Mapping of HPV transcripts in four human cervical lesions using RNAseq suggests quantitative rearrangements during carcinogenic progression

Jinmiao Chen a, Yuezhen Xue b, Michael Poidinger a, Thimothy Lim c, Sung Hock Chew c, Chai Ling Pang a, Jean-Pierre Abastado a,*, François Thierry b

a SlgN (Singapore Immunology Network), Biopolis, 8A Biomedical Grove, Immunos, Singapore 138648, Singapore
b Institute of Medical Biology A*STAR, Biopolis, 8A Biomedical Grove, Immunos, Singapore 138648, Singapore
c KK Hospital, Singapore 229899, Singapore

A R T I C L E  H I S T O R Y

Article history:
Received 21 April 2014
Returned to author for revisions 2 May 2014
Accepted 16 May 2014
Available online 14 June 2014

Keywords:
HPV
RNAseq
Patient samples
Transcripts
Quantification
Carcinogenesis
Promoter usage

A B S T R A C T

Two classes of Human papillomaviruses (HPV) infect the anogenital track: high risk viruses that are associated with risk of cervical cancer and low risk types that drive development of benign lesions, such as condylomas. In the present study, we established quantitative transcriptional maps of the viral genome in clinical lesions associated with high risk HPV16 or low risk HPV6b. Marked qualitative and quantitative changes in the HPV16 transcriptome were associated with progression from low to high grade lesions. Specific transcripts encoding essential regulatory proteins such as E7, E2, E1widehatE4 and E5 were identified. We also identified intrinsic differences between the HPV6b-associated condyloma transcript map and that of the HPV16-associated low grade CIN specifically regarding promoter usage. Characterization and quantification of HPV transcripts in patient samples thus establish the impact of viral transcriptional regulation on the status of HPV-associated lesions and may therefore help in defining new biologically-relevant prognosis markers.

© 2014 Elsevier Inc. All rights reserved.

Introduction

Human papillomaviruses (HPV) of high risk type, such as HPV16, can directly transform mucosal epithelia and are considered important etiologic agents of cervical, anal and oropharyngeal cancers. More than 99% of cervical cancers contain HPV genomes and express HPV transcripts (Walboomers et al., 1999). In these lesions, and in cell lines established there from, HPV transcripts are initiated at the viral regulatory region and contain the E6 and E7 open reading frames (ORFs) (Schwarz et al., 1986, 1985; Seedorf et al., 1987). Indeed, viral E6 and E7 oncogenic proteins are predominantly responsible for degradation of the cell cycle regulatory factors p53, TAp63 and pocket proteins, amongst other processes contributing to carcinogenic transformation (Ben Khalifa et al., 2011) (reviewed in (Munger and Howley, 2002; McLaughlin-Drubin and Munger, 2009; Howie et al., 2009)). In addition, continuous expression of E6 and E7 is necessary for maintenance of the transformed phenotype in cell culture. In patients, the carcinogenic progression induced by HPV is a multi-step process, starting with infection of the keratinocytes forming the basal layer of the epithelium, followed by the development of low and high grade precancerous lesions (Cervical Intraepithelial Neoplasia or CIN), and eventually fully-fledged carcinoma of either Squamous Cell Carcinoma (SCC) or adenocarcinoma (ADC) types. Carcinogenic progression requires persistent HPV infection and is thought to be associated with major rearrangements in the HPV transcriptome, leading eventually to sole expression of E6 and E7. However, little is known about the nature of these rearrangements. To date, viral transcription within the high risk genotypes HPV16, 18 and 31 has most commonly been studied in vitro in cell lines or keratinocytes differentiated in 3D raft cultures (Schwarz et al., 1983; Milligan et al., 2007; Baker et al., 1987; Zheng and Baker, 2006; Smotkin and Wettstein, 1986; Gray et al., 2010; Wang et al., 2011; Schmitt et al., 2011). A small number of cell lines have also been derived from lower grade lesions containing non-integrated episomal high risk HPV genomes or from benign condylomas associated with low risk HPV types genomes (Milligan et al., 2007; Bedell et al., 1991); HPV16 transcripts have been mapped in one such line (Milligan et al., 2007; Doobar et al., 1990) but the extent to which these data reflect the in vivo situation remains unclear. Moreover, these studies cannot address the important question of how the transcript map might evolve during carcinogenic progression, and so any association between...
changes in HPV transcription and in vivo oncogenesis remains undefined. Recent data from our lab suggest that loss of E2 protein expression, which is common in cervical cancer (Xue et al., 2010), is unexpectedly independent of the disruption of the E2 gene that occurs during integration of the viral genome in ~40% of HPV16-associated cancers (Xue et al., 2012; Das Ghosh et al., 2012). To understand this, better knowledge of the structure of viral transcripts is needed, and of their temporal regulation during cancer progression, in particular regarding expression of E1\widehat{E}4 and E2 transcripts relative to that of E6 and E7. The E1\widehat{E}4 protein is highly expressed in the upper layers of differentiated cells in CIN but not in SCC (Doorbar, 2005); the E2 protein is also expressed in CIN in intermediate and upper epithelial layers, while expression of E7 mainly occurs in lower layers and in the full thickness of the SCC. This pattern of expression thus links E1\widehat{E}4 and E2 to cellular differentiation, while in contrast E7 expression occurs in undifferentiated proliferative epidermal cells. However, how transcripts of these different viral proteins are expressed and regulated in vivo has not been resolved. In this study we present a comprehensive description of HPV transcriptomes using RNAseq analyses of HPV16-associated patient lesions and a HPV6b-associated condyloma.

Results and discussion

Comparison of transcript maps from HPV16-associated cervical lesions with published in vitro HPV16 transcript maps

RNA was extracted from ~20 fresh biopsies obtained from patients with HPV-associated lesions. Histological grades and HPV genotypes of all samples were determined with three HPV16-associated lesions (one CIN2, one CIN3 and one SCC), and one HPV6b-associated low grade condyloma selected for deep sequencing. Serial sections of the HPV16-associated CIN2 and SCC and HPV6b-condyloma samples used for the deep RNA sequencing were subjected to both IHC to detect viral E2 protein and the p16\textsuperscript{INK4} cellular marker (a surrogate marker indicating expression of the viral E7 protein), and in situ hybridization (ISH) for viral DNA detection, as previously described (Xue et al., 2010, 2012) (Fig. 1). These results confirmed our previous observations of E2 protein expression in the intermediate and upper layers of CIN2 but not in SCC, while in contrast p16\textsuperscript{INK4} is detected in the 2/3 lower layers of CIN2 and uniformly in SCC samples. HPV DNA detected by ISH showed episomal amplification in the upper layers of CIN2 (uniform labeling of the entire nuclei) and an integrated pattern in the SCC (punctate labeling). These labeling patterns were also remarkably similar in the HPV6b-associated condyloma and HPV16-associated CIN2, especially regarding the concomitant expression of E2 and viral DNA replication, which is shown here in the HPV6v-associated condyloma using an HPV16 E2 antibody that cross reacts with HPV6 E2 protein as reported previously (Lai et al., 2013). The CIN3 lesion sample was not of sufficient size for sectioning.

In the three HPV16-associated lesions, a total of 21 distinct viral transcripts were identified (Figs. 2 and 3, Table 1). Thirteen (A–Q, Fig. 2) exclusively covered the early region of the HPV16 genome and used the early polyadenylation site ~nt 4215 of the viral genome. These “early transcripts” have been reported previously (Zheng and Baker, 2006). In addition 8 “late transcripts” were detected, of which 2 were already known (P and Q), and 6 were newly identified here (represented with the prefix “cuff”). Seven of these late transcripts used the polyadenylation site at nt 7321, while the remaining transcript (cuff21) corresponded to an almost full-length covering of the genome with undefined coding capacity (Fig. 3). The total absolute abundance of all viral transcripts was ~1200 FPKM (Fragments Per Kilobase of transcript per Million fragments mapped) in CIN2, ~120 in CIN3 and ~800 in SCC (Table 1). Remarkably, 16 of the 21 transcripts used the strong E1 splice donor at nt 880 (Figs. 2 and 3).

Comparison of the HPV16 transcriptome in CIN and SCC lesions

Our data confirmed that all HPV16 transcripts are polycistronic, as reported earlier (Zheng and Baker, 2006). To determine the coding capacity of the 20 defined transcripts, we searched for translation initiation sequences (AUG with Kozak’s motif) located at least 18 nt downstream from the 5’ end of the transcripts. The latter length is considered the minimal size of mRNA 5’UTR required for efficient landing of the ribosomal complex, while the average size 5’UTR in eukaryotic transcripts is ~200 nt (Chatterjee and Pal, 2008). In addition, and in accordance with the general scanning model of RNA translation in eukaryotic cells, we only considered the 5’ most proximal AUG with “Kozak motif”. The coding capacity of the 20 mRNAs is given in Fig. 2 for the “early transcripts” and in Fig. 3 for the “late transcripts”. The relative abundance of transcripts in the three analyzed HPV16-associated lesions is given in the right part of Figs. 2 and 3, with the absolute values presented in Table 1.

Eight of the 20 HPV16 transcripts were putative mRNAs encoding the E7 protein. The most frequently used transcription start sites (TSSs) map to around nt 97 and nt 670, which have both been previously described. P\textsubscript{97} is the main promoter for transcription of E6 and E7 in HPV16-associated cervical carcinoma cell lines SiHa and Caski (Smotkin and Wettstein, 1986; Smotkin et al., 1989) and in transfected cells (Romanczuk et al., 1990). The P\textsubscript{670} promoter has also been defined as a differentiation inducible promoter (Grassmann et al., 1996). The use of upstream start sites such as 4, 10 or 36 (Fig. 3) has also been proposed but while our approach did not focus on precise determination of the start site for every transcript, we did observe that 5–7% of all transcripts started upstream of P\textsubscript{97} in SCC and CIN3 (discussed below). Interestingly, while transcription starting at P\textsubscript{97} is repressed by E2 (Romanczuk et al., 1990; Thierry, 2009), initiation at the other promoters, either downstream or upstream, was predicted to be insensitive to E2 repression, being located away from the E2 binding sites.

Remarkably, none of the “late transcripts”, defined as transcripts terminating at the “late” polyadenylation site, used the P\textsubscript{97} TSS. Instead, they made use of putative TSSs at upstream or downstream positions (Fig. 3). These “late transcripts” represented a relatively small fraction of the total viral RNAs and are relatively more represented in the CIN2 lesion (~10% of the total transcripts) than in CIN3 and SCC (~3–4%) (Table 1).

Description of HPV16 transcripts encoding early proteins

E6: None of the late transcripts and only three early transcripts (A, F and K) contained a full length E6 ORF, as all other transcripts were interrupted within E6 at the splice donor located at nt 226 (Fig. 4A). These transcripts started at TSS 97, which is only 7 nt upstream of the E6 AUG initiation codon and thus is unlikely to be an efficient translation initiation site for E6 protein. Moreover these transcripts were only evident in the SCC and CIN3 samples where they represent only ~4% of the total transcripts, with transcript A being the most abundant (Fig. 2). As approximately 7–8% of total viral transcripts began upstream of P\textsubscript{97}, especially visible in the SCC sample (Fig. 4B), it is therefore possible that the E6 transcripts may also have been initiated upstream of P\textsubscript{97}. In fact there are two major putative upstream putative TSSs, at ~nt 36 or nt 46, that could give rise to transcripts containing leader sequences of ~70 nt that would allow efficient translation of E6. The relative scarcity of full length E6 transcripts in HPV-associated genital
lesions has been reported before (Bohm et al., 1993; Nilsson et al., 1996).

E7: In sharp contrast to E6, E7 transcripts are highly represented both in terms of absolute number (8) and relative abundance in the three clinical samples (Table 1). As expected for the major onco-
genic protein, the highest expression of E7 transcripts was observed in CIN3 and SCC, where they accounted for ~45% and 56% of all viral transcripts respectively (Fig. 2 and Table 1). Five E7 transcripts started at P97 and terminated at the “early” polyA site, with transcript B being the most prominent (44% in SCC) (Fig. 2). These transcripts are predicted to be sensitive to E2 repression, and they appeared to be amplified in CIN3 and SCC where E2 is either not expressed or is expressed but at low levels compared to in CIN2. However, in CIN1 and CIN2, expression of E2 and E7 proteins is normally spatially separated within the epithelium thus making E7 transcription via the use of the P97 promoter possible (Xue et al., 2010). Interestingly, three putative E7 transcripts were also found among the late transcripts described in Fig. 3, although they were less abundant and all contained the downstream ORFs encoding capsid proteins (Fig. 3 and Table 1). When the levels of E6 and E7 transcripts were compared, it was apparent that E7 transcripts are well expressed in all lesions where they are over 10-fold more abundant than E6 transcripts in high grade lesions while E6 transcripts are undetectable (under the threshold) in CIN2 (Table 1) in accordance with the semi-quantitative data obtained by in situ hybridization reported earlier (Nilsson et al., 1996).

E2: We recently described detection of the E2 protein in HPV16-associated clinical samples and found high levels of expression in CIN3 and SCC (Xue et al., 2010). However, in only 3 of these transcripts (Q, R and S), was the E1\hat{E}4 AUG in a favorable position to be translated (Figs. 2 and 3). The E1\hat{E}4 protein is highly expressed specifically in the most differentiated layers of low grade lesions and is used as a marker of productive viral infection (Doorbar, 2005). Therefore, and as for the E2 mRNAs, the quantitative data showing higher expression of the E1\hat{E}4 transcripts in CIN2 compared to CIN3 and SCC is in agreement with the known protein expression pattern (Fig. 2 and Table 1). In the “early” transcripts map, the E1\hat{E}4 messenger Q was the most highly expressed transcript in CIN2, representing ~80% of all transcripts, while its expression was
Transcription of the capsid proteins

Transcripts able to drive expression of the L1 and L2 capsid proteins were terminated at the “late” Poly A site ~200 nt downstream of the termination codon of the L1 ORF (Fig. 3). Six out of the 8 late transcripts described in Fig. 3 possessed a complete L1 ORF, using a splice donor at nt 3632 after the E4 and the splice acceptor at nt 5638, corresponding exactly to the L1 AUG (Fig. 3). L1 is expected to be expressed only in low grade CIN, where productive infection occurs and new viral particles are formed. In only two transcripts (cuff16 and cuff127) was the L1 AUG the most 5’ AUG for translation initiation, and both were expressed at very low levels in CIN2 and CIN3, but not in SCC. The most highly expressed putative L1 transcript (S) was initiated at TSS 670, and could encode both E1\widehat{E}4 and L1 but with a much more favorable translational efficiency for E1\widehat{E}4 than L1 due to better availability of its AUG initiation codon (Fig. 3). A similar transcript, not using the splice donor after E4, could also code for L2 but was expressed at very low levels and only in CIN2, where again the E1\widehat{E}4 translation would be more favorable.

A major transcriptional rearrangement occurs before CIN3

The relative distribution of the HPV16 transcripts was quite similar in the CIN3 and SCC lesions: L1-encoding transcripts were barely detectable in these 2 lesions (Fig. 3), while E7-encoding transcripts constituted around half of all viral transcripts, followed by E1\widehat{E}4- and E5-encoding transcripts (Fig. 2). Transcripts were differently represented in the CIN2 lesion where E7-encoding transcripts comprised only ~5.3% of the total, while in contrast E1\widehat{E}4 transcripts were preponderantly expressed.
at 80% (Fig. 2). This indicated that progression from CIN2 to CIN3 is associated with major qualitative and quantitative rearrangements of HPV16 transcription as illustrated by the comparison of depth read profiles encompassing only the proximal 1000 nt region of HPV16 in the three types of lesions (Fig. 4A). These comparable quantifications clearly highlight the much lower transcription covering of the E7 area (between nt 500 and 800) as well as the almost undetectable transcription over the specific E6 area (between nt 226 and 409) in CIN2 when compared to CIN3 and SCC. In addition, it illustrates the conserved use of previously published splice donors and acceptors and TSS in all lesions, as indicated at the top of the read depth profiles (Fig. 4A). Quantitative distribution of transcripts given by read depth profiles over the full HPV16 sequence obtained in CIN2 and SCC confirmed very different patterns (Fig. 4B): the E6 and E7 ORFs were more transcribed in SCC, while the strong splicing for E1\textsuperscript{\textdagger}E4 was used in both samples. Transcripts in the late region, between ~nt 4000 and late polyA at 7158 were only detected in the CIN2 sample (Fig. 4B).

Further analyses of the transcripts in HPV16 samples confirmed that integration of the HPV16 viral genome had occurred only in the SCC sample, in agreement with the ISH shown in Fig. 1. However the SCC sample was heterogeneous and only 20% of the cells exhibited overt viral genomic integration signal. Nevertheless, sequences of the transcripts in the SCC sample allowed us to determine the integration sites in both the viral and cellular genomes. Two integration sites were found in the viral genome: one at nt 3505 in E2 and a second at nt 3913 in E5. These integration events gave rise to fusion transcripts between the early HPV16 region and cellular genome, as depicted in Fig. 3B.
which were present in amounts below the threshold used for quantification of the viral transcripts as reported in Table 1 and Figs. 2 and 3.

Minor transcripts

No transcripts coding for E1 were detected in our data set. An E1-encoding transcript has yet to be described for HPV16 and the corresponding transcript in HPV18, grown in vitro in raft cultures, has only recently been identified (Wang et al., 2011). This HPV18 transcript is spliced in E6 and is among the rare transcripts ending with E2, for viral genome amplification (Fig. 1). The absence of E1 transcript is particularly puzzling in the case of the CIN2 lesion where viral DNA replication has only recently been identified (Wang et al., 2011). The absence of E1 transcript is spliced in E6 and is among the rare transcripts ending with E2, for viral genome amplification.

Mapping of viral transcripts of a HPV6b-associated condyloma and comparison with HPV16-associated CIN2

We used the same deep sequencing method to map the viral transcripts in a HPV6b-associated benign condyloma that expressed the E2 protein and exhibited viral DNA amplification (Fig. 1). The HPV6b transcripts map was constructed using previously published data including data on the closely-related HPV11 (Baker and Calef, 1996; Chow et al., 1987; Renaud and Cowsert, 1992). As noted in these previous reports, the most abundant transcript was transcript A, encoding the E1\widehat{E}4 protein, followed by transcript M, which we allocated to E5.

Surprisingly, the total amount of viral RNA extracted from this lesion was ~14-fold greater than from the HPV16-associated CIN2 lesion, which exhibited the highest transcript levels among the HPV16-associated lesions (~1200 total transcripts in CIN2 vs 17000 in condyloma) (Table 1 and Fig. 5). The second unexpected result was that the main early promoter, around nt 90 in front of E6, did not seem to be highly used as mapping of the 5’ ends of the transcripts gave rise to continuous flat readings from an upstream start site at ~nt 16 (Fig. 5B).
steric hindrance via binding of E2, as was demonstrated for HPV16 and HPV18 early promoters (Thierry, 2009) (Fig. 5).

Using the same translation rules as described above for the HPV16 samples, we found that the prominent transcript (A) codes for E1>E4 and represents ~75% of the total viral transcripts, and that the second most abundant transcripts (M), coding for E5 represents ~10% of all transcripts. Interestingly, the two E1 and E2 transcripts were expressed at similarly low abundance of ~2% of total and were initiated at a putative TSS that is not sensitive to E2 repression, which allowed co-expression of the two proteins in the same cells that is consistent with their concerted roles in viral DNA replication. The HPV16-associated CIN2 and HPV6-associated condyloma samples used for deep sequencing show quite comparable levels of expression of both p16 and E2 as well as similar HPV DNA replication in the upper most differentiated layers (Fig. 1). When depth read maps of these samples were compared, transcripts distribution were similar with E1>E4 transcripts being the most abundant (~80%) while E7 is expressed at ~10-fold lower levels (Fig. 5B). However, in contrast with the HPV16-CIN2 the full-length E1 transcript was found in HPV6b-condyloma, while the level of expression of E2 (~2%) was comparable in both types of lesions (Fig. 5A).
Fig. 5. Map of the HPV6b transcripts in the condyloma sample. (A) Analysis of the transcripts was carried out as for Figs. 2 and 3 with additional instruction from published data for the closely-related HPV11 (transcripts A–S) (Baker and Calef, 1996) and description of new transcripts (cuff). (B) Comparison of the HPV6b transcripts quantitative profile with that of HPV16-associated CIN2 where total transcript quantities were equalized for better comparison. The HPV6b-associated condyloma gave rise to a total of ~17,000 transcripts compared to ~600 from HPV16-associated CIN2.
Comparing read depth transcripts in HPV6b-condyloma and HPV16-CIN2 clearly showed that in both cases there are transcripts covering the L2-L1 region, although they were surprisingly relatively less expressed in condyloma than in CIN2 (Fig. 5B). The relative amounts of E6–E7 transcripts were also comparatively low, with a slightly different pattern due to alternative splicing in HPV16 E6 but not in HPV6b. A specific E7 transcript initiated at the E6 putative internal TSS–nt 280, as previously described in an HPV11-associated cell line (Smoatin et al., 1989), was relatively abundant (~6%) compared to the putative E6 transcripts which altogether represented only ~1% of the transcripts (Fig. 5A). In conclusion, all viral genes are transcribed in this HPV6b-associated condyloma, leading to the hypothesis that all viral proteins could be expressed in low-risk HPV-associated lesions (Fig. 5). It will be interesting to compare viral expression in various layers of the lesions in both CIN and condyloma as well as methylation status of the viral genome which is thought to play a crucial role in transcriptional regulation (Vinokurova and von Knebel Doeberitz, 2011; Kalantari et al., 2014) to better understand differential expression of the viral genes and its relationship with epidermal proliferation and differentiation.

Comparison of HPV and cellular transcript abundance

Both HPV and cellular transcripts were detected by sequencing. The abundance of total transcripts followed a similar distribution pattern in CIN2, CIN3, SCC and HPV6b-associated condyloma (Supplementary Fig. S1). In particular, HPV transcript abundance was much higher than the 95% quantile of the entire transcriptome in all four samples (Supplementary Fig. S1). This observation suggested that HPV transcript expression was not minor compared to cellular transcripts, especially in low grade lesion CIN2 and benign condyloma.

Conclusion

We used deep sequencing to construct a quantitative map of HPV transcripts in three HPV16-associated lesions with increasing histological grades: CIN2, CIN3 and SCC, and one benign lesion associated with low-risk HPV6b. Our results show that the HPV16 transcriptome in cervical lesions significantly differs from that proposed by studies in cell lines. Moreover, we uncovered important differences in the transcriptome between the lower grade CIN2, and the higher grade CIN3 and SCC. We identified specific transcripts for the HPV16 early ORFs including E1widehatE4 and E2 transcripts that were more highly expressed in the CIN2 than in the SCC sample. As expected, transcription of E2 found in CIN2 correlated with relatively lower transcription of E7, while this situation was reversed in CIN3 and SCC where E7 transcripts were amplified by ~8–10-times compared to CIN2, and the specific E2 transcript was absent in SCC. We further compared the HPV16 transcriptome in the CIN2 lesion with the transcriptome of low-risk HPV in an HPV6b-associated condyloma. Comparative preponderant (~80%) expression of E1widehatE4 transcripts initiated at the downstream TSS in E7 (nt 670 in HPV16 and nt 715 in HPV6) was evident, thus representing a hallmark of low-grade lesions in both high-risk and low-risk viruses. Interestingly the main E2 transcript was also initiated at the same downstream putative TSS, although less abundantly than E1widehatE4. There have been 3 previous reports on RNAseq of HPV lesions. However the first one looked at a single sample containing HPV58, and did not use splice read mapping nor transcript assembly/abundance software (Li et al., 2013); the second looked for HPV reads without consideration of specific transcripts (Arron et al., 2011); and the third was concerned with viral DNA integration sites in the cellular genome and again not with specific viral transcripts (Khoury et al., 2013). Thus, this report is the first analysis of the structures and relative quantifications of the viral transcripts in patients’ lesions aiming to facilitate interpretation of the regulatory pathways leading to oncogenic progression.

Materials and methods

Ethics statement

The study was approved by institutional review boards of KK Hospital SingHealth (CIRB 2010/517/B). The present study only included adult subjects who provided written informed consent.

Tissue specimens

Patient specimens were obtained from the KK Women’s and Children’s Hospital (Singapore) and fresh biopsies were immediately treated for RNA extraction or fixed in neutral buffered formalin before embedding in paraffin for histology. Diagnosis of specimens was validated by 2 pathologists according to the classification criteria of World Health Organization.

Immunohistochemistry

Consecutive sections of paraffin-embedded tissue samples were labeled with purified anti-HPV16E2C (1/50) and anti-p16Rac4 (clone JC8, Santa Cruz) (1/100) antibodies as previously described (Xue et al., 2010, 2012), in situ hybridizations (ISH) were performed with broad spectrum HPV biotinylated DNA probe sets able to detect 11 types of anogenital HPV (ref: Y1401, Dako), according to the manufacturer’s guidelines. Two different ISH detection kits were used: one for detection of episomal HPV DNA (ref: K0601, Dako) and the other for detection of integrated copies, Genpoint Tyramide Signal Amplification System (ref: K0620, Dako).

DNA extraction and HPV genotyping

DNA extraction was carried out using a column-based protocol with QiAamp DNA FFPE Tissue Kit (QIAGEN, Germany), according to the manufacturer’s instructions. Genotyping was performed using the commercial HPV GenoArray test kit (Hybribio Limited, Hong Kong).

Preparation of RNA samples for sequencing

Total RNA was extracted using the Ambion mirVana miRNA Isolation Kit. The integrity of total RNA was assessed by Agilent Bioanalyzer; only samples with RIN ≥ 7.6 were selected for the analysis. cDNA libraries were generated with 300 ng of total RNA and 2 μl of a 3:1000 dilution of ERCC RNA Spike in Controls (Ambion). The Illumina TruSeq™ RNA sample preparation kit version 2 was used according to the manufacturer’s protocol, with the following modifications: 12 PCR cycles of amplification and the addition of 2 rounds of Agencourt Ampure XP SPRI bead purification (Beckman Couter). The Agilent Bioanalyzer DNA 1000 kit was used to verify the length of the cDNA libraries. The four libraries were sequenced as multiplexes (3 samples/lane) on an Illumina HiSeq2000 using a paired-end run (2 × 51) and the SBS kit v3.

Data analysis

Curation of known transcript annotations

The annotations of known HPV16 and HPV6b transcripts were obtained from in vitro studies (Zheng and Baker, 2006; Baker and
Calef, 1996). However, initial analysis suggested these datasets may be incomplete, so we carried out manual curation of the annotations by searching for reads spanning the splice junctions and examining the donor–acceptor site sequences. We curated the known transcription map of HPV11. We inspected all the known splice sites of HPV11 and discovered that our HPV6b sequencing data did not have read support for any of these splice sites: 1272–3325 (transcript F), 1272–3377 (transcript H), and 3593–5771 (transcript J). In addition, these loci of HPV6b did not have the donor (GU) and acceptor (AG) sequences expected of a splice site. In contrast, our HPV6b data had reads spanning 1278–3325, 1278–3377 and 3596–5789 splice junctions, which do contain the canonical GU donor site and AG acceptor site. Therefore, we changed the 1272–3325 junction to 1278–3325 for transcript F; the 1272–3377 junction to 1278–3377 for transcript H; and the 3593–5771 junction to 3596–5789 for transcript J. The curated annotations of known HPV16 and HPV6b transcripts are given in Supplementary Tables S1 and S2.

Mapping and transcript assembly

Quality check (QC) of raw reads was performed using FASTQC and all libraries passed. A hybrid reference genome, named hg19HPV, was built by combining hg19 and genome sequences of 28 HPV subtypes (Supplementary Table S3) downloaded from NCBI. Similarly, a hybrid gtf file containing annotations of known gene models and transcripts was generated by combining UCSC known human genes and the above curated HPV transcripts. Raw reads were mapped to hg19 HPV using Tophat (Trapnell et al., 2009). Tophat alignment was run allowing novel junction discovery and guided by the annotations of known gene models and transcripts. With the aligned reads, we estimated the relative abundance of transcripts using Cufflinks (Trapnell et al., 2010) guided by annotation of known transcripts and allowing for novel transcript assembly. The FPKM (Fragments Per Kilobase of transcript per Million fragments mapped = [number of fragments][length of transcript in kilo base] / [million mapped fragments]) values calculated by Cufflinks were used as measures of transcript expression abundance. In RNA-Seq, the relative expression of a transcript is proportional to the count of cDNA fragments that originate from it. These counts must be normalized to the length of the transcripts and the sequencing depth of a sample. FPKM provides such normalized count of fragments. Coding potential of the transcripts was predicted by six-frame translation. Read depth graphs were generated with Integrated Genome Browser (Nicol et al., 2009) to visualize the number of reads covering each nucleotide position.

HPV16

Transcripts of HPV16 were first assembled using Cufflinks guided by the curated transcription map (Zheng and Baker, 2006) for CN2, CN3 and SCC independently. Novel transcripts assembled from CN2, CN3 and SCC were then merged using Cuffmerge guided by known transcripts. The merged transcripts comprised of both known and novel transcripts were used as guide to re-run Cufflinks with novel transcript discovery disabled. The results from the re-assembling comprise our final transcript map of HPV16. Detection of viral–cellular fusion transcripts

Given that we had paired-end data, we analyzed the reads for the presence of read-pairs where one read aligns to human genome and another to a viral sequence, as this would be an indication of viral–cellular fusion transcripts. We detected 152 such read-pairs in the SCC sample by mapping against the hg19HPV hybrid reference genome using Bowtie2 (Langmead and Salzberg, 2012). To further define the viral–cellular fusion point to base-pair resolution, we used FusionMap (Ge et al., 2011) to align reads spanning the fusion junctions.

Acknowledgments

We thank Francesca Zolezzi, and the functional Genomics Platform of SfGn and Liew Meng Li Nancy and Sim Kah Cheng for their help in fresh samples collection from the hospitals. We are in debt to Lucy Robinson of insight editing London for careful reading of the manuscript.

Appendix A. Supplementary information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.virol.2014.05.026.

References


Baker, C.C., Calef, C., 1996. Maps of Papillomavirus mRNA Transcripts; Laboratory LAN. Los Alamos National Laboratory, NM.


Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.

Virology 384, 375–379.