FULL PAPER

BJC

British Journal of Cancer (2015) 112, 1958–1965 | doi: 10.1038/bjc.2015.121

Keywords: HPV detection; human cancer; next-generation sequencing (NGS)

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]

[1]*Advanced 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). Other techniques used include the hybridisation-based SPF LiPA method, signal-amplification assays (Hybrid Capture 2 and Cervista) and nucleic-acid-based amplification-like microarray, real-time PCR-based methods (COBAS 4800 real-time test) (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

*Correspondence: Dr A Dutt; E-mail: adutt@actrec.gov.in
[2]These authors contributed equally to this work.

Received 31 July 2014; revised 3 March 2015; accepted 7 March 2015; published online 14 May 2015

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 intregration (Johansson *et al*, 2013; Xu *et al*, 2013; Li *et al*, 2013b; Ameur *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 (Naeem *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 the HPV genome detected. 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).

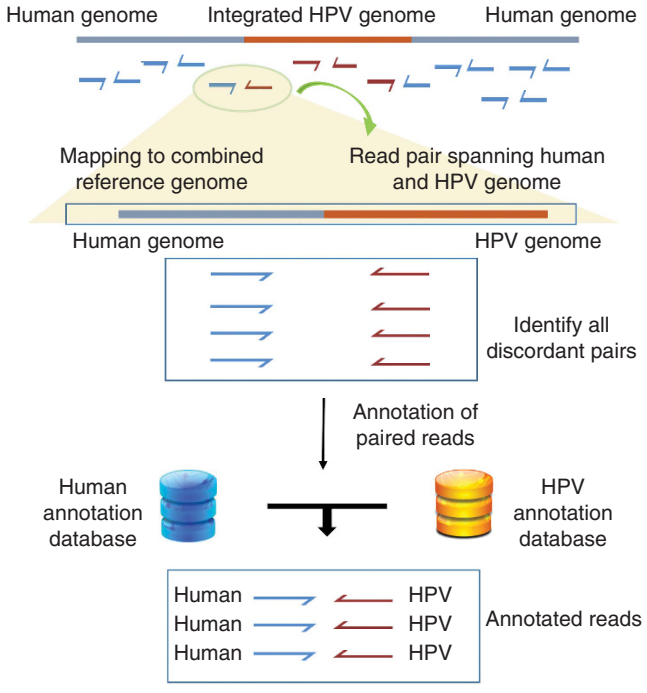

Figure 1. Conceptual workflow of the HPVDetector. The flowchart represents workflow for HPVDetector. Paired-end reads obtained from next-generation sequencing data are aligned to a combined Human–HPV reference database. All discordant read pairs with one read aligning to human and other to the HPV genome are identified and annotated utilising human and HPV database using an inbuilt annotator module.

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 $>36\times$ genome coverage which was further downsampled to $1\times$, $2\times$, $3\times$, $4\times$, $5\times$, $10\times$, $15\times$, $20\times$, $25\times$ and $30\times$ 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

containing 10 μl (2 ×) Biomix-Red master mix (Biomix, Port Orange, FL, USA; cat Bio25005), 5 μM each primer, 50 ng gDNA. PCR condition was: initial denaturation: 95 °C, denaturation: 94 °C for 1 min, annealing: 55 °C, (MY09/11, GAPDH), extension: 72 °C for 1 min and final extension at 72 °C for 5 min on a PCR machine (Veriti 96-Well Fast Thermal Cycler, Applied Biosystems, Carlsbad, CA, USA). Primers were designed to amplify 122-, 126- and 120-bp read sequences of HPV16 identified by HPVDetector in SiHa cell line. Primers flanking the human reads were designed to amplify 119 and 290 bp, respectively. These sequences were further validated by Sanger sequencing. All the experiments were repeated at least twice independently. The details of the primers used in the study are provided in Supplementary Table 2.

RESULTS

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 gene. 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 (Liaw *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). Interesting 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).

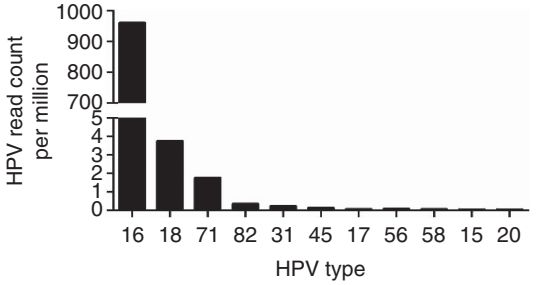

Figure 2. Quantitative representation, by number of reads, of HPV types detected in cervical tumours. The graph represents distribution of total number of HPV reads per million of total reads for all HPV types detected across 17 cervical samples. HPV16 has the highest number of reads across 17 samples followed by HPV(18, 71, 45, 31, 82, 17, 56, 58, 15, 20) in the decreasing order of their read counts.

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*, 1997) of seven cell lines analysed was found to be positive for HPV71. Next, we analysed whole-transcriptome data of 17 TSCC and 6 TSCC 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).

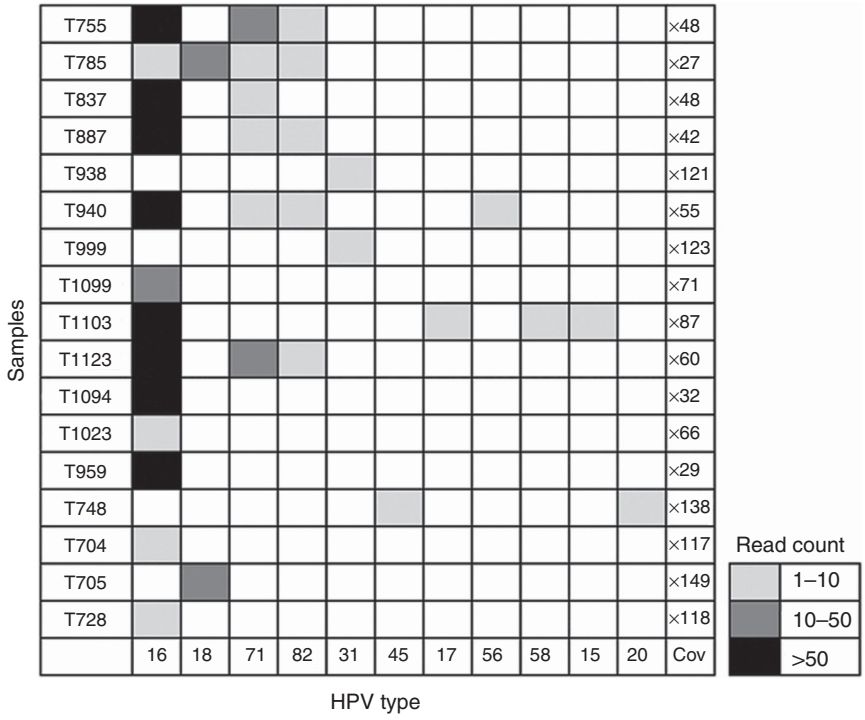

Figure 3. HPV gene integration frequency across different cervical cancer samples. Heatmap representation of HPV types detected across 17 cervical cancer samples. HPV16 in 13 samples, HPV18 in 2 samples, HPV71 in 6 samples, HPV82 in 5 samples, HPV31 in 2 samples, HPV45 in 5 samples, and HPV17, 56, 58, 15, 20 were detected in 1 sample, each. Total coverage of the exome sequencing is indicated in the last column 'cov'. On the basis of read count, the abundance of HPV is graded in different samples: read counts 1–10 as light grey; read counts 10–50 as dark grey; and read counts >50 as black.

Sesnsitivity. 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.

Specificity. Having benchmarked the HPVDetector against SiHa for sensitivity, next we tweaked the SiHa whole-genome sequence data to test specificity of the tool by taking reverse (not complement) of the SiHa genome to simulate as a random sequence but retaining composition of nucleotides and genome complexity, using an in-house perl script. We found no spurious HPV reads when the SiHa whole-genome sequence was reversed, suggesting the HPVDetector is specific to detect true HPV traces. Further, to address the issue of specificity among primary tumours, we performed another round of functional validation on tongue squamous tumours that were found HPV negative based on HPVDetector (Supplementary Table 1) and validated by My09/11 primers using genomic DNA. We analysed the expression of HPV E6 (Supplementary Figure 2b) in these samples. All samples were found negative for HPV presence. This suggests that the tool has low false negative.

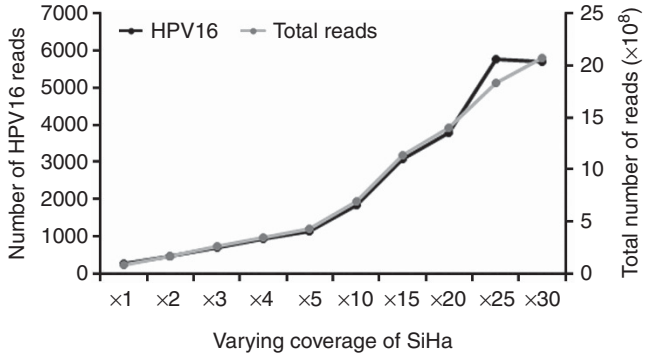

Figure 4. Sensitivity of HPVDetector as a function of increasing genome sequence coverage. SiHa WGS data were downsampled to the coverage of 1 ×, 2 ×, 3 ×, 4 ×, 5 ×, 10 ×, 15 ×, 20 ×, 25 ×, and 30 ×. HPV16 reads (black) were counted using HPVDetector at varying coverage and plotted along with total number of reads (grey).

Annotation of the HPV genome integrated in the host genome. 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). Interesting 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).

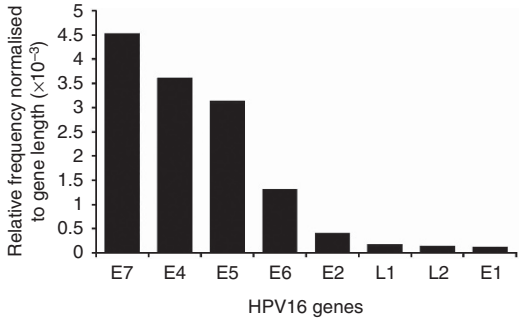

Figure 5. Relative frequency of integration of HPV genes in cervical carcinoma. HPV16 reads were annotated using an inbuilt annotation module of the HPVDetector to identify the viral genes. Number of reads per viral gene were normalised to the gene length, and frequency reads for individual genes are plotted, as shown.

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

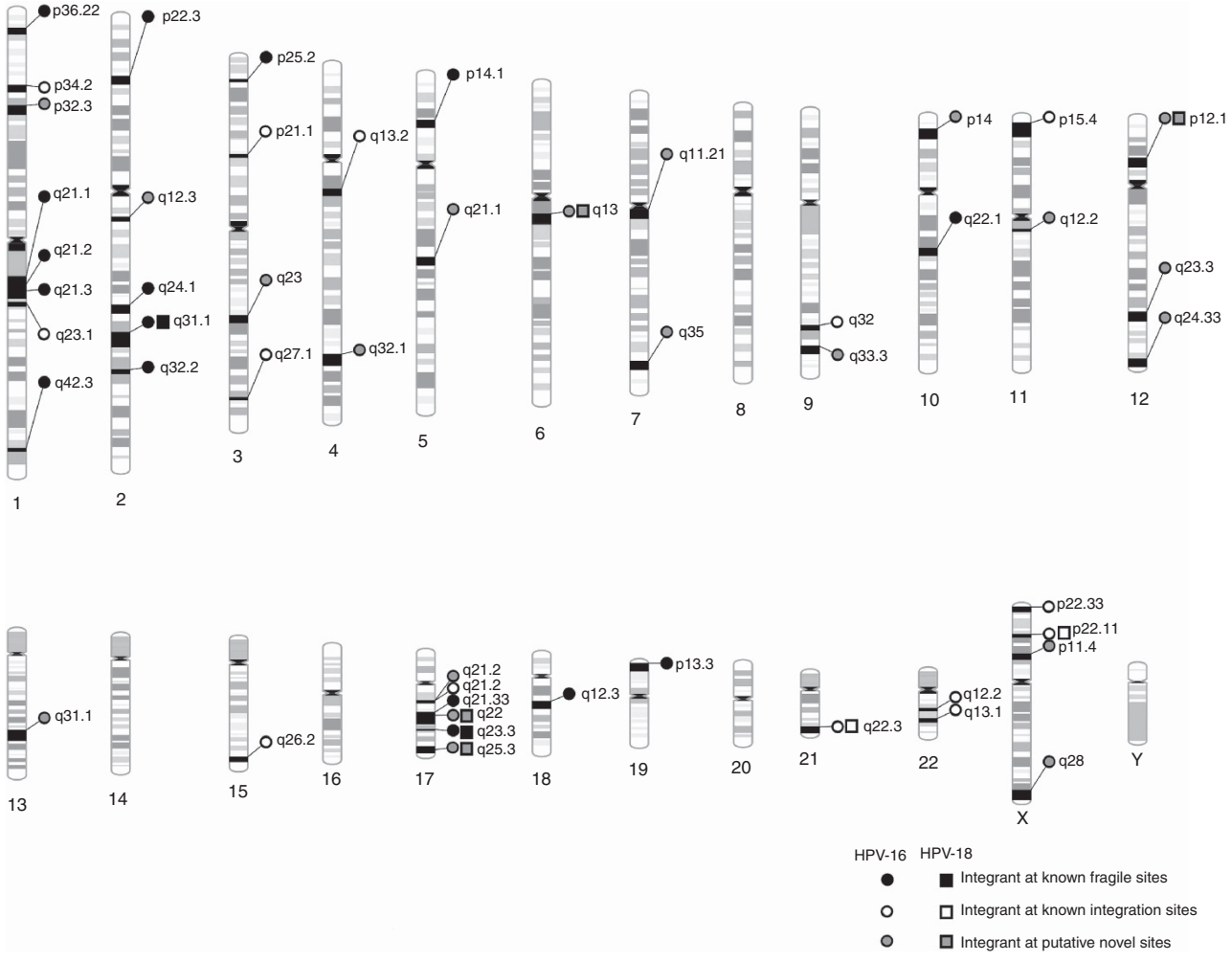

Figure 6. Schematic representation of all HPV16 and 18 integration sites in the human genome detected across cervical cancer samples using HPVDetector. Site of integration as determined by HPVDetector in cervical cancer samples is shown. HPV16 integration sites are depicted by circles and HPV18 by rectangles. Black, open and grey circles or rectangles represent integrations at known fragile sites, at known integration sites and at novel sites, respectively.

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.

informatics tool to rapidly detect the presence of HPV in these genomes, in an exclusive 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 intense 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; Tsimplaki *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. We analysed 13 gall bladder exome, 1 gall bladder transcriptome and 1 liposarcoma whole-genome sequencing data using HPVDetector. No HPV reads were found in these samples in this study.

Another critical feature of the HPVDetector is determination of HPV integration sites at the host genome. These integrations are known to occur at preferred regions of the genome (Thorland *et al*, 2003; Matovina *et al*, 2009). Using the integration site detection feature of the HPVDetector, we detected integration at various chromosomal locations (for e.g., 1p, 2p, 2q, 3p, 3q), some with significant overlap to the known fragile sites in literature and at several novel sites as summarised in Figure 6 and Supplementary Table 1. In summary, HPVDetector is a simple yet precise and robust tool for detecting HPV from tumour samples using variety of NGS 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 with minimal computational knowledge (as described in the attached supplementary 'HPVDetector User Guide' including Supplementary Figures 4 and 5).

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/I/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, Consolaro 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.

Ameur 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, Hollema 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, Ungewickell 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.

Chaturvedi AK, Katki HA, Hildesheim A, Rodriguez AC, Quint W, Schiffman M, Van Doorn LJ, Porras C, Wacholder S, Gonzalez P, Sherman ME, Herrero R. CVT Group (2011) Human papillomavirus infection with multiple types: pattern of coinfection and risk of cervical disease. *J Infect Dis* **203**(7): 910–920.

Chaudhary AK, Singh M, Sundaram S, Mehrotra R (2009) Role of human papillomavirus and its detection in potentially malignant and malignant head and neck lesions: updated review. *Head Neck Oncol* **1**: 22.

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

Das P, Thomas A, Mahantshetty U, Shrivastava SK, Deodhar K, Mulherkar R (2012) HPV genotyping and site of viral integration in cervical cancers in Indian women. *PLoS One* **7**(7): e41012.

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

Harari A, Chen Z, Burk RD (2014) Human papillomavirus genomics: past, present and future. *Curr Probl Dermatol* **45**: 1–18.

Hu Z, Zhu D, Wang W, Li W, Jia W, Zeng X, Ding W, Yu L, Wang X, Wang L, Shen H, Zhang C, Liu H, Liu X, Zhao Y, Fang X, Li S, Chen W, Tang T, Fu A, Wang Z, Chen G, Gao Q, Li S, Xi L, Wang C, Liao S, Ma X, Wu P, Li K, Wang S, Zhou J, Wang J, Xu X, Wang H, Ma D (2015) Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism. *Nat Genet* **47**(2): 158–163.

Johansson H, Bzhalava D, Ekstrom J, Hultin E, Dillner J, Forslund O (2013) Metagenomic sequencing of 'HPV-negative' condylomas detects novel putative HPV types. *Virology* **440**(1): 1–7.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**(Database issue): D493–D496.

Kleter B, van Doorn LJ, ter Schegget J, Schrauwen L, van Krimpen K, Burger M, ter Harmsel B, Quint W (1998) Novel short-fragment PCR assay for highly sensitive broad-spectrum detection of anogenital human papillomaviruses. *Am J Pathol* **153**(6): 1731–1739.

Kostic AD, Ojesina AI, Pedamallu CS, Jung J, Verhaak RG, Getz G, Meyerson M (2011) PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nat Biotechnol* **29**(5): 393–396.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754–1760.

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

Li W, Zeng X, Lee NP, Liu X, Chen S, Guo B, Yi S, Zhuang X, Chen F, Wang G, Poon RT, Fan ST, Mao M, Li Y, Li S, Wang J, Jianwang, Xu X, Jiang H, Zhang X (2013b) HIVID: an efficient method to detect HBV integration using low coverage sequencing. *Genomics* **102**(4): 338–344.

Liaw KL, Hildesheim A, Burk RD, Gravitt P, Wacholder S, Manos MM, Scott DR, Sherman ME, Kurman RJ, Glass AG, Anderson SM, Schiffman M (2001) A prospective study of human papillomavirus (HPV) type 16 DNA detection by polymerase chain reaction and its association with acquisition and persistence of other HPV types. *J Infect Dis* **183**(1): 8–15.

Lu DW, El-Mofty SK, Wang HL (2003) Expression of p16, Rb, and p53 proteins in squamous cell carcinomas of the anorectal region harboring human papillomavirus DNA. *Mod Pathol* **16**(7): 692–699.

Matovina M, Sabol I, Grubisic G, Gasperov NM, Grce M (2009) Identification of human papillomavirus type 16 integration sites in high-grade precancerous cervical lesions. *Gynecol Oncol* **113**(1): 120–127.

Mendez F, Munoz N, Posso H, Molano M, Moreno V, van den Brule AJ, Ronderos M, Meijer C, Munoz A. Instituto Nacional de Cancerologia Human Papillomavirus Study Group (2005) Cervical coinfection with human papillomavirus (HPV) types and possible implications for the prevention of cervical cancer by HPV vaccines. *J Infect Dis* **192**(7): 1158–1165.

Meynert AM, Ansari M, FitzPatrick DR, Taylor MS (2014) Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics* **15**: 247.

Mulherkar R, Goud AP, Wagle AS, Naresh KN, Mahimkar MB, Thomas SM, Pradhan SA, Deo MG (1997) Establishment of a human squamous cell carcinoma cell line of the upper aero-digestive tract. *Cancer Lett* **118**(1): 115–121.

Munoz N, Bosch FX, de Sanjose S, Herrero R, Castellsague X, Shah KV, Shah KV, Snijders PJ, Meijer CJ. International Agency for Research on Cancer Multicenter Cervical Cancer Study Group (2003) Epidemiologic classification of human papillomavirus types associated with cervical cancer. *N Engl J Med* **348**(6): 518–527.

Naeem R, Rashid M, Pain A (2013) READSCAN: a fast and scalable pathogen discovery program with accurate genome relative abundance estimation. *Bioinformatics* **29**(3): 391–392.

Ogura H, Yoshinouchi M, Kudo T, Imura M, Fujiwara T, Yabe Y (1993) Human papillomavirus type 18 DNA in so-called HEP-2, KB and FL cells–further evidence that these cells are HeLa cell derivatives. *Cell Mol Biol (Noisy-le-grand)* **39**(5): 463–467.

Pannone G, Santoro A, Papagerakis S, Lo Muzio L, De Rosa G, Bufo P (2011) The role of human papillomavirus in the pathogenesis of head & neck squamous cell carcinoma: an overview. *Infect Agent Cancer* **6**: 4.

Patel KR, Vajaria BN, Begum R, Desai A, Patel JB, Shah FD, Shukla SN, Patel PS (2014) Prevalence of high-risk human papillomavirus type 16 and 18 in oral and cervical cancers in population from Gujarat, West India. *J Oral Pathol Med* **43**(4): 293–297.

Schiffman M, Clifford G, Buonaguro FM (2009) Classification of weakly carcinogenic human papillomavirus types: addressing the limits of epidemiology at the borderline. *Infect Agent Cancer* **4**: 8.

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

Shukla S (2009) Infection of human papillomaviruses in cancers of different human organ sites. *Indian J Med Res* **130**: 222–233.

Siebers TJ, Merkx MA, Slootweg PJ, Melchers WJ, van Cleef P, de Wilde PC (2008) No high-risk HPV detected in SCC of the oral tongue in the absolute absence of tobacco and alcohol–a case study of seven patients. *Oral Maxillofac Surg* **12**(4): 185–188.

Smeets SJ, Hesselink AT, Speel EJ, Haesevoets A, Snijders PJ, Pawlita M, Meijer CJ, Braakhuis BJ, Leemans CR, Brakenhoff RH (2007) A novel algorithm for reliable detection of human papillomavirus in paraffin embedded head and neck cancer specimen. *Int J Cancer* **121**(11): 2465–2472.

Smith DI, Zhu Y, McAvoy S, Kuhn R (2006) Common fragile sites, extremely large genes, neural development and cancer. *Cancer Lett* **232**(1): 48–57.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehväslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* **12**(10): 1611–1618.

Tatake RJ, Rajaram N, Damle RN, Balsara B, Bhisey AN, Gangal SG (1990) Establishment and characterization of four new squamous cell carcinoma cell lines derived from oral tumors. *J Cancer Res Clin Oncol* **116**(2): 179–186.

Thorland EC, Myers SL, Gostout BS, Smith DI (2003) Common fragile sites are preferential targets for HPV16 integrations in cervical tumors. *Oncogene* **22**(8): 1225–1237.

Trottier H, Mahmud S, Costa MC, Sobrinho JP, Duarte-Franco E, Rohan TE, Ferenczy A, Villa LL, Franco EL (2006) Human papillomavirus infections with multiple types and risk of cervical neoplasia. *Cancer Epidemiol Biomarkers Prev* **15**(7): 1274–1280.

Tsimplaki E, Argyri E, Xesfyngi D, Daskalopoulou D, Stravopodis DJ, Panotopoulou E (2014) Prevalence and expression of human papillomavirus in 53 patients with oral tongue squamous cell carcinoma. *Anticancer Res* **34**(2): 1021–1025.

Vaccarella S, Franceschi S, Snijders PJ, Herrero R, Meijer CJ, Plummer M. IARC HPV Prevalence Surveys Study Group (2010) Concurrent infection with multiple human papillomavirus types: pooled analysis of the IARC HPV Prevalence Surveys. *Cancer Epidemiol Biomarkers Prev* **19**(2): 503–510.

Van Doorslaer K, Tan Q, Xirasagar S, Bandaru S, Gopalan V, Mohamoud Y, Huyen Y, McBride AA (2013) The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. *Nucleic Acids Res* **41**(Database issue): D571–D578.

Wang Q, Jia P, Zhao Z (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**(5): e64465.

Wentzensen N, Vinokurova S, von Knebel Doeberitz M (2004) Systematic review of genomic integration sites of human papillomavirus genomes in epithelial dysplasia and invasive cancer of the female lower genital tract. *Cancer Res* **64**(11): 3878–3884.

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)