NGS-based approach to determine the presence of HPV and their sites of integration in human cancer genome

P Chandrani 1,2, V Kulkarni 1,2, P Iyer 1, P Upadhyay 1, R Chaubal 1, P Das 1, R Mulherkar 1, R Singh 1 and A Dutt*,1

1Advanced Centre for Treatment, Research and Education in Cancer, Tata Memorial Centre, Kharghar, Navi Mumbai, Maharashtra 410210, India

Background: Human papilloma virus (HPV) accounts for the most common cause of all virus-associated human cancers. Here, we describe the first graphic user interface (GUI)-based automated tool ‘HPVDetector’, for non-computational biologists, exclusively for detection and annotation of the HPV genome based on next-generation sequencing data sets.

Methods: We developed a custom-made reference genome that comprises of human chromosomes along with annotated genome of 143 HPV types as pseudochromosomes. The tool runs on a dual mode as defined by the user: a ‘quick mode’ to identify presence of HPV types and an ‘integration mode’ to determine genomic location for the site of integration. The input data can be a paired-end whole-exome, whole-genome or whole-transcriptome data set. The HPVDetector is available in public domain for download: http://www.actrec.gov.in/pi-webpages/AmitDutt/HPVdetector/HPVDetector.html.

Results: On the basis of our evaluation of 116 whole-exome, 23 whole-transcriptome and 2 whole-genome data, we were able to identify presence of HPV in 20 exomes and 4 transcriptomes of cervical and head and neck cancer tumour samples. Using the inbuilt annotation module of HPVDetector, we found predominant integration of viral gene E7, a known oncogene, at known 17q21, 3q27, 7q35, Xq28 and novel sites of integration in the human genome. Furthermore, co-infection with high-risk HPVs such as 16 and 31 were found to be mutually exclusive compared with low-risk HPV71.

Conclusions: HPVDetector is a simple yet precise and robust tool for detecting HPV from tumour samples using variety of next-generation sequencing platforms including whole genome, whole exome and transcriptome. Two different modes (quick detection and integration mode) along with a GUI widen the usability of HPVDetector for biologists and clinicians with minimal computational knowledge.

Human papilloma viral (HPV) infections has been associated with various types of cancer. Epidemiological studies indicate that about 90% of cervical cancers, 90–93% of anal canal cancers, 12–63% of oropharyngeal cancers, 36–40% of penile cancers, 40–64% of vaginal cancers and 40–51% of vulvar cancers are attributable to HPV infection (Munoz et al, 2003; Shukla, 2009). Currently, HPV detections are primarily carried out using PCR-based MY09/11 and CPI/II systems (Kleter et al, 1998; Brink et al, 2007; Abreu et al, 2012). These technologies come with limitations to detect minor, low-abundance HPV genotypes and a complex mixture of co-infections that can be a negative determinant of the clinical outcome (Mendez et al, 2005; Trottier et al, 2006). Next-generation sequencing (NGS) technologies overcomes such limitations, as evident from the recently described
High-risk HPV genotyping assay for primary cervical cancer screening based on self-collection (Yi et al., 2014), using TEN16 or HIVID methodology, and to determine co-infection among the HPV types probed along with their sites on integration (Johansson et al., 2013; Xu et al., 2013; Li et al., 2013b; Ameer et al., 2014; Hu et al., 2015). However, there is an unmet need for a simplified tool for biologists with no previous experience or knowledge of informatics to analyse the data generated by whole-exome, transcriptome or genome sequencing using NGS technology to detect the presence of HPV sequences along with their integration sites. There are a variety of gene integration finding tools available that can detect different pathogen insertions in the human genome such as ViralFusionSeq (Li et al., 2013a), VirusSeq (Chen et al., 2013), VirusFinder (Wang et al., 2013), Path-Seq (Kostic et al., 2011), RINS (Bhaduri et al., 2012), and ReadSCAN (Naem et al., 2013). These tools have their specific third-party needs, and are not specific for HPV detection. They can detect presence of a HPV sequence along with other viruses, but lack information to annotate the region of detection. They can detect presence of a HPV sequence along with other viruses, but lack information to annotate the region of detection. Here, we describe ‘HPVDetector’ as a specific in silico automated tool that is capable of multi-HPV type detection, their annotation and determination of site of HPV integration utilising raw exome, transcriptome, or whole-genome data as input with minimal requirement for third-party tools.

MATERIALS AND METHODS

HPV detection involves a computational subtraction-based approach, where NGS data are used for alignment against custom-made HPV multi-reference genome sequences to detect the traces of multiple HPV types using an automated pipeline (Figure 1).

HPV reference sequences and annotation. As a first step of the pipeline, HPV genomes in fasta format is required. We have acquired GenBank (.gb) files of 143 types of HPVs from a web resource Papillomavirus Episteme (PAVE) (Van Doorslaer et al., 2013). We converted these GenBank (.gb) files into fasta files. All these reference sequences were concatenated to compose a multi-fasta sequence using bio-perl modules (Stajich et al., 2002). Apart from this, we also parsed the GenBank (.gb) files to generate a HPV gene reference having nucleotide intervals for each gene of each HPV type. This gene reference file was used to annotate the HPV gene.

HPV type and HPV-aligned reads detection. Evaluating the HPV type and HPV-aligned reads is crucial to find HPV in the respective sample. For HPV type detection, we indexed the multi-fasta HPV reference file using BWA aligner followed by alignment of reads to indexed genome (Li and Durbin, 2009). The aligned reads were extracted from the same file using a utility ViewSam from Picard Tools package (http://broadinstitute.github.io/picard/). The alignment files were parsed using UNIX shell program to detect the type of HPV as well as number of reads that align to a particular HPV type. Number of HPV reads were normalised to the total depth of coverage per sample and with respect to different HPV gene sizes.

Assessment of specificity and sensitivity of HPVDetector. We downloaded SiHa whole-genome sequence from Sequence Read Archive database of DDBJ (https://trace.ddbj.nig.ac.jp/DRASearch/study; SRP048769). The data were converted from SRA to FASTQ using SRAtoolkit. The resulting FASTQ files represents >36 x genome coverage which was further downsampled to 1 x, 2 x, 3 x, 4 x, 5 x, 10 x, 15 x, 20 x, 25 x and 30 x using Picard Toolkit’s DownsampleSam function (http://broadinstitute.github.io/picard/). The resulting FASTQ files were used for testing HPV detection using HPVDetector.

Human–HPV integration loci detection. To detect integration sites, we created a custom reference genome comprised of human chromosomes and HPV fasta sequences as pseudochromosomes. HPV genomes were appended to human chromosomes to compose a multi-fasta reference genome. This custom Human–HPV reference genome was then used for aligning reads with short-read aligner BWA. The alignment files were parsed for the reads where one mate is aligned to human chromosome and another to HPV. The Human chromosomal positions, HPV type and HPV reference position were parsed and annotated with a gene reference annotation file acquired from UCSC table browser (Karolchik et al., 2004) to get a list of integration sites.

RNA extraction, cDNA synthesis and E6-specific PCR. Total RNA extraction was performed from primary tumours and cell lines using Trizol reagent (Invitrogen, Grand Island, NY, USA) as per the manufacture’s instruction and later resolved on 1.2% agarose gel to confirm the RNA integrity. DNase treatment was done using the DNase Free kit (Ambion, Foster City, CA, USA; cat AM1906) followed by first-strand cDNA synthesis taking 2 µg of total RNA using Superscript III kit (Invitrogen, 18080-051). E6 (HPV-16) and GAPDH expression were checked as described previously (Smeets et al., 2007).

HPV detection using MY09/11 and PCR primers. MY09/11 primer sequences were taken from previously reported literature (Baay et al., 1996). All samples were screened by PCR first using MY09/11 primer. GAPDH was used as internal control for each sample. SiHa cell line (Adler et al., 1997) was used as a positive control for HPV and AW13516 cell line (Tatake et al., 1990) as a negative control. The PCR reaction was performed in 20 µl volume
HPVDetector is a tool to quickly detect hundreds of HPV types from next-generation sequence data without any prerequisite knowledge about virus types. It runs on paired-end sequenced samples. It is composed of two modes or sub pipelines as quick detect and integration detect mode.

**Quick detect mode.** This mode is to quickly determine the HPV type or types to check whether multiple HPV co-infections are existing or not in a given sample. Quick detect mode starts with alignment of raw paired-end sequencing reads against the custom-made multi-HPV genome using BWA aligner. Computational subtraction of the reads is then carried out, in which HPV-aligned reads are retained using Picard Tools and further processed using UNIX shell program to distinguish reads mapping to different HPV types. Finally, HPVDetector outputs a result file, which enlists one or more HPV type(s) and number of HPV reads.

**Integration detect mode.** This mode of HPVDetector determines the genomic location of HPV integrant, annotate with HPV gene, human chromosomal loci and human genome. This mode of HPVDetector pipeline starts with alignment of raw reads against a custom-made reference including a pseudochromosome such as the multi-fasta reference genome containing 143 HPV reference sequences and the HG19 human reference genome. Computational subtraction was carried out to retain discordant read pairs where the sequences are aligned to both human as well as HPV genomes. Finally, HPVDetector outputs a result file, which enlists HPV integration loci on the human genome, annotation of HPV genes, human genes and human genome cytobands.

**Detection of HPV type integrated in the host genome**

**Cervical cancer exome sequencing data.** We analysed 22 cervical cancer exome sequencing data (generated in-house at ACTREC, unpublished data) to detect the presence of HPV. Among the 22 samples analysed, HPV was detected in 18 cervical samples, with maximum number of reads supporting the HPV16 sequence (Figure 2) (Das et al., 2012). We also detected the presence of additional HPV types such as HPV71 (in six samples), HPV82 (in five samples) and HPV31 (in two samples) with variable number of supporting reads as shown in (Figure 3). Co-infection with more than one HPV type is known to be associated with significantly increased risk of cervical intraepithelial neoplasia 2 + and found in 43.2% of HPV-positive women (Liu et al., 2001; Mendez et al., 2005; Vaccarella et al., 2010; Chaturvedi et al., 2011). Six of 22 cervical cancer patients (43%) were found to be co-infected with one or more HPV subtypes in this study using HPVDetector (Figure 3). Interestingly to note, based on phylogenetic analysis of HPV types, HPV16 and HPV31 of the virulent alpha 7 group infection occurred in a mutually exclusive manner (in 13 of 22 samples), whereas HPV71 of the alpha 15 subgroup, known to be involved in commensal infections that infected 6 of 14 cervical tumour samples, invariably co-occurred with other HPV subtypes (Schiffman et al., 2009; Harari et al., 2014). The HPV sequence detected in primary cervical tumour sample were independently validated by directed sequencing in T1094, the only sample with sufficient quality DNA (as shown in Supplementary Figure 1).

**Tongue squamous cell carcinoma exome and transcriptome data.** HPV is an independent risk factor in head and neck squamous cell carcinoma (HNSCC), in particular for oral and oropharyngeal carcinomas (Chaudhary et al., 2009; Pannone et al., 2011). We analysed whole-exome data from 23 paired and one orphan tongue squamous cell carcinoma (TSCC) sample and 7 HNSCC cell lines (generated in-house at ACTREC, unpublished data). None of the TSCC primary tumours were found to be HPV positive, as reported earlier (Siebers et al., 2008; Patel et al., 2014; Tsimplaki et al., 2014). The absence of HPV infection were further validated by PCR using MY09/11 and E6 on genomic DNA and cDNA, respectively, suggesting a low false-negative feature of the HPVDetector primers (Supplementary Figure 2). At the same time, among the cell lines, NT8e cells (Mulherkar et al., 1993) were HPV positive, as reported earlier (Siebers et al., 2008; Patel et al., 2014). The HPV18 genes (E1, E6, and E7) were validated in 11 cell lines (generated in-house at ACTREC, unpublished data) using the HPVDetector. Three of 17 primary tumours were found to be HPV18 positive. In addition, HPV18 reads were found in HEP2 cell line, consistent with earlier reports in literature (Ogura et al., 1993). The HPV18 genes (E1, E6, and E7) were validated in HEP2 cell line by PCR and Sanger sequencing (as shown in Supplementary Figure 1 and Table 2).

**Gall bladder and liposarcoma exome and whole-genome data.** We analysed 13 gall bladder cancer whole-exome, 1 gall bladder cancer whole-transcriptome and 1 liposarcoma whole-genome sequence data (generated in-house at ACTREC, unpublished data). No trace of the HPV sequence was detected in these samples.

**Assessment of specificity and sensitivity of HPVDetector.** SiHa cell line developed from a cervical squamous cell carcinoma patient represents single-copy integration of HPV16 (el Awady et al., 1987). We analysed SiHa whole-genome sequence using HPVDetector. Consistent with a published report (Hu et al., 2015), HPVDetector could detect integration at chr13 intragenic location of KLF5—KLF12 genes and other regions (Supplementary Table 1). The integration was validated by PCR followed by sequencing (Supplementary Figure 3).
reads with as low as 1–10 reads (grey). HPV16 reads (black) were counted using HPVDetector against SiHa whole-genome sequence coverage. Beyond 25 ×, no significant increase in HPV reads were found, suggesting saturation of genome coverage (Figure 4). In addition, among primary tumours, two pairs of HPV56 reads detected by the HPVDetector in T9440 as described in Supplementary Table 1 were validated earlier by Luminex array and SPF1/2 (Das et al., 2013). Thirty-two proteins E6 and E7 function as oncogenes by regulating the known genes found to be predominantly represented among the cervical squamous tumours infected with HPV16, followed by E4, E5 and E6, in decreasing order (Figure 5). Of these genes found to be enriched among all the integrants, it is interesting to note that the viral proteins E6 and E7 function as oncogenes by regulating the known

HPVDetector in T9440 as described in Supplementary Table 1 and read counts

To determine the sensitivity of HPVDetector, we downsampled the SiHa genome using a ‘downsampling’ method, a Picard Toolkit’s DownsampleSam function (http://broadinstitute.github.io/picard/) (Meynert et al., 2014) to generate varying coverage of the SiHa whole-genome data ranging from 1 × to 30 × coverage, and analysed using HPVDetector. Reads supporting presence of HPV reads linearly increased as a function of increasing coverage from 1 × to 25 × coverage. Beyond 25 ×, no significant increase in HPV reads were found, suggesting saturation of genome coverage (Figure 4). In addition, among primary tumours, two pairs of HPV56 reads detected by the HPVDetector in T9440 as described in Supplementary Table 1 were validated earlier by Luminex array and SPF1/2 (Das et al., 2012). Taken together, this suggests HPVDetector could detect reads with as low as 1 × genome coverage with reads supported by as low as just two paired reads.

To enable accurate gene annotation of the HPV genome sequenced, we prepared a gene annotation database of 143 HPV types from PAVE database (Van Doorslaer et al., 2013). Thirty-two reads of viral ORFs were found in 5 of 11 cervical tumours positive for HPV-16. Following the normalisation for the total number of reads against the length of individual genes, the viral gene E7 was found to be predominantly represented among the cervical tumours infected with HPV16, followed by E4, E5 and E6, in decreasing order (Figure 5). Of these genes found to be enriched among all the integrants, it is interesting to note that the viral proteins E6 and E7 function as oncogenes by regulating the known
human tumour suppressors, p53 and pRb, respectively (Lu et al., 2003; Yim and Park, 2005).

**Determination of the HPV integration sites in the host genome.** We identified 55 integration sites in 7 cervical cancer tumour samples T1099, T1123, T755, T887, T938, T1094, and T959 and 1 head and neck tumour sample using the HPVDetector (Supplementary Table 1). In this study, chromosomal loci 17q21, 3q27, 7q35, and Xq28 were observed with higher frequency compared with other loci for HPV integration, as reported earlier (Thorland et al., 2003). Interestingly to note, we found HPV integration in the following fragile regions—(1p, 1q, 2p, 2q, 3p, 3q, 4p, 4q, 5p, 5q, 7q, 9q, 10q, 11p, 11q, 12p, 12q, 13q, 15q, 17q, 18q, 22q, Xp and Xq) that are prone to chromosome breaks to facilitate foreign DNA integration (Figure 6) (Smith et al., 2006).

In T1123 and T755 HPV16 integration sites were detected at chr1q42.3 and chr3q23, respectively, identical to as reported earlier (Wentzensen et al., 2004; Schmitz et al., 2012). In addition, in T755 integration of HPV16 were found within the coding region at SLC25A36, a pyrimidine nucleotide carrier. This site of integration were also determined in T755 and T1123 samples using the APOT assay, as described earlier (Das et al., 2012) (Supplementary Table 1).

In total, we analysed 116 exome, 23 transcriptome and 2 whole-genome sequencing data, out of which we have detected presence of HPV in 20 exome and 4 transcriptome data (Table 1).

**DISCUSSION**

HPV accounts for the most common cause of all virus-associated human cancers. However, despite large-scale genome-wide DNA sequencing efforts of the cancer genome, there is no dedicated
informatics tool to rapidly detect the presence of HPV in these genomes, in an inclusive manner. There are indeed a variety of gene integration finding tools available that can detect different pathogen insertions in the human genome, such as ViralFusionSeq, VirusSeq, VirusFinder, Path-Seq, RINS, and ReadSCAN. These sophisticated tools although have their specific third-party needs, necessitate extensive computational infrastructure, cannot be run without specialised and advanced computational expertise of the researcher, and more importantly are not specific for HPV detection, per se—for example, lacks information to annotate the region of the HPV genome to predict the integrated viral gene, of which some are known to function as oncogenes.

We present a new user-friendly in silico tool ‘HPVDetector’ as a unique tool to analyze NGS data to detect HPV sequences for non-computational biologists. Using the HPVDetector tool, we have detected 55 integration sites from the cervical exome and head and neck transcriptome data set. The tool allowed us to perform a comprehensive analysis to generate the information for co-occurrence of HPV subtypes across cervical cancer patients that is known to affect the clinical outcome of the disease. In addition, our finding of significant enrichment of viral gene E7 > E4 > E5 > E6 reads among the cervical tumour samples, using the inbuilt annotation module of the HPVDetector, is consistent with the known biology of HPV genes and their role in carcinogenesis, as E6 and E7 are known viral oncogenes. This unique feature of the HPVDetector with an inbuilt HPV annotation module could potentially be helpful to understand the function of other HPV ORFs with unknown function by studying their incidence against varying tumour stage and types. Although the analysis of cervical tumours were restricted to its exome data set, a complete spectrum of the load of viral genes present in a sample can similarly be determined using the whole-genome data as input to the HPVDetector.

HPVDetector demonstrate a low false-negative and false-positive rate that can detect HPV reads at as low as 1 × genome coverage. Reads supported by even two paired reads were found to be credible. No viral reads were detected across 54 head and neck primary tumour samples of Indian origin, as reported earlier (Siebers et al, 2008; Patel et al, 2014; Tsimpaki et al, 2014), but detected a low-risk HPV71 in a cell line that could be validated by performing MY09/11 PCR on the primary tumours as shown in Supplementary Figure 2. On the other hand, all the four HPV reads detected across different tumour types using HPVDetector could be validated by directed PCR followed by Sanger sequencing. One interesting utility of the HPVDetector would be to explore for HPV reads in NGS data from different cancer types.

### Table 1. Summary of HPV detection in all samples

| Study type                  | Samples tested | Presence of virus in samples |
|-----------------------------|----------------|-------------------------------|
| Cervical cancer exome       | 36 (22 tumour, 14 paired normal) | 18                            |
| SiHa cell line WGS          | 1              | 1                             |
| HNSCC cell line exome       | 7              | 1                             |
| TSCC exome                  | 47 (24 tumour, 23 paired normal) | 0                             |
| Gall bladder exome          | 26 (13 tumour, 13 paired normal) | 1                             |
| TSCC transcriptome          | 17 (11 tumour, 6 paired normal) | 3                             |
| HNSCC cell line transcriptome| 5              | 1                             |
| Gall bladder transcriptome   | 1              | 0                             |
| Liposarcoma WGS             | 1              | 0                             |
| Total number of samples     | 141            | 25                            |

Abbreviations: HNSCC = head and neck squamous cell carcinoma; HPV = human papilloma virus; TSCC = tongue squamous cell carcinoma; WGS = whole-genome sequencing.

### ACKNOWLEDGEMENTS

Prasad Kanvinde for the help with web hosting of HPVDetector and other IT-related support at ACTREC, Tata Memorial Centre, Mumbai. Dhwanit Shah for help with coding of the tool; all members of the Dutt laboratory for critically reviewing the manuscript. Genotypic Inc., SciGenome Pvt. Ltd and Sandor Proteomics for providing sequencing services. AD is supported by an Intermediate Fellowship from the Wellcome Trust/DBT India Alliance (IA/1/11/2500278), by a grant from DBT (BT/PR2372/AGR/36/696/2011), and intramural grants (IRB project 92 and 55). PC and PI are supported by a senior research fellowship from ACTREC. PU is supported by a senior research fellowship from CSIR. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

### CONFLICT OF INTEREST

The authors declare no conflict of interest.

### REFERENCES

Abreu AL, Souza RP, Gimenes F, Consolario ME (2012) A review of methods for detect human Papillomavirus infection. Virol J 9: 262.

Adler K, Erickson T, Bobrow M (1997) High sensitivity detection of HPV-16 in SiHa and CaSki cells utilizing FISH enhanced by TSA. Histochem Cell Biol 108(4-5): 321–324.

Amour A, Meiring TL, Bunikis I, Haggqvist S, Lindau C, Lindberg JH, Gustavsson I, Mbulawa ZZ, Williamson AL, Gyllensten U (2014) Comprehensive profiling of the vaginal microbiome in HIV positive women using massive parallel semiconductor sequencing. Sci Rep 4: 4398.

Baay MF, Quint WG, Koudstaal J, Hollemma H, Duk JM, Burger MP, Stolz E, Herbrink P (1996) Comprehensive study of several general and type-specific primer pairs for detection of human papillomavirus DNA by PCR in paraffin-embedded cervical carcinomas. J Clin Microbiol 34(3): 745–747.

Bhaduri A, Qu K, Lee CS, Ungevikell A, Khavari PA (2012) Rapid identification of non-human sequences in high-throughput sequencing datasets. Bioinformatics 28(8): 1174–1175.

Brink AA, Snijders PJ, Meijer CJ (2007) HPV detection methods. Dis Markers 23(4): 273–281.
Xu B, Chotewutmontri S, Wolf S, Klos U, Schmitz M, Durst M, Schwarz E (2013) Multiplex identification of human papillomavirus 16 DNA integration sites in cervical carcinomas. PLoS One 8(6): e66693.

Yi X, Zou J, Xu J, Liu T, Liu T, Hua S, Xi F, Nie X, Ye L, Luo Y, Xu L, Du H, Wu R, Yang L, Liu R, Yang B, Wang J, Belinson JL (2014) Development and validation of a new HPV genotyping assay based on next-generation sequencing. Am J Clin Pathol 141(6): 796–804.

Yim EK, Park JS (2005) The role of HPV E6 and E7 oncoproteins in HPV-associated cervical carcinogenesis. Cancer Res Treat 37(6): 319–324.

Supplementary Information accompanies this paper on British Journal of Cancer website (http://www.nature.com/bjc)