd0eac77bd)	Left tonsil and soft palate	Left tonsil and soft palate	NA	Yes	Male
8 (c7df3466-b9a7-4818-883b-d0cd08483570)	Right tonsil	Right tonsil	NA	Yes	Female
9 (4bfbce2b-9d0b-4e8a-950f-fd8e0ba3e05a)	Right base of the tongue	Right base of tongue	NA	Yes	Female
10 (387db1df-ebaa-41c2-b036-f46ad61e313a)	Base of the tongue	Base of the tongue	NA	Yes	Male
11 (9d469689-0413-4898-aa83-c6756cbfe117)	Right soft palate	Right soft palate	P 16/18 *in situ* hybridization positive	Yes	Male
12 (54b295d4-6315-4416-bde7-a221859f965c)	Left glossotonsillar sulcus extending into lingual tonsil	Left glossotonsillar sulcus extending into lingual tonsil	NA	Yes	Male
13 (b3631718-9e0a-454c-bee1-8f36ebc509d8)	Left tonsil		NA	Yes	Male
14 (27c28c89-f4e7-4aec-a806-c0da7756e47f)	Right floor of mouth	Right floor of mouth	NA	Yes	Male
15 (03c3ae62-d0aa-412e-bd3c-4577fc9f919c)	Right tonsil	Right tonsil	Positive for p16 and HPV16.	Yes	Male
16 (9f89510c-ed07-471f-b35e-7c87c237b9fe)	Left tonsil	Left tonsil	NA	Yes	Male
17 (602f2512-00b6-44b6-9ed6-f8b010224f8c)	Pharynx, oropharynx	Pharynx, oropharynx	P16 IHC positive, P16 ISH positive	Yes	Male
18 (180036a2-3b56-405e-a1fe-d5932517b6c7)	Right tonsil	Right tonsil	NA	Yes	Male
19 (10f53522-9ae7-47b9-80aa-b4b481561465)	Right tonsil	Right tonsil	NA	Yes	Male
20 (a0b136fb-3a0a-4411-8907-4ca775c7d04e)	Left tonsil	Left tonsil	P16 staining positive	Yes	Male

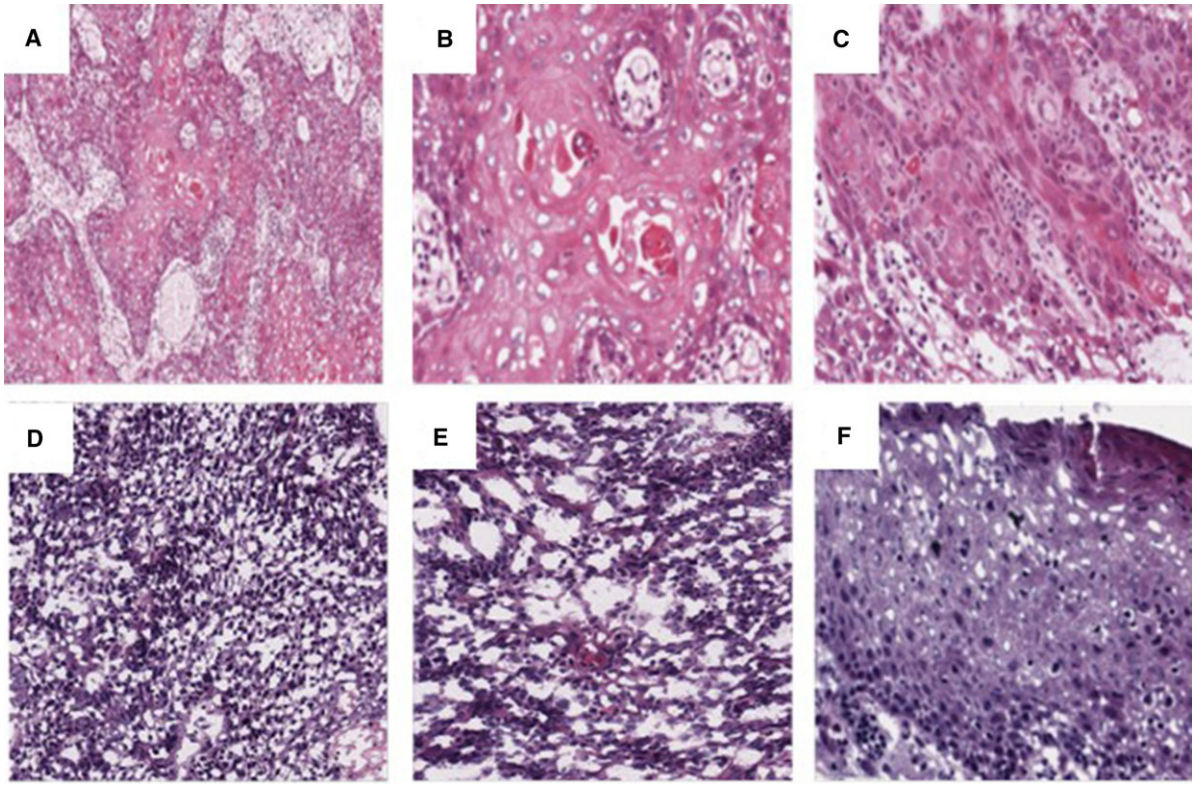

Fig. 2. Histopathology confirms *in silico* results of HPV infection. (A–C) Histopathology of cancer specimens without *in silico* HPV detection. (A) Moderately differentiated invasive squamous cell carcinoma (20×). (B) Higher magnification showing invasive tumor islands with central keratin pearl formation and individual cell dyskeratosis (40×). (C) Cellular atypia, pleomorphism, nucleoli, and mitotic activity are observed regularly (40×). (D–F) Histopathology of *in silico* HPV-positive specimens which have characteristic nonkeratinizing tumor morphology. (D) A typical HPV16-positive specimen (20×) detected by VirTect and confirmed by serological test. (E) Higher magnification (40×) shows infiltrating tumor islands lacking squamous maturation and comprising a cell population of ovoid to spindle-shaped cells with indistinct borders or gap junctions, and hyperchromatic nuclei that lack prominent nucleoli. (F) Overlying mucosa and significant koilocytosis indicating HPV infection. Viral cytopathic effect can be seen in the form of keratinocyte nuclear enlargement and hyperchromasia with perinuclear clearing or halo.

mapped regions. A HNSCC sample with detectable HPV33 viral genes is shown in Fig. 3B. The key viral oncoprotein E7 was also expressed in this sample. The HPV+ samples we identified were all oropharyngeal in origin, which further supports the link between HPV infection and oncogenesis since nearly all oral HPV+ cases are oropharyngeal in origin anatomically. Pathology reports in some of these cases included *in situ* hybridization for HPV high-risk subtypes which indicated HPV 16 infection, providing further molecular support for viral etiology.

Although pathology analysis is the gold standard for HPV infection, it is too labor-intensive for screening a large number of patients. With VirTect, we were able to screen 363 patients for HPV infection within a week. Among the 363 HNSCC patients, HPV 16 viral genes were detected in 20 patients (3 female and 17 male; see Table 1). Although the sample size is small, it appears that in this database, HPV-induced HNSCC was not equally distributed between males and females. Also, as expected, the L1 and L2 viral genes were not detected in any patient, and E2 and E8^E2 were detected at high levels in all HPV16+ patients.

3.2. HPV gene expression profiles in HNSCC

In cervical carcinoma, integration of the HPV16 viral genome into the host genome often leads to the disruption of the E1 and E2 open reading frame, resulting in unregulated expression of E6 and E7 (Schwarz *et al.*, 1985; Smotkin and Wettstein, 1986). However, this does not appear to be the case in HNSCC (Fig. 4). Among the 20 patients with detectable HPV16 viral sequences, the number of detectable viral genes ranged from three genes (patients 9, 11, and 15) to 10 genes (patient 19). E2 and E7 were the most common, detected in 85% (17/20) and 90% (18/20) of patients, respectively. Most patients (15/20) had both

Table 2. Comparison of virus detection results among VirTect, VirusSeq, and VirusFinder.

Method	Precision	Recall	Accuracy
VirTect	1.00	1.00	1.00
VirusSeq	1.00	0.087	0.344
VirusFinder	1.00	0.913	0.938

There were 32 RNA-seq data for analysis including 21 HPV16+ RNA-seq data and 11 HPV16− RNA-seq data. Precision is the number of HPV16+ RNA-seq data where HPV16 was detected by a tool divided by the number of RNA-seq data where HPV16 was detected, recall is the number of HPV16+ RNA-seq data where HPV16 was detected by a tool divided by the number of HPV16+ RNA-seq data, while accuracy is the number of correctly identified RNA-seq data, no matter HPV16+ or HPV16−, divided by the total number of RNA-seq data.

detectable E2 and E7. The alternatively spliced E8^E2 was detected in 90% (18/20) of patients, and the two patients (patients 9 and 15) without E8^E2 were among those without detectable E2; these patients only had three detectable viral genes (E1, E1^E4, and E7). E1^E4 and E4 were detected in 95% (19/20) and 85% (17/20) of patients, respectively, in agreement with a previous report where E1^E4 was found to be the most abundantly expressed HPV protein in infected epithelial cells (Raj *et al.*, 2004). E6 (sequence range 83–559) was only detected in 30% (6/20) of patients. Fragments of this gene, namely E6*(sequence range 83–226) and E6*# (sequence range 409–417), were detected in 25% (5/20) and 15% (3/20) of patients, respectively. Patients 16 and 17 had detectable E6 and E6*, while all E6 fragments were detected in patient

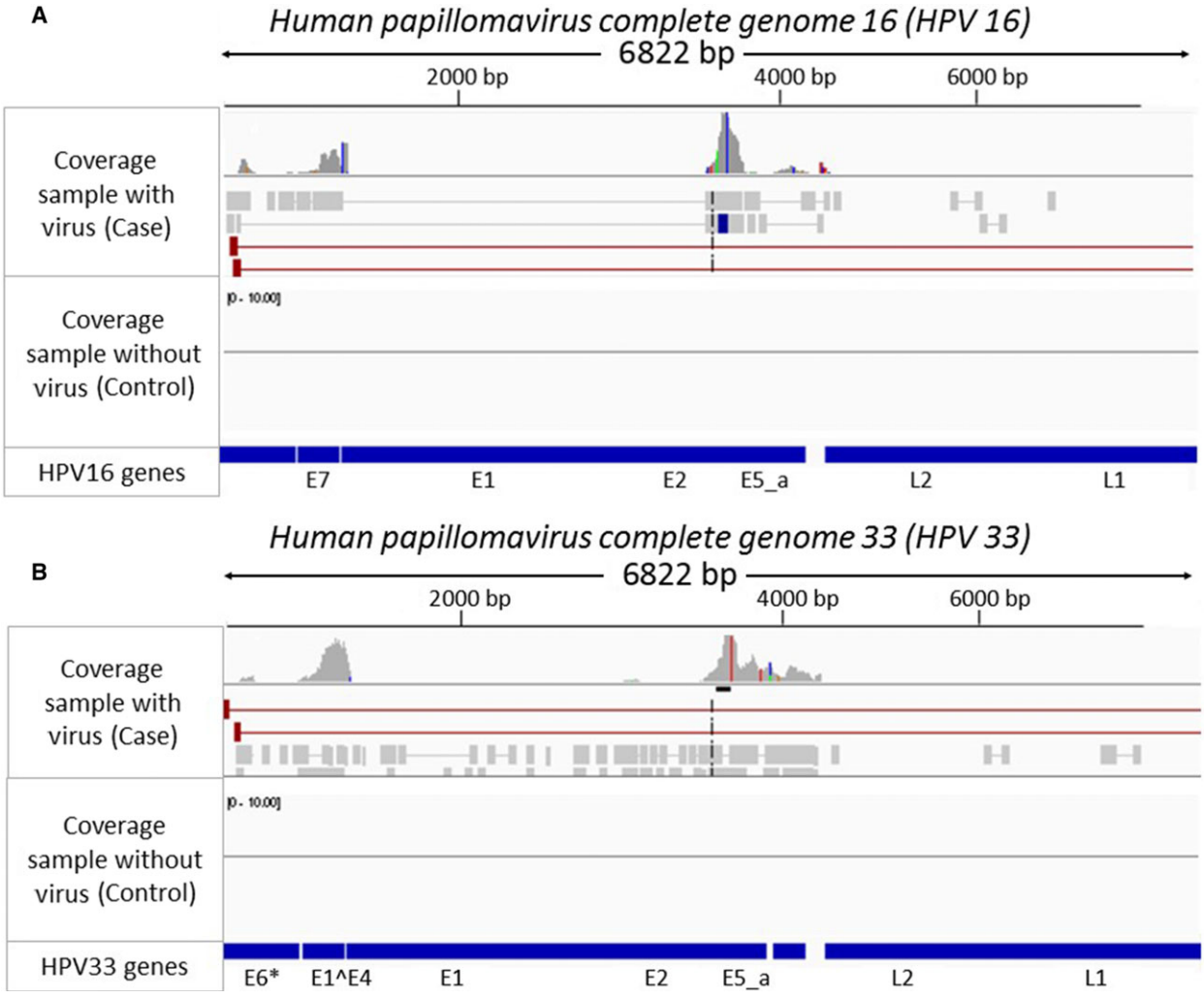

Fig. 3. The IGV figures of the nonhuman reads of two head and neck cancer samples affected by HPV16 (A) and HPV33 (B) showing the coverage of a sample affected by HPV16/HPV33 and control.

19. This RNA-seq analysis suggested that both E2 and E7 were expressed at high levels in HNSCC. On the other hand, E6 was only detected in 30% (6/20) of patients, and the expression level was not as high as that of E7 or E2. In addition to HPV16 and HPV18, which have been reported to be causative agents of carcinoma (Howley, 1986), we also found two HPV33-infected patients. The viral genes E7, E4, E1^E4, and E8^E2 were expressed in both patients, while HPV16 genes were not detected. These findings together suggest that HPV-induced HNSCC has a different molecular foundation from that of HPV-induced cervical carcinoma.

3.3. Molecular pathways in HPV16-induced HNSCC

Genetic instability is caused by steady accumulation of DNA damage and genetic variants, which leads to the activation of proto-oncogenes or the inactivation of tumor suppressor genes. A frequent event in cancers of the head and neck is the deletion of the short arm of chromosome 9, which results in inactivation of the host *p16* gene. In HPV-induced carcinogenesis, the most important initiating factor is the expression of the viral proteins E6 and E7, as they lead to the inactivation of the cellular tumor suppressor proteins p53 and retinoblastoma (Rb; Moll and Petrenko, 2003). The p16 expression assay is currently used clinically to identify HPV infection in HNSCC patients, because HPV+ individuals have very different prognosis and treatment options from those with non-virus-associated HNSCC. We identified highly significant upregulation of the tumor suppressor proteins CDKN2A (p16), CDKN2D (p19), and Tp53 in the samples from head and neck cancers with detected HPV16 infection when compared to samples without HPV16.

With ingenuity pathway analysis, we identified a significant enrichment of proteins involved in the G1/S checkpoint in HPV+ samples (Fig. 5). We first performed gene differential expression (DE) analysis by comparing 20 HPV16+ and 339 HPV16− HNSCC samples. We then used HTSeq for counting reads (Anders *et al.*, 2015) and DESeq for detecting DE genes (Anders and Huber, 2010). We detected 437 genes with adjusted $P < 0.05$ calculated based on negative binomial distribution. These genes showed significantly different expression levels between HPV16+ and HPV− groups. For DE genes, we performed enrichment analysis using DAVID (Huang *et al.*, 2009), which reported 22 pathways involving the DE genes that were differentially expressed between HPV+ and HPV− samples. The most probable pathway is illustrated in Fig. 4, which shows how the G1/S cell checkpoint was altered in HPV16+ HNSCC patients when compared to HPV− samples. Several molecules in this pathway were differentially expressed, with large differences found in CDKN2D (p16), CDKN2A (p19), E2F1, and TP53 expression. Tp53 is the degradation target of E6, so a higher mRNA level of Tp53 suggests a compensation mechanism of P53 expression to offset the E6 degradation. The expression level of TP53 in the HPV+ HNSCC samples was slightly higher than

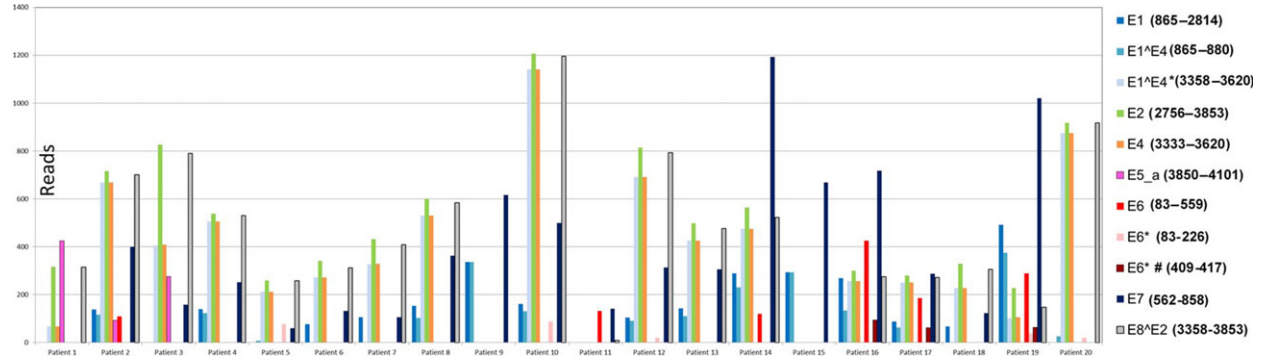

Fig. 4. HPV16 viral genes detected in HNSCC patients. Among the 20 patients with detectable HPV16 viral sequence, the number of detectable HPV16 genes ranged from 3 (patients 9, 11, and 15) to 10 (patient 19). E2 and E7 were most common, detected in 85% (17/20) and 90% (18/20) of patients, respectively. Most patients (15/20) had both detectable E2 and E7. The alternatively spliced E8^E2 was detected in 90% (18/20) of patients. The two patients (patients 9 and 15) without E8^E2 only had 3 detectable viral genes (E1, E1^E4, and E7). E1^E4 and E4 were detected in 95% (19/20) and 85% (17/20) of patients, respectively. However, E6 (83–559) was only detected in 30% (6/20) of patients. Fragments of E6*(83–226) and E6*# (409–417) were detected in 25% (5/20) and 15% (3/20) of patients, respectively. Patients 16 and 17 had detectable E6 (83–559) and E6*(83–226), while patient 19 had all E6 fragments detected. This viral expression pattern suggests that E2 and E7 are the major players in HPV-induced HNSCC.

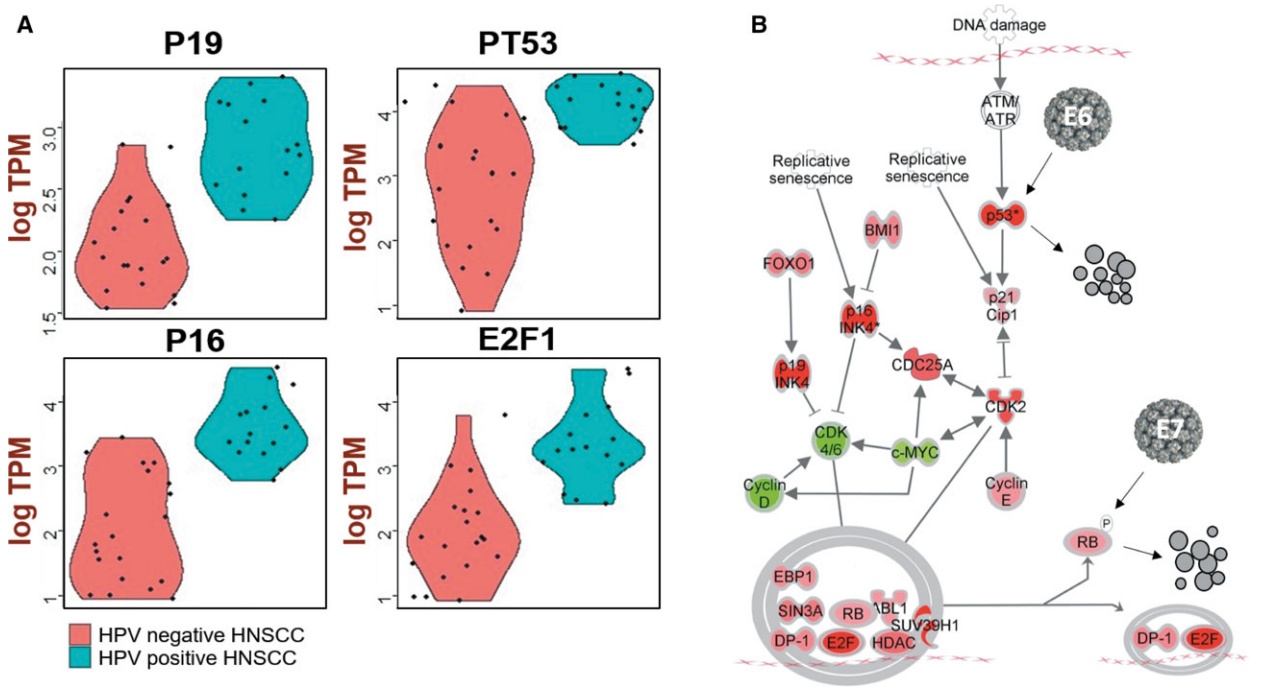

Fig. 5. Carcinogenesis pathway of HPV-induced HNSCC. (A) HNSCC patients with ($n = 20$) and without ($n = 20$) HPV infection had significantly different expression values for p16 ($P = 4.88E-7$), p19 ($P = 7.6E-6$), E2F1 ($P = 3.63E-5$), and TP53 ($P = 9.75E-7$). (B) Ingenuity pathway analysis (IPA) identified the molecules involved in the G1/S checkpoint to be the most significantly ($P = 3.28E-18$) active molecular pathway, including 17 genes with significantly different expression values between HPV+ and HPV− HNSCC ($P < 0.05$). Red: upregulated in HPV+ HNSCC; green: downregulated in HPV+ HNSCC. While E6 plays a minor role, E7 plays a major role in HPV carcinogenesis in HNSCC.

that in the HPV− HNSCC samples, which correlated well with the low level of E6 detected. This result further suggests that the carcinogenesis of HNSCC is different from the mechanism established for the better-studied cervical cancer.

4. Discussion

Here, we reported a novel virus detection method called VirTect that uses RNA-seq data as input. We tested and analyzed VirTect's performance on several datasets and identified several viruses from this analysis. Several groups have developed and discussed virus detection methods based on subtraction of nonhuman sequences from human samples (Chen *et al.*, 2013; Isakov *et al.*, 2011; Kostic *et al.*, 2011). For example, VirusSeq (Chen *et al.*, 2013) uses RNA-seq as input and distinguishes viral and nonviral sequences using the requirement of the minimum number of reads aligned to a viral sequence (set at 1000 reads). However, RNA-seq data from a sample might have thousands of nonhuman sequences, which may have repeat sequences that align to a virus but do not represent real viral genomes. Also, it is hard to only use the coverage of RNA-seq data for determining whether the number of reads were required to detect the existence of virus: the cutoff of 1000 reads (which recommended by VirusSeq) might be higher and cannot detect the existence of viruses. That is, the total number of mapped reads might not be a good filter to detect virus existence in RNA-seq. Therefore, we sought to develop a new method that would detect viruses from NGS data with greater sensitivity and accuracy, by designing a new multilayered filtering strategy. To confirm that this approach performs better, we compared VirTect with existing virus detection methods, namely VirusSeq (Chen *et al.*, 2013) and VirusFinder (Wang *et al.*, 2013). We found that VirTect could more accurately detect and discriminate noise and artifacts from real viral sequences. For example, for one head and neck cancer sample, VirusSeq wrongly attributed a poly(A) sequence to hepatitis C and tick-borne encephalitis, again due to the high coverage (more than 30×) and more than 1000 reads mapped to these viruses; however, it did not detect the real HPV16 sequences in this sample, since the number of mapped reads to HPV16 was 926, just below the threshold. In comparison, VirTect detected both the poly(A) sequence and the real HPV16 sequences (Fig. 3); after filtration, VirTect indicated the presence of HPV16.

With VirTect, we also analyzed 363 HNSCC samples available in the TCGA database. We detected HPV in 22 of these samples; 20 had detectable HPV16 genes and 2 had detectable HPV33 genes. We confirmed the viral etiology of these cancers by correlating RNA-seq and VirTect data to clinical and histopathologic findings. HPV-mediated carcinogenesis is thought to work through viral oncogenic proteins E6 and E7 via the disruption of cell cycle regulatory components (Sherman *et al.*, 1997). Specifically, E6 directly binds P53 and complexes it with the E6-AP ubiquitin ligase. This ubiquitination targets P53 for proteolytic degradation (Scheffner *et al.*, 1990) and impairs host cell G2/M checkpoint and apoptosis, resulting in uncontrolled proliferation and genomic instability (Taylor and Stark, 2001). Similarly, HPV E7 targets Rb for degradation (Boyer *et al.*, 1996) and results in a deregulated cell cycle and unstable genome. However, our data suggest that E7 plays a major role in the carcinogenesis of HNSCC, while E6 only plays a minor role. The E7 viral gene was detected in all 22 HPV+ HNSCC samples, but E6 was only detected in 6 of the 22 patients (20 HPV16+ and 2 HPV33+). Our findings are supported by a recent study of HPV16-driven oropharyngeal squamous cell carcinoma, which showed that HPV16 E6 seropositivity has low sensitivity (50%, 95% CI19–81) but is highly specific (100%, 95% CI 96–100) (Holzinger *et al.*, 2017).

To our surprise, E2 was detected in almost all HPV+ HNSCC patients. For cervical cancer, it has been reported that the integration of the HPV genome into the host genome often interrupts the E2 promoter and leads to unregulated E6 and E7 expression (Schwarz *et al.*, 1985; Smotkin and Wettstein, 1986). E2 is not detected in most cervical cancer cell lines with HPV infection (Vinokurova *et al.*, 2008). Furthermore, re-introduction of E2 to cervical cancer cell lines and HeLa cells was found to repress E6/E7 expression and lead to cellular senescence and apoptosis (Desaintes *et al.*, 1997; Dowhanick *et al.*, 1995; Goodwin *et al.*, 2000). In our study, E2 and E1^E4 were detected at high levels, which agreed with the previous report that E1^E4 stabilizes E2 (Davy *et al.*, 2009). The detection of E2 and E8^E2 viral fragments in RNA-seq data from HNSCC patients suggested that HPV-induced HNSCC has different molecular mechanisms of carcinogenesis from those of HPV-induced cervical cancers. Also, the cervical cancer high-risk strain HPV18 was not detected in any of the tested samples, but HPV33 was detected. Although only two patients had detectable HPV33 fragments, this is consistent with the report that HPV33 rarely induces HNSCC (Brennan *et al.*, 2017).

In light of these findings, we compared the gene expression profiles of HPV+ and HPV− HNSCC to identify a potential molecular pathway of HPV carcinogenesis in HNSCC. Our results agreed with the previous report that not all HPV16 infections associated with HPV16 E6 seropositivity (Maura *et al.*, 2015). This underscores that HPV16 E6 seropositivity is not a suitable assay for HPV-induced HNSCC screening. Importantly, we presented evidence here that, although cervical and oropharyngeal mucosal cancers are generally thought to have the same mechanisms of carcinogenesis given that they are associated with similar HPV subtype infections (e.g., HPV16), in fact these two anatomical sites may experience different paths to carcinogenesis, even with the same HPV16 subtype. This potentially has profound translational implications for diagnostics and therapeutics in head and neck oncology. The main limitation of VirTect is that it depends on a database of all known viruses, which is used to nominate candidate viruses in human cancer tissue. Though the database is extensible, this approach cannot detect novel viruses that are not in the database. However, VirTect can be used as a tool to screen known pathogens in RNA-seq or DNA-seq data for better understanding the role of viral infection in carcinogenesis.

5. Conclusions

We tested and evaluated VirTect on different datasets from TCGA, and we believe that VirTect can accurately detect viruses in NGS samples with greater accuracy than other methods. The ability to identify viral infection accurately may have important translational value for oncology, as we have demonstrated with a HNSCC dataset. This work has significant clinical relevance because of the importance of viral identification and characterization for precision medicine. For example, head and neck cancers which are associated with HPV16 have a better prognosis and are more sensitive to radiation therapy than non-HPV cancers. Therefore, accurate detection and knowledge of HPV status is essential for diagnosis, targeted therapeutic approaches, and prognostication to ultimately improve patient outcomes (Sedghizadeh *et al.*, 2016).

Acknowledgements

The results published here are in part based upon data generated by TCGA, managed by the NCI and NHGRI. We are grateful to TCGA for this source of data. Information about TCGA can be found at http://cancergenome.nih.gov. We also thank the Wang

lab members for helpful comments and feedback on VirTect. This work was supported by NIH grants CA197903 (JZ) and HG006465 (KW).

Conflict of interest

The authors declare no conflict of interest.

Author contributions

AK developed methods, analyzed data, interpreted data, and wrote the manuscript. QL contributed to method development, data analysis, and manuscript revision. KW and JZ coordinated the study. PPS and XZ evaluated and interpreted clinical data, prepared the pathology figures, as well as edited the manuscript. All authors discussed the biological findings and read/approved the final version of the present manuscript.

References

Anders S and Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* **11**, R106.

Anders S, Pyl PT and Huber W (2015) HTSeq-a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169.

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Arbyn M, Castellsague X, de Sanjose S, Bruni L, Saraiya M, Bray F and Ferlay J (2011) Worldwide burden of cervical cancer in 2008. *Ann Oncol* **22**, 2675–2686.

Bosch FX, Manos MM, Munoz N, Sherman M, Jansen AM, Peto J, Schiffman MH, Moreno V, Kurman R and Shah KV (1995) Prevalence of human papillomavirus in cervical cancer: a worldwide perspective. International biological study on cervical cancer (IBSCC) Study Group. *J Natl Cancer Inst* **87**, 796–802.

Boyer SN, Wazer DE and Band V (1996) E7 protein of human papilloma virus-16 induces degradation of retinoblastoma protein through the ubiquitin-proteasome pathway. *Can Res* **56**, 4620–4624.

Brennan K, Koenig JL, Gentles AJ, Sunwoo JB and Gevaert O (2017) Identification of an atypical etiological head and neck squamous carcinoma subtype featuring the CpG island methylator phenotype. *EBioMedicine* **17**, 223–236.

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

Chiu CY (2013) Viral pathogen discovery. *Curr Opin Microbiol* **16**, 468–478.

Davy C, McIntosh P, Jackson DJ, Sorathia R, Miell M, Wang Q, Khan J, Soneji Y and Doorbar J (2009) A novel interaction between the human papillomavirus type 16 E2 and E1–E4 proteins leads to stabilization of E2. *Virology* **394**, 266–275.

Desaintes C, Demeret C, Goyat S, Yaniv M and Thierry F (1997) Expression of the papillomavirus E2 protein in HeLa cells leads to apoptosis. *EMBO J* **16**, 504–514.

Dowhanick JJ, McBride AA and Howley PM (1995) Suppression of cellular proliferation by the papillomavirus E2 protein. *J Virol* **69**, 7791–7799.

Dunne WM Jr, Westblade LF and Ford B (2012) Next-generation and whole-genome sequencing in the diagnostic clinical microbiology laboratory. *Eur J Clin Microbiol Infect Dis* **31**, 1719–1726.

Goodwin EC, Yang E, Lee CJ, Lee HW, DiMaio D and Hwang ES (2000) Rapid induction of senescence in human cervical carcinoma cells. *Proc Natl Acad Sci USA* **97**, 10978–10983.

Holzinger D, Wichmann G, Baboci L, Michel A, Hofler D, Wiesenfarth M, Schroeder L, Boscolo-Rizzo P, Herold-Mende C, Dyckhoff G *et al.* (2017) Sensitivity and specificity of antibodies against HPV16 E6 and other early proteins for the detection of HPV16-driven oropharyngeal squamous cell carcinoma. *Int J Cancer* **140**, 2748–2757.

Howley PM (1986) On human papillomaviruses. *N Engl J Med* **315**, 1089–1090.

Huang W, Sherman BT and Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**, 1–13.

Isakov O, Modai S and Shomron N (2011) Pathogen detection using short-RNA deep sequencing subtraction and assembly. *Bioinformatics* **27**, 2027–2030.

Kostic AD, Ojesina AI, Pedamallu CS, Jung J, Verhaak RGW, Getz G and Meyerson M (2011) PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nat Biotechnol* **29**, 393–396.

Leung TH, Tang HW, Siu MK, Chan DW, Chan KK, Cheung AN and Ngan HY (2017) HPV-E6 protein enriches the CD55(+) population in cervical cancer cells promoting radio-resistance and cancer aggressiveness. *J Pathol* **244**, 151–163.

Leyden WA, Manos MM, Geiger AM, Weinmann S, Mouchawar J, Bischofj K, Yood MU, Gilbert J and Taplin SH (2005) Cervical cancer in women with comprehensive health care access: attributable factors in the screening process. *J Natl Cancer Inst* **97**, 675–683.

Li H and Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G and Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079.

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal* **17**, 10.

Maura LG, Anil KC, William FA and Carole F (2015) Epidemiology of human papillomavirus-positive head and neck squamous cell carcinoma. *J Clin Oncol* **33**, 3235–3242.

Moll UM and Petrenko O (2003) The MDM2-p53 interaction. *Mol Cancer Res* **1**, 1001–1008.

Quinlan AR and Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842.

Raj K, Berguerand S, Southern S, Doorbar J and Beard P (2004) E1 empty set E4 protein of human papillomavirus type 16 associates with mitochondria. *J Virol* **78**, 7199–7207.

Relman DA (1999) The search for unrecognized pathogens. *Science* **284**, 1308–1310.

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

Schwarz E, Freese UK, Gissmann L, Mayer W, Roggenbuck B, Stremlau A and zur Hausen H (1985) Structure and transcription of human papillomavirus sequences in cervical carcinoma cells. *Nature* **314**, 111–114.

Sedghizadeh PP, Billington WD, Paxton D, Ebeed R, Mahabady S, Clark GT and Enciso R (2016) Is p16-positive oropharyngeal squamous cell carcinoma associated with favorable prognosis? A systematic review and meta-analysis. *Oral Oncol* **54**, 15–27.

Sherman L, Jackman A, Itzhaki H, Stoppler MC, Koval D and Schlegel R (1997) Inhibition of serum- and calcium-induced differentiation of human keratinocytes by HPV16 E6 oncoprotein: role of p53 inactivation. *Virology* **237**, 296–306.

Smotkin D and Wettstein FO (1986) Transcription of human papillomavirus type 16 early genes in a cervical cancer and a cancer-derived cell line and identification of the E7 protein. *Proc Natl Acad Sci USA* **83**, 4680–4684.

Taberna M, Mena M, Pavon MA, Alemany L, Gillison ML and Mesia R (2017) Human papillomavirus related oropharyngeal cancer. *Ann Oncol* **28**, 2386–2398.

Taylor WR and Stark GR (2001) Regulation of the G2/M transition by p53. *Oncogene* **20**, 1803–1815.

Trapnell C, Pachter L and Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111.

Vinokurova S, Wentzensen N, Kraus I, Klaes R, Driesch C, Melsheimer P, Kisseljov F, Durst M, Schneider A and von Knebel Doeberitz M (2008) Type-dependent integration frequency of human papillomavirus genomes in cervical lesions. *Can Res* **68**, 307–313.

Wang QG, Jia PL and Zhao ZM (2013) VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. *PLoS One* **8**, e64465.

Wang QG, Jia PL and Zhao ZM (2015) VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. *Genome Med* **7**, 2.

Weinstock H, Berman S and Cates W (2004) Sexually transmitted diseases among American youth: incidence and prevalence estimates, 2000. *Perspect Sex Reprod Health* **36**, 6–10.

Zur Hausen H (2009) The search for infectious causes of human cancers: where and why. *Virology* **392**, 1–10.