The Journal of Molecular Diagnostics, Vol. 21, No. 5, September 2019

ELSEVIER

the Journal of Molecular Diagnostics

jmd.amjpathol.org

Broad-Range Papillomavirus Transcriptome as a Biomarker of Papillomavirus-Associated Cervical High-Grade Cytology

Philippe Pérot,* Anne Biton,† Jacques Marchetta,‡ Anne-Gaelle Pourcelot,§ André Nazac,§ Henri Marret,¶ Thomas Hébert,¶ Delphine Chrétien,* Marie-Christine Demazoin,‖ Michaël Falguières,‖ Laurence Arowas,‖ Hélène Laude,‖ Isabelle Heard,‖ and Marc Eloit*,**

From the Pathogen Discovery Laboratory, Biology of Infection Unit, the Bioinformatics and Biostatistics Hub (C3BI),† and the French Human Papillomavirus Reference Laboratory (Centre National de Référence des Papillomavirus Humains),‖ Institut Pasteur, Paris; the Centre Hospitalier Universitaire,‡ Angers; the Hôpital Le Kremlin-Bicêtre,§ Le Kremlin-Bicêtre; the Centre Olympe de Gouges,¶ Centre Hospitalier Universitaire Bretonneau, Tours; and the National Veterinary School of Alfort,** Paris-Est University, Maisons-Alfort, France*

Accepted for publication April 2, 2019.

Address correspondence to Marc Eloit, D.V.M, Ph.D., Institut Pasteur, 28 rue du Dr Roux, 75015 Paris, France. E-mail: marc.eloit@pasteur.fr.

Human papillomaviruses (HPVs) are responsible for >99% of cervical cancers. Molecular diagnostic tests based on the detection of viral DNA or RNA have low positive predictive values for the identification of cancer or precancerous lesions. Triage with the Papanicolaou test lacks sensitivity; and even when combined with molecular detection of high-risk HPV, this results in a significant number of unnecessary colposcopies. We have developed a broad-range detection test of HPV transcripts to take a snapshot of the transcriptome of 16 high-risk or putative high-risk HPVs in cervical lesions (HPVs 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66, 68, 73, and 82). The purpose of this novel molecular assay, named HPV RNA-Seq, is to detect and type HPV-positive samples and to determine a combination of HPV reads at certain specific viral spliced junctions that can better correlate with high-grade cytology, reflecting the presence of precancerous cells. In a proof-of-concept study conducted on 55 patients, starting from cervical smears, we have shown that HPV RNA-Seq can detect papillomaviruses with performances comparable to a widely used HPV reference molecular diagnostic kit; and a combination of the number of sequencing reads at specific early versus late HPV transcripts can be used as a marker of high-grade cytology, with encouraging diagnostic performances as a triage test. *(J Mol Diagn 2019, 21: 768—781; https://doi.org/10.1016/j.jmoldx.2019.04.010)*

Human papillomavirus (HPV) infections are associated with the development of cervical carcinoma, one of the most common cancers among women, and other cancers like anal cancer[1] and head and neck cancer.[2] HPVs are the etiological agents responsible for >99% of all cervical cancers.[3] HPVs are small, nonenveloped DNA viruses commonly transmitted through sexual contact, which infect basal cells and replicate in the nucleus of squamous epithelial cells. HPVs include >200 genotypes characterized by their oncogenic potential, with highly oncogenic HPV types (high-risk HPV) having a unique ability to drive cell proliferation.[4]

The genomic organization of papillomaviruses is divided into functional early and late regions. The model of HPV infection, which is mainly derived from knowledge on HPV16, is that following the infection of basal cells in the cervical epithelium, the early HPV genes (*E6*, *E7*, *E1*, *E2*, *E4*, and *E5*) are expressed and the viral DNA replicates from the episomal form of the viral DNA. As the cells divide, in the upper layers of the epithelium, the viral genome is replicated further, and the late genes (*L1* and *L2*) and *E4* are expressed. Viral shedding then further initiates new infections.[5]

HPV infection during the development of cervical cancer is associated with a shift from productive infection (which in most of the cases will be cleared by the immune system)

Supported by Institut Pasteur grants InnovDiag and Pasteur Innov.

Disclosures: P.P., A.B., I.H., and M.E. have submitted patent application EP 19 305 934.9 covering the findings of this work.

https://doi.org/10.1016/j.jmoldx.2019.04.010

toward nonproductive persistent and transforming infection (in a minority of cases) characterized in particular by a high level of E6 and E7 mRNAs and low expression of *E2* and late genes such as *L1*.[6,7] High-risk HPV infection may result in low-grade lesions, with highly productive infection and a high rate of spontaneous regression. In contrast, high-risk persistent HPV infection is responsible for high-grade lesions, the true precancerous lesions.

Cervical cancer screening allows detection and treatment of precancerous lesions before the development of cervical cancer. Screening is based on different algorithms, some allowing detection of HPV and others identifying abnormal cells. Despite the role of high-risk HPV in cervical cancer, screening tests of cancer or precancerous lesions remain in many countries, mainly based on the Papanicolaou cytology test and do not include molecular virology tests.[4] This is largely due to the low positive predictive value (PPV) of current molecular tests. Indeed, because most of the current molecular diagnostic methods rely on the detection of HPV genome (DNA) and do not address the patterns of viral expression (RNA), they remain weak predictors of the evolution from low-grade squamous intraepithelial lesion (LSIL) to high-grade squamous intraepithelial lesion (HSIL) of the cervix.[8] In addition, DNA identification of high-risk HPV is not fully predictive of cancer because only persistence for years of high-risk HPV is associated with an increased risk of cancer development.[4] Thus, the use of HPV DNA tests, as a screening assay, is currently increasing worldwide and shows high sensitivity[9] but low PPV for HSIL detection.[10]

HPV RNA tests and, in particular, the detection of E6 and E7 mRNAs of high-risk HPV have been proposed as better molecular markers of cancer development, but *E6* and *E7* are also expressed during HPV transient infection, so it remains difficult to define a threshold of expression associated with the persistence and evolution to high-grade lesions and cancer. There is no consensus that HPV RNA tests have a better diagnostic accuracy compared with HPV DNA tests and cytology for the detection of cervical precancerous lesions.[11–13] There is, therefore, a need for a novel generation of molecular diagnostic tests that not only can detect HPV infection, but also have the ability to accurately predict precancerous stages to offer a better and cost-saving medical benefit.[14–16]

We took advantage of next-generation sequencing (NGS) technologies that now make it possible to study populations of transcripts as a whole, instead of focusing only on one or two specific mRNAs, as done with former techniques, such as quantitative RT-PCR used in HPV RNA tests. A multiplexed amplification system targeting the virus splice junctions, coupled with NGS analysis, was developed, tentatively named HPV RNA-Seq (based on the AmpliSeq technology; Thermo Fisher Scientific, Waltham, MA), which allows describing fine equilibrium among transcript species of 13 high-risk HPVs (HPVs 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, and 66) plus 3 putative high-risk HPVs (HPVs 68, 73, and 82), in a single reaction. This molecular approach, in particular, makes it possible to take a snapshot of the early versus late populations of HPV transcripts and to define a model based on a combination of reads that reflects the biology of the virus, which can then be correlated with the evolution of cervical lesions. The ultimate goal is to replace the current combination of cytology (Papanicolaou smear) and HPV molecular screening by a single molecular test for both the detection of high-risk or putative high-risk HPV and the triage of women at risk of transforming infection, before colposcopy.

In this proof-of-concept study conducted on 55 patients (27 with HSIL and 28 with LSIL), starting from cervical smears conserved at room temperature, we have shown that HPV RNA-Seq can detect papillomaviruses with performances comparable to an HPV DNA-based reference diagnostic kit.[17] Also, a combination of the number of sequencing reads at specific early versus late HPV RNA spliced junctions can be used as a marker of high-grade cytology, with encouraging diagnostic performances as a triage test.

Materials and Methods

Ethics Approval and Consent to Participate

This work was approved by the Institutional Review Board Ile de France 1 and by the French National Agency for the Safety of Medicines and Health Products (ANSM) et des Produits de Santé. The data processing was authorized by the National Commission on Informatics and Liberty (CNIL). Patients provided written informed consent to participate in the study.

Evaluation of Transport Medium for RNA Conservation

HPV16-positive cervical squamous cell carcinoma SiHa cells were cultivated and inoculated at a final concentration of 7×10^4 cells/mL in four transport media: PreservCyt Solution (Hologic, Bedford, MA), NovaPrep HQ+ Solution (Novaprep, Vélizy-Villacoublay, France), RNA Protect Cell Reagent (Qiagen, Hilden, Germany), and NucliSens Lysis Buffer (BioMérieux, Marcy l'Etoile, France). The mixtures were aliquoted in 1-mL tubes and kept at room temperature for 2, 48, 168, 336, and 504 hours. In parallel, 7×10^4 cell pellets without transport medium were kept frozen at −80°C for 2, 48, 168, 336, and 504 hours as a control. At 2, 48, 168, 336, and 504 hours, room temperature aliquots were centrifuged, the medium was removed, and the pellets were frozen at −80°C for a short time (<1 hour) before proceeding with RNA extraction. In the particular case of the NucliSens Lysis Buffer, because the cells were lysed, the entire 1-mL aliquot was frozen at −80°C for a short time without prior centrifugation. For each sample, RNA was extracted using the PicoPure RNA Isolation kit (Thermo Fisher Scientific), together with the corresponding (time match) frozen control, so that all samples have undergone

one freezing cycle. Quantitative RT-PCR (RT-qPCR) was performed to quantify the expression of the two human genes *G6PD* (forward primer, 5′-TGCA-GATGCTGTGTCTGG-3′; and reverse primer, 5′-CGTACTGGCCCAGGACC-3′) and *GAPDH* (forward primer, 5′-GAAGGTGAAGGTCGGAGTC-3′; and reverse primer, 5′-GAAGATGGTGATGGGATTTC-3′) and the expression of the two viral genes HPV16 *E6* (forward primer, 5′-ATGCACCAAAAGAGAACTGC-3′; and reverse primer, 5′-TTACAGCTGGGTTTCTCTAC-3′) and *E7* (forward primer, 5′-GTAACCTTTTGTTGCAAGTGT-GACT-3′; and reverse primer, 5′-GATTATGGTTTCTGA-GAACAGATGG-3′) (Supplemental Figure S1). RNA integrity was assessed on a Bioanalyzer instrument (Agilent Technologies, Santa Clara, CA) (Supplemental Figure S2).

HPV Selection and Splice Site Analysis

HPV reference clones, made available by the International Human Papillomavirus Reference Center (Karolinska University, Stockholm, Sweden), served as reference genomes, except for HPV68, which was retrieved from Chen et al.[18] Accession numbers used in this study were as follows: K02718 (HPV16), X05015 (HPV18), J04353 (HPV31), M12732 (HPV33), X74477 (HPV35), M62849 (HPV39), X74479 (HPV45), M62877 (HPV51), X74481 (HPV52), X74483 (HPV56), D90400 (HPV58), X77858 (HPV59), U31794 (HPV66), KC470267 (HPV68), X94165 (HPV73), and AB027021 (HPV82). HPV genomes were multiply aligned with ClustalW version 2.1 using Geneious[19] version 10. Previously known splice donor (SD) and splice acceptor (SA) sites for HPV16[20] and HPV18[21] were reported on the alignment, and unknown SD and SA sites were predicted manually for the other genotypes by sequence analogy (Figure 1A).

HPV RNA-Seq AmpliSeq Custom Panel

A custom AmpliSeq panel was designed to be used on both PGM and Ion Proton instruments (Thermo Fisher Scientific). Five categories of target sequences were defined as described below.

HPV spliced junctions are a set of target sequences, which are specific HPV splice events, involving a pair of SD and SA sites. The nomenclature includes a spliced junction tag. For example, 31_sp_1296_3295_J43-46 stands for HPV31, splice junction, SD at position 1296 on the HPV31 genome, SA at position 3295 on the HPV31 genome, and junction at position 43 to 46 on the amplicon. The junction coordinates are given in a four-base interval, where the first two bases correspond to the donor part (or left part) and the last two bases correspond to the acceptor part (or right part) of the sequence.

HPV unspliced junctions are a set of target sequences that are specific HPV genomic regions spanning either SD or SA sites, in the absence of any splice event. The nomenclature includes an unspliced junction tag. For example, 31_unsp_1296_1297_J43-46 stands for HPV31, unspliced junction, last base of the left part of the amplicon at position 1296 on the HPV31 genome, first base of the right part of the amplicon at position 1297 on the HPV31 genome, and junction at position 43 to 46 on the amplicon. In this context, the term junction refers to the exon-intron interface (ie, the position where a donor or acceptor site would be found in case of a splice event), and the associated junction coordinates are used to characterize unspliced sequences bioinformatically as described in *Sequencing Data Processing*.

HPV genome away from spliced junctions is a set of target sequences that are specific HPV genomic regions, away from any SD or SA sites. The nomenclature includes a genome away from spliced junctions tag. For example, 45_gen_1664_1794_NoJ stands for HPV45, HPV genomic region, and amplicon coordinates from position 1664 to position 1794 on the HPV45 genome.

HPV-human fusion sequences are a set of hypothesis-driven viral-cellular fusion transcripts, based on previous descriptions.[22–26] For each HPV, 18 fusion sequence candidates involving SA2 or putative breakpoint 1 or 2 (Figure 1B) for the viral part and specific exons from *MYC* or *PVT1* oncogenes for the cellular part were added to the design. For example, 18_fus_929_MYC_001_exon3_J37-40 stands for HPV18, candidate fusion transcript, break/fusion at position 929 on the HPV18 genome, fused with MYC mRNA isoform 001 exon 3, and junction at positions 37 to 40 on the amplicon.

Human sequences are a set of 30 human sequences used as internal controls retrieved from publicly available AmpliSeq projects and representing housekeeping genes (*ACTB*, *B2M*, *GAPDH*, *GUSB*, and *RPLP0*), epithelial markers (*KRT10*, *KRT14*, and *KRT17*), oncogenes, tumor suppressor genes, and direct or indirect downstream effectors of HPV oncoproteins (*AKT1*, *BCL2*, *BRAF*, *CDH1*, *CDKN2A*, *CDKN2B*, *ERBB2*, *FOS*, *HRAS*, *KRAS*, *MET*, *MKI67*, *MYC*, *NOTCH1*, *PCNA*, *PTEN*, *RB1*, *STAT1*, *TERT*, *TOP2A*, *TP53*, and *WNT1*). The nomenclature for these sequences includes a human sequence tag. For example, hg_TOP2A_E21E22 stands for human topoisomerase 2A mRNA exons 21 to 22.

In total, 750 target sequences were included into the panel (Table 1) and can be amplified with a pool of 525 unique primers (Supplemental Table S1). The custom panel is registered under number WG_WG00141 (Ion AmpliSeq Designer; Thermo Fisher Scientific). The average amplicon size of the panel (primers included) is 141 bp (range, 81 to 204 bp). A detailed table, including amplicons names and characteristics along with their corresponding primers and amplicon sequences, is given in the Supplemental Table S1.

Study Participants

Study participants were women, aged from 25 to 65 years, referred for colposcopy consultation in French hospitals.

The patients were referred for colposcopy in the context of an LSIL or an HSIL result at their cytology test performed in accordance with French recommendations regarding the cervical cancer screening program. Patients provided written informed consent, according to French legislation.

Specimen Collection

Genital samples were collected just before performing colposcopy using a cervical sampling device, immersed and rinsed in a vial filled with 20 mL of PreservCyt Solution (Hologic) and sent at room temperature to the HPV National Reference Center at Institut Pasteur (Paris, France). From July 2014 to April 2015, 84 patients were enrolled in the study, coming from three different French centers: Centre Hospitalier Universitaire (CHU) Angers ($n = 66$); CHU Kremlin-Bicêtre ($n = 10$); and CHU Tours ($n = 6$). Samples were removed from the study because of technical reasons (sample leakage, $n = 1$), because of legal issues ($n = 7$), or because they were used for initial technical tests (RNA conservation or RNA extraction and amplification, $n = 4$). The remaining 72 samples (HSIL = 37; LSIL = 35) were processed (Supplemental Table S2).

Data Collection and Availability

The following bioclinical data were collected: date and results of the cytology test, age at the time of the cytology test, and date and results of all available histologic results posterior to colposcopy. As colposcopy was performed in the context of routine health care, biopsies were not performed in case of normal colposcopy. The data sets supporting the conclusions of this article are included within the article and its additional files. The AmpliSeq custom panel is registered under number WG_WG00141. Raw sequencing data are available on the National Center for Biotechnology Information Sequence Read Archive database under BioProject accession number PRJNA525642 (*https://www.ncbi.nlm.nih.gov/sra*).

HPV DNA Detection Using the PapilloCheck Test Kit (HPV DNA)

On reception at the HPV National Reference Center, 16 mL of cytologic sample was transferred into a 50-mL Falcon tube and centrifuged at 4500 × *g* for 10 minutes. The supernatant was removed, and the pellet was washed with 1 mL of phosphate-buffered saline. Sample was then centrifuged again at 5000 × *g* for 10 minutes, and the supernatant was removed. The pellet was frozen at −80°C before DNA extraction. After DNA extraction (Macherey Nagel, Düren, Germany), HPV was detected using the PapilloCheck Test Kit (Greiner Bio-One GmbH, Frickenhausen, Germany), according to the manufacturer's instructions (Supplemental Table S2).

RNA Extraction and Characterization

In parallel to the HPV DNA procedure, 3 × 1-mL aliquots of cytologic specimen were centrifuged at 14,000 rpm for 7 minutes, the supernatant was removed, and the pellet was washed with 1 mL of phosphate-buffered saline. Sample was then centrifuged again at 14,000 rpm for 7 minutes, and the supernatant was removed. The pellet was frozen at −80°C before RNA extraction. RNA samples were extracted using the PicoPure RNA Isolation kit, including on-column DNase treatment, with a final elution volume of 30 μL. Total RNA was quantified on a Nanodrop (Life Technologies, Carlsbad, CA); and RNA integrity was evaluated on a Bioanalyzer RNA 6000 pico chip (Agilent) using the RNA integrity number, a quality score ranging from 1 (strongly degraded RNA) to 10 (intact RNA). For each sample, RT-qPCR, targeting mRNA from housekeeping genes *ACTB* (forward primer, 5′-CATCGAGCACGGCATCGTCA-3′; and reverse primer, 5′-TAGCACAGCCTGGATAGCAAC-3′; amplicon size = 210 bp), and *GAPDH* (forward primer, 5′-GAAGGTGAAGGTCGGAGTC-3′; and reverse primer, 5′-GAAGATGGTGATGGGATTTC-3′; amplicon size = 226 bp), was done in an SYBR Green format with 45 cycles of amplification. Reverse transcription−negative PCR was run to evaluate the presence of residual DNA after RNA extraction (Supplemental Table S2).

Amplification and Sequencing

Starting from RNA, cDNA was generated using the SuperScript III ($n = 17$ samples) or SuperScript IV ($n = 55$ samples) first-strand synthesis system (Thermo Fisher Scientific) with random hexamers and a final RNase H treatment. Libraries were prepared using the Ion AmpliSeq Library Kit 2.0 and the AmpliSeq custom panel WG_WG00141, with 21 cycles of amplification before adapter's ligation. Each sample was barcoded individually. Only positive libraries were sequenced (Supplemental Table S2). In total, 55 clinical samples plus 1 cellular model (SiHa) were sequenced on four Ion Proton runs. Raw data (.fastq files) are available on the National Center for Biotechnology Information Sequence Read Archive database under BioProject accession number PRJNA525642 (*https://www.ncbi.nlm.nih.gov/sra*).

Sequencing Data Processing

Reads were aligned to the reference sequences of the amplicons using STAR[27] version 2.5.3a in local alignment mode (parameter –alignEndsType EndToEnd), by only reporting uniquely mapped reads (–outFilterMultimapNmax 1) and turning off splicing alignment (–alignIntronMax 1). The expression of each amplicon was evaluated by the number of sequencing reads uniquely mapping to their respective sequence (read counts). For reference sequences

containing a splice junction, only reads mapping at the junction site and encompassing at least 10 bases before and 10 bases after the junction were kept. Read counts for each sequence and each sample are provided in Supplemental Table S3.

HSIL Prediction Model

Selection of Amplicons

Read counts were normalized by the size of the library (each read count was divided by a ratio of the library size for a given sample/that of the average library size across samples), and the 215 amplicons capturing splice junctions of the 16 high-risk or putative high-risk HPVs were selected. These amplicons have been annotated with generic names with respect to the type of transcripts they capture, which are shared across HPV species (eg, SD1-SA1) (Figure 1B and Supplemental Table S1). Amplicons capturing splice junctions conserved across the 16 HPV species were summed up, leading to the definition of 18 variables used as predictors in the model. Of the 55 clinical samples, 33 have been selected as presenting enough coverage of these specific amplicons (20 monoinfected and 13 multi-infected samples). The remaining 22 samples of the data set were not used in the logistic regression analysis because they had missing or too low expression signal at spliced junctions for the prediction, reflecting, for example, HPV-negative samples.

Logistic Regression Model

Calling high-grade cytology Y as taking the value 1 for high-grade (HSIL) and 0 for low-grade (LSIL) and a set of amplicons x, a logistic regression model was used to predict the probability that a given observation belongs to the 1 class versus the probability that it belongs to the 0 class. Logistic regression models the log odds of the event (herein, the grade of the cytology) as a function of the predictor variables (herein, the amplicon expression estimated by its read count). Formally, the logistic regression model assumes that the log odds is a linear function of the predictors:

$$\text{logit}(\pi) = \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta' x \qquad (1)$$

where $\pi = P(y = 1|x)$ indicates the probability of the event (being of high grade), β_i are the regression coefficients, and x_i the explanatory variables, in our case the log2 number of reads mapping to the amplicons.

Solving for π, this gives:

$$\pi = \frac{1}{1 + e^{-(\beta_0 + \beta x)}} \qquad (2)$$

Implementation of the Logistic Regression Model

To limit overfitting, L2-norm (ridge) regularization was used, which allows shrinking the magnitudes of the regression coefficients such that they will better fit future data. The logistic model was estimated using the R software version 3.5.1 (*http://www.r-project.org*; last accessed January 29, 2019) package *glmnet* version 2.0-16.[28] Leave-one-out cross validation was used to pick the regularization parameter λ; the one that gives minimum mean cross-validated misclassification error was used. Using λ as the regularization parameter, the model output consisted of an estimate of a coefficient value β for each variable in the logistic regression model. This model was then used to predict the grade of the multi-infected observations, by treating each HPV species separately.

Training Set and Test Set

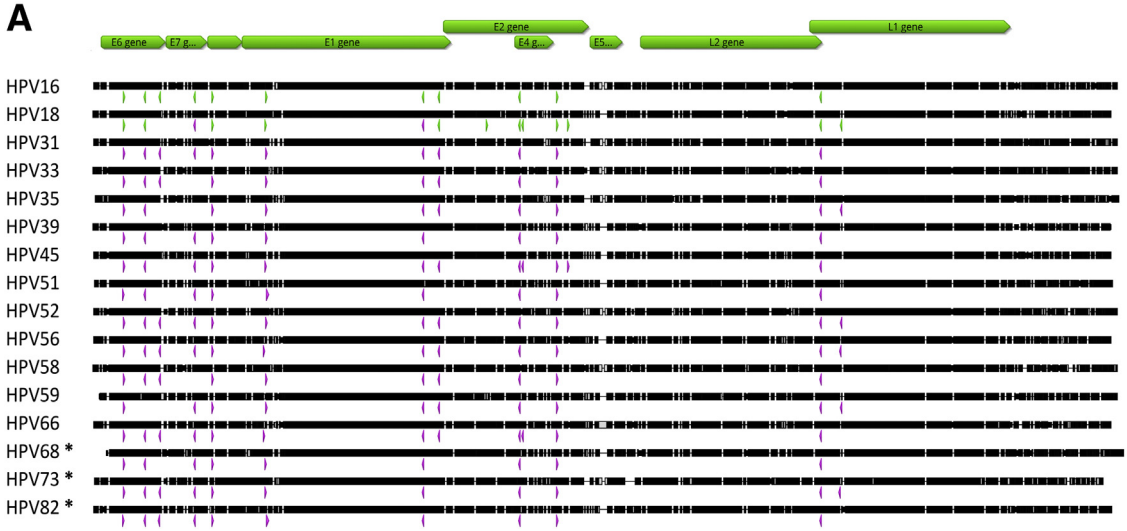
The model was built on the clinical outcome LSIL or HSIL, obtained from the cytologic analysis and estimated on a training set consisting of 20 mono-infected samples (5 LSIL and 15 HSIL) to avoid a confusion bias. It is anticipated that, in the case of multi-infected samples, several HPVs could contribute differently to the progression of the lesion, or to a mix of several grades within the same sample, because they are engaged in different stages of their cycle. The performance of the model was then evaluated on a test set consisting of 13 multi-infected samples. In this case, the set of amplicons of each HPV species was used separately to classify the multi-infected samples, to get one prediction per HPV, as done for the monoinfected samples. For example, if a sample had expression of amplicons from both HPV16 and HPV32, two predictions were given: one using only sequencing reads mapping to HPV16 and one using only sequencing reads mapping to HPV32. Like this, it became possible to interpret the results finely from a virological point of view, as it could be discriminated which HPV was responsible for the lesion.

Results

Evaluation of Transport Medium for RNA Conservation

The stability of total RNA from cervical cells at room temperature was evaluated in four solutions: PreservCyt, the most widely used solution for gynecologic specimen collection; NovaPrep HQ+ Solution, a competitor product used for cells and DNA recovery but never evaluated for RNA conservation; RNA Protect Cell Reagent, a popular solution for RNA stability; and NucliSens Lysis Buffer, a lysis buffer part of the NucliSens automated acid nucleic procedure that has been described as an RNA stabilizer (unpublished data). The amount of spiked HPV16-positive cervical squamous cell carcinoma cells (SiHa) was calibrated to be representative of a cervical smear. After 48 hours at room temperature, RT-qPCR measurement of cellular and viral transcripts showed no or little RNA loss in PreservCyt, only limited RNA degradation (<1 log) in RNA Protect and NucliSens Lysis Buffer, and a marked RNA loss in NovaPrep HQ+ Solution (>2 log) (Supplemental

A



B

5' -	SD1	SA1	SA2	SA3	SD2	SD3	SA4	SA5	put. bkpt	SD4	SA6	SA7	SA8	put. bkpt	SD5	pA early	SD6	SA9	SA10	pA late	- 3'
HPV16	226	409	526	742	880	1301	2581	2708	2869		3357			3619	3631	4218		5637		7265	
HPV18	233	416		791	929	1357	2651	2779	2943	3165	3434	3440	3465	3684	3696	4240	3786	5613	5776	7283	
HPV31	230	413	530	740	877	1296	2518	2646	2807		3295			3578	3590	4143		5552		7232	
HPV33	231	414	531		894	1316	2575	2702	2863		3351			3577	3589	4181		5594		7291	
HPV35	232	415			883	1305	2543	2670	2831		3319			3584	3596	4191		5601	5767	7292	
HPV39	235	418		802	943	1368	2636		2927		3418	3424		3677	3689	4248		5643		7266	
HPV45	230	413		791	929	1357	2610	2737	2901		3392	3398	3423	3648	3660	4233	3750	5608		7315	
HPV51	217	402		751	886	1302	2548		2834		3319			3572	3584	4109		5521		7174	
HPV52	224	407	524	738	879	1301	2569	2696	2857		3345			3613	3625	4231		5643	5810	7344	
HPV56	233	416	533	773	910	1295	2572	2699	2861		3349			3629	3641	4189		5597	5759	7282	
HPV58	232	415	532		898	1320	2579	2706	2867		3355			3596	3608	4214		5643		7301	
HPV59	183			749	887	1306	2571	2698	2862		3353	3359		3615	3627	4228		5606	5769	7260	
HPV66	233	416	533	773	910	1290	2555	2682	2843		3331		3362	3605	3617	4241		5647		7300	
HPV68 *	129	312		697	838	1233	2510		2801		3292			3551	3563	4096		5488		7153	
HPV73 *	227	410	527	727	862	1287	2570		2858		3346			3560	3572	4059		5494	5647	7204	
HPV82 *	222	407	521	753	888	1316	2575		2860		3345			3601	3613	4140		5571		7219	

Figure 1 Transcription maps of 16 HPVs. **A:** Alignment of 16 high-risk or putative high-risk HPV reference sequences, showing splice donor and acceptor sites known previously for HPV16[20] and HPV18[21] (**green marks**) or predicted by sequence homology for other genotypes (**pink marks**). Protein coding genes are indicated on top (**green arrows**). Putative high-risk HPVs are indicated by **asterisks**. **B:** Genomic coordinates of splice donor (SD) and splice acceptor (SA) sites. **Green:** previously known sites.[20,21] **Pink:** predicted sites based on sequence alignment. pA early, polyA signal (early sites; **gray**); pA late, polyA signal (late sites; **gray**); put. bkpt, putative break point for HPV DNA genome integration into the host genome.

Figure S1). After 7 days and up to 21 days, only the PreservCyt solution provided RNA quality with a limited RNA degradation pattern, as indicated by the detection of 18S and 28S rRNA (Supplemental Figure S2). PreservCyt solution was, therefore, used to collect the gynecologic specimens of the study.

HPV RNA-Seq AmpliSeq Custom Panel

Transcriptomic maps known for HPV16[20] and HPV18[21] were used to predict unknown but likely splice donor and splice acceptor sites for HPVs 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66, 68, 73, and 82 (Figure 1). The resulting reconstructed transcripts, as well as HPV genomic sequences, were used as a template for the design of an Ion AmpliSeq panel targeting 16 high-risk or putative high-risk HPVs and named HPV RNA-Seq. Putative break points in HPV genomes, and 30 human cellular genes used as internal controls, were also added to the design. In total, 750 sequences are targeted by a single mix made of 525 unique primers (Table 1 and Supplemental Table S1).

Samples, RNA, and Sequencing

Seventy-two gynecologic samples (HSIL = 37; LSIL = 35) coming from three different French centers (Angers, Kremlin-Bicêtre, and Tours) and collected in PreservCyt solution were processed with RNA extraction using a method designed to

Table 1 HPV RNA-Seq AmpliSeq Custom Panel Contents

Variable	HPV spliced junctions	HPV unspliced junctions	HPV genome away from spliced junctions	HPV-human fusion sequences	Human sequences	
HPV16	14	11	4	18	0	
HPV18	18	12	4	18	0	
HPV31	14	11	4	18	0	
HPV33	13	9	4	18	0	
HPV35	14	10	4	18	0	
HPV39	10	8	4	18	0	
HPV45	14	10	4	18	0	
HPV51	10	9	4	18	0	
HPV52	16	11	4	18	0	
HPV56	16	10	4	18	0	
HPV58	13	8	4	18	0	
HPV59	14	8	4	18	0	
HPV66	15	8	4	18	0	
HPV68*	10	9	4	18	0	
HPV73*	13	10	4	18	0	
HPV82*	11	9	4	18	0	
Human	0	0	0	0	30	
Total	215	153	64	288	30	750

The number of target amplicons is indicated for each category (HPV spliced junctions, HPV unspliced junctions, HPV genome away from spliced junctions, HPV-human fusion sequences, and human sequences) and for each viral and cellular origin.

*Putative high-risk HPV.

recover total RNA from as little as a single cell (PicoPure RNA Isolation kit) (Supplemental Table S2). In most of the cases, total RNA was measurable using a Nanodrop (70/72 positive, average on-positive RNA eluates = 18 ng/μL) and was detectable on a Bioanalyzer pico RNA chip with a pattern indicating RNA degradation (63/72 positive, average RNA integrity number on positive = 2.2). RT-qPCR performed for all samples on *ACTB* mRNA (amplicon size = 210 bp) and *GAPDH* mRNA (amplicon size = 226 bp) indicated that RNA quality was compatible with amplification of 200 to 250 bp size fragments (*ACTB* mRNA average Ct = 27.8; *GAPDH* mRNA average Ct = 30.1). Samples that failed passing this initial RT-PCR quality control were not sequenced. Quantitative PCRs performed after omitting the reverse transcription step were also run and showed in general no or little traces of residual genomic DNA (ACTB DNA average Ct = 38.4; GAPDH DNA average Ct = 35.6). The presence of residual cellular DNA or HPV DNA in RNA preparation is not a major concern because the AmpliSeq assay can differentiate between HPV transcripts and genomic sequences. AmpliSeq libraries were initiated from total RNA and were positive after 21 cycles of amplification for 55 samples (ie, detectable on a Bioanalyzer HS DNA chip). Attempts to add one or two amplification cycles did not bring any significant improvement to the results (data not shown).

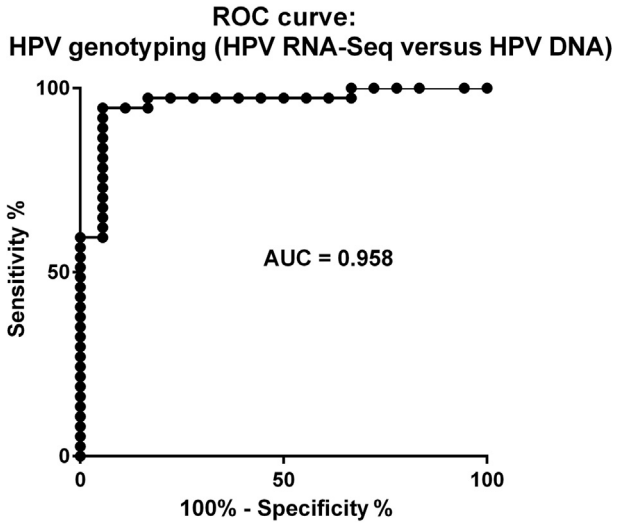


Figure 2 Receiver-operating characteristic (ROC) curve. HPV DNA (PapilloCheck) was used as a reference to evaluate the performances of HPV RNA-Seq for the detection of at least one HPV genotype in a sample. $n = 55$ patients. AUC, area under the curve.

In total, 55 patients (HSIL = 27; LSIL = 28), plus SiHa HPV16-positive cells as a control, were sequenced on Ion Proton. The sequencing reads were aligned to the target sequences, and read counts were generated (Supplemental Table S3). An average of 2.4 million useable reads per sample was reached (minimum = 0.02 million; maximum = 8.36 million), among which an average of 2.1 million reads mapped to the human sequences were used as internal controls (minimum = 0.01 million; maximum = 8.06 million) (Supplemental Table S3). The detection of highly expressed human sequences in all samples, although intersample variations were observed, contributed to validate the sequencing procedure, which is important, especially for the interpretation of HPV-negative samples. Rare nonzero values were also observed for some of the numerous HPV-human fusion sequences that were hypothesized (Supplemental Table S3) but were all false positives,

Table 2 Performances of HPV RNA-Seq for HPV Detection

		HPV DNA			Value, %
		HPV^+	HPV^-	$Se_{(HPV\text{-}DNA)}$	97.3
HPV RNA-Seq	HPV^+	36	3	$Sp_{(HPV\text{-}DNA)}$	83.3
	HPV^-	1	15	$PPV_{(HPV\text{-}DNA)}$	92.3
				$NPV_{(HPV\text{-}DNA)}$	93.8

Performances of HPV RNA-Seq at a threshold value of 150 reads versus HPV DNA (PapilloCheck) for HPV detection. HPV^+ means that at least one HPV genotype is identified in a patient. Data: 55 patients.

NPV, negative predictive value; PPV, positive predictive value; Se, sensitivity; Sp, specificity.

identified as such because only half of the reference sequences were covered by reads.

HPV RNA-Seq Used for HPV Detection and Genotyping

The first application of HPV RNA-Seq is to detect the presence in a given sample of any of the 16 high-risk or putative high-risk HPVs targeted by the panel. The number of reads mapping to HPV-specific amplicons (ie, the sum of categories HPV spliced junctions, HPV unspliced junctions, and HPV genome away from spliced junctions) was used to detect the presence of a given HPV genotype. To help determining a threshold for detection, an HPV DNA test validated for clinical use (PapilloCheck; Greiner Bio-One GmbH) was used as a reference. The best sensitivity and specificity values between the two tests were obtained for a threshold of 100 to 200 reads (Figure 2). For example, a threshold value of 150 reads resulted in a sensitivity [$Se_{(HPV\text{-}DNA)}$] of 97.3% and a specificity [$Sp_{(HPV\text{-}DNA)}$] of 83.3%, leading to a $PPV_{(HPV\text{-}DNA)}$ of 92.3% and a negative predictive value [$NPV_{(HPV\text{-}DNA)}$] of 93.8% for detecting high-risk HPV in this population composed of approximately 50% of HSIL and 50% of LSIL (Table 2, with raw data in Supplemental Tables S2 and S3).

A more detailed view of the genotypes identified by both techniques is given in Figure 3. The number of monoinfected, multi-infected, or HPV-negative samples identified by the two tests is summarized in Table 3. Because the HPV DNA test can detect the 16 high-risk or putative high-risk HPVs captured by HPV RNA-Seq (HPVs 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66, 68, 73, and 82) plus 8 additional low-risk HPVs (HPVs 6, 11, 40, 42, 43, 44/45, 53, and 70), the comparison was based only on the 16 HPVs common to both tests. Using a threshold value of 150 reads, HPV RNA-Seq detected two more positive patients than the HPV DNA test ($n = 39$ versus $n = 37$) (Table 2). HPV RNA-Seq identified the presence of more than one HPV for three more patients than the HPV DNA test ($n = 13$ versus $n = 10$ multi-infected samples) (Table 3). Globally, HPV16 was found at a slightly weaker occurrence by HPV RNA-Seq ($n = 18$ versus $n = 19$) in favor of other genotypes, such as HPV 31, 33, 45, 52, 56, 58, or 66, which were less commonly found by the HPV DNA test (HPV31, $n = 5$ versus $n = 4$; HPV33, $n = 3$ versus $n = 1$; HPV45, $n = 3$ versus $n = 2$; HPV52, $n = 5$ versus $n = 3$; HPV56, $n = 4$ versus $n = 2$; HPV58, $n = 5$ versus $n = 4$; HPV66, $n = 2$ versus $n = 1$) (Figure 3). Apart from HPV16, only HPV51 was less frequently found by HPV RNA-Seq than by HPV DNA ($n = 2$ versus $n = 3$). The cellular model (SiHa) gave only HPV16 signal in both tests, as expected (Supplemental Table S3).

HPV RNA-Seq Used as a Marker of High-Grade Cytology

An exploratory analysis was conducted on 20 of the monoinfected samples in which HPV RNA spliced junctions could be used to predict high-grade cytology. Amplicons capturing splice junctions were studied to ensure detection of HPV transcripts. However, the number of monoinfected samples ($n = 20$) used as the training set was small, in particular the number of samples of LSIL ($n = 5$). In this configuration, a fully accurate variable selection (ie, to

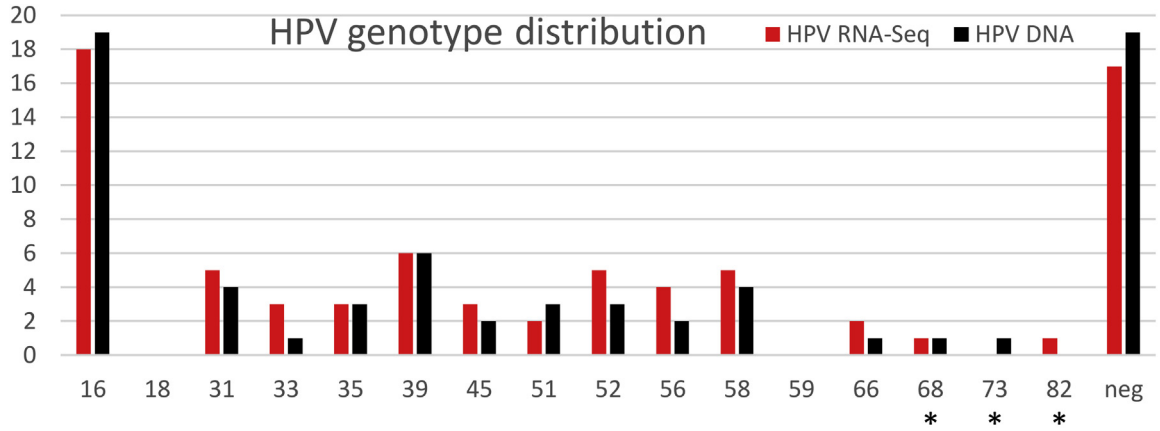


Figure 3 Comparison of the number of HPV genotypes identified by HPV RNA-Seq and HPV DNA. Vertical bars represent the number of HPV genotypes identified by HPV RNA-Seq at a threshold value of 150 reads (red) versus HPV DNA (PapilloCheck; black). Putative high-risk HPVs are indicated by **asterisks**. $n = 55$ patients. neg, negative.

Table 3 Comparison of HPV RNA-Seq and HPV DNA for the Classification of Samples

Sample	HPV RNA-Seq	HPV DNA
Monoinfected, *n*	26	27
Multi infected, *n*	13	10
HPV negative, *n*	16	18

Number of monoinfected, multi-infected, and HPV-negative samples by HPV RNA-Seq at a threshold value of 150 reads, versus HPV DNA (PapillоCheck). $n = 55$ patients.

select the strict minimum number of amplicons that were necessary for HSIL versus LSIL prediction and set the others to zero coefficient) was not feasible. In addition, overfitting, as using only five LSIL and 15 HSIL samples, did not allow capturing the diversity of the whole population. Leave-one-out cross validation was used to pick the λ giving the minimum cross-validated error using ridge regularization. $\lambda = 0.08$ gave a mean cross-validated error of 15%. A 20% prediction error was also computed using nested cross validation. This error rate can be seen as an indicator of how the model could fit future data sets. The corresponding parameter was used to fit a regularized logistic regression model, assigning a coefficient to each amplicon (Table 4) and a probability of being of high grade to each sample (Table 5). The grade of the 20 monoinfected samples was classified correctly, except for one observation (Table 5). This unique misclassified sample (IonXpress_019_2613), which was classified LSIL by the cytologic analysis, was further found as containing a mixture of LSIL and HSIL lesions after histologic examination performed >1 year after the sampling done for HPV RNA-Seq/cytology.

Table 4 Coefficients of the (Ridge) Logistic Regression

Junction	Coefficient	Transcript category	Transcript contents
Intercept	0.468298365	NA	NA
SD2-SA10	−0.693322203	Late	L1
SD3-SA4	0.545728771	Early	(E1) E4 E5
SD1-SA4	0.387642812	Early	(E6) E2 E5
SD2-SA4	−0.262522618	Early	(E7) E2 E5
SD1-SA2	0.146954179	Early	E6 E7
SD2-SA5	0.12050536	Early	(E7) E2 E5
SD1-SA6	0.107204358	Early	(E6) E4 E5
SD5-SA10	0.096088118	Late	L1
SD3-SA6	0.093052957	Early	(E1) E4 E5
SD1-SA5	0.092877361	Early	(E6) E2 E5
SD2-SA6	−0.088655106	Early	(E7) E4 E5
SD1-SA1	0.07669912	Early	E6 E7
SD1-SA3	0.069688722	Early	E6 E7
SD2-SA8	0.061867993	Early	(E7) E4 E5
SD3-SA5	0.051702326	Early	(E1) E4 E5
SD2-SA9	−0.040972141	Late	L1
SD5-SA9	−0.026083777	Late	L1
SD3-SA8	0	Early	(E1) E4 E5

The first and fourth columns give the identification of the splice junction captured by the amplicon, the second column gives the coefficient assigned by the logistic regression, and the third column indicates whether the splice junction comes from a late or an early transcript.

NA, not applicable; SA, splice acceptor; SD, splice donor.

The estimated model was then used to classify the 13 multi-infected samples, with each HPV species present within one sample being classified individually for its implication in HSIL development. If at least one HPV species gave an HSIL prediction, the sample was considered to be HSIL. Performances for HSIL prediction were calculated for all samples, considered as not being of high grade, both the 6 samples without sufficient coverage of the splice junctions and the 16 HPV-negative samples not exceeding the threshold of HPV detection. The calculated performances for HSIL prediction in comparison to cytology for the 55 patients (monoinfected, multi-infected, and HPV-negative patients) were as follows: $Se_{(cyto)} = 66.7\%$, $Sp_{(cyto)} = 85.7\%$, $PPV_{(cyto)} = 81.8\%$, and $NPV_{(cyto)} = 72.7\%$ (Table 6). The performances were also calculated for the subset of 39 samples having at least one high-risk (HR+) HPV identified by HPV RNA-Seq, giving in this case $Se_{(cyto/HR+)} = 94.7\%$, $Sp_{(cyto/HR+)} = 80.0\%$, $PPV_{(cyto/HR+)} = 81.8\%$, and $NPV_{(cyto/HR+)} = 94.1\%$ (Table 6). The ratio of HSIL/LSIL remained similar between these two populations (approximately 1:1), making the comparison of the PPV and the NPV possible. Finally, a summary of the results for HPV detection and genotyping (HPV RNA-Seq versus HPV DNA) and high-grade cytology prediction (HPV RNA-Seq versus cytology), including posterior histologic data of cervix biopsies when available, is presented in Table 7.

HPV RNA-Seq Used as a Triage Test

The performances of HPV RNA-Seq as a triage test were evaluated using histology as gold standard. Results from histologic examination were, however, not available for all patients. The time interval separating HPV RNA-Seq/cytology tests from histologic analysis, varying between 0 and 780 days, was another limitation in this study. To try to overcome these drawbacks, the performances of HPV RNA-Seq versus histology were compared with the performances of cytology versus histology, considering all available samples, only samples for which histology was performed <3 months after HPV RNA-Seq/cytology, or only samples for which histology was performed <6 months after HPV RNA-Seq/cytology (Supplemental Table S4). In addition and for each category, distinction was made between the performances obtained when HPV RNA-Seq HPV-positive and HPV-negative patients were grouped together or when only HPV-positive patients were considered (Supplemental Table S4). Calculation of the PPV as a function of HSIL prevalence in the population was also performed (Supplemental Figure S3 and Supplemental Table S4).

Table 5 Classification Results of the (Ridge) Logistic Regression

Sample ID	HSIL predicted probability	Predicted grade	Cytology grade	Agreement
IonXpress_039_115	0.115	LSIL	LSIL	True
IonXpress_033_730	0.204	LSIL	LSIL	True
IonXpress_038_114	0.259	LSIL	LSIL	True
1492	0.425	LSIL	LSIL	True
IonXpress_019_2613	0.562	HSIL	LSIL	False
IonXpress_027_598	0.653	HSIL	HSIL	True
729	0.716	HSIL	HSIL	True
567	0.718	HSIL	HSIL	True
IonXpress_018_2439	0.902	HSIL	HSIL	True
610	0.904	HSIL	HSIL	True
1066	0.911	HSIL	HSIL	True
IonXpress_034_758	0.919	HSIL	HSIL	True
1122	0.934	HSIL	HSIL	True
25	0.944	HSIL	HSIL	True
IonXpress_037_1267	0.947	HSIL	HSIL	True
IonXpress_024_26	0.965	HSIL	HSIL	True
IonXpress_025_538	0.97	HSIL	HSIL	True
752	0.976	HSIL	HSIL	True
IonXpress_021_443	0.984	HSIL	HSIL	True
2612	0.993	HSIL	HSIL	True

The first column gives the sample identification, the second column gives the probability estimate that the sample is HSIL, the third and fourth columns give the corresponding prediction, and the fifth column contains true if the prediction is consistent with the grade evaluated by cytology.

HSIL, high-grade squamous intraepithelial lesion; LSIL, low-grade squamous intraepithelial lesion.

Discussion

We have developed a highly multiplexed RT-PCR assay coupled with NGS (HPV RNA-Seq) combining HPV detection and genotyping together with predicting high-grade cytology, starting from cervical specimens conserved at room temperature. A pilot study was conducted on 55 patients.

The performances of HPV RNA-Seq used as an HPV detection and genotyping assay were evaluated in comparison to the HPV DNA PapilloCheck kit (HPV DNA; Greiner Bio-One GmbH), which is officially approved for clinical use. A good concordance of the results was observed between the two assays (area under curve $>$ 0.95) (Figure 2). A positive threshold of 150 reads resulted in high sensitivity and negative predictive value of HPV RNA-Seq [$Se_{(HPV\text{-}DNA)}$ = 97.3%, $NPV_{(HPV\text{-}DNA)}$ = 93.8%] (Table 2), along with a relatively high but lower specificity and positive predictive value [$Sp_{(HPV\text{-}DNA)}$ = 83.3%, $PPV_{(HPV\text{-}DNA)}$ = 92.3%] (Table 2) linked to the identification of additional genotypes by HPV RNA-Seq not detected by HPV DNA. Because cervical samples were split before independent extractions of RNA (HPV RNA-Seq) and DNA (HPV DNA), the few differences observed between the two tests can reflect a nonhomogeneous distribution of infected cells. Also, PapilloCheck, like other HPV DNA tests, is not 100% accurate,[29,30] so it remained difficult to identify potential false-positive results of HPV RNA-Seq versus better sensibility. For example, three patients were classified as HPV negative by PapilloCheck but not by HPV RNA-Seq. The number of RNA-Seq reads associated with HPV species in these three potential false HPV-positive patients was close to the limit of detection for some of them ($\leq$400 reads) but not for all (eg, 39,527 reads mapped to HPV58 for sample 2065) (Supplemental Table S3). The calculated sensitivity and specificity may, therefore, not reflect optimally the added value of HPV RNA-Seq. These limitations are common for any novel diagnostic test when compared with older references.

Effective cervical cancer screening requires high Se and NPV for high-risk HPV detection, as women with a

Table 6 Performances of HPV RNA-Seq for the Prediction of High-Grade Cytology

n = 55	Cytology			Value, %
HPV RNA-Seq	HSIL	LSIL	$Se_{(cyto)}$	66.7
HSIL	18	4	$Sp_{(cyto)}$	85.7
Not HSIL*	9	24	$PPV_{(cyto)}$	81.8
			$NPV_{(cyto)}$	72.7
n = 39	Cytology			
HPV RNA-Seq HR+	HSIL	LSIL	$Se_{(cyto/HR+)}$	94.7
HSIL	18	4	$Sp_{(cyto/HR+)}$	80.0
Not HSIL*	1	16	$PPV_{(cyto/HR+)}$	81.8
			$NPV_{(cyto/HR+)}$	94.1

Performances of HPV RNA-Seq versus cytology for HSIL detection for both the 55 total patients and for the subset of 39 patients having at least one HPV identified by HPV RNA-Seq.

*Either no HPV was detected in the sample by HPV RNA-Seq or none of the HPV genotypes detected were given HSIL prediction.

HR, high-risk; HSIL, high-grade squamous intraepithelial lesion; LSIL, low-grade squamous intraepithelial lesion; NPV, negative predictive value; PPV, positive predictive value; Se, sensitivity; Sp, specificity.

Table 7 HPV Detection and Genotyping and HSIL Prediction for the 55 Clinical Samples

		HPV RNA-Seq							
		Genotyping	Marker of HSIL						
		Per patient	Per HPV			Per patient			
Sample name	HPV DNA	Detection	Not enough coverage on splice junctions	Not HSIL	HSIL	Prediction	Cytology	Histology	Time between cytology-histology, days
D-15-0041_1066_BC13	16	16			16	HSIL	HSIL	HSIL	55
D-15-0041_1122_BC14	16	16			16	HSIL	HSIL	HSIL	130
D-15-0041_1124_BC5	16, 39	16, 39	39	16		Not HSIL	LSIL	HSIL	70—434
D-15-0041_1490_BC6	16, 39	16, 35, 39		39	16, 35	HSIL	LSIL	HSIL	67
D-15-0041_1492_BC7	16	16		16		Not HSIL	LSIL	LSIL	81
D-15-0041_151_BC15	16, (53)	16			16	HSIL	LSIL	HSIL	130
D-15-0041_152_BC16	16, (42)	16, 52, 82	16, 52, 82			Not HSIL	LSIL	LSIL	41
D-15-0041_2209_BC11	16, (42), 52	16, 39, 52	39	16, 52		Not HSIL	LSIL	HSIL	ND
D-15-0041_250_BC12	16, 39, (42)	16, 39		16, 39		Not HSIL	LSIL	LSIL	55
D-15-0041_25_BC4	16	16			16	HSIL	HSIL	HSIL	75
D-15-0041_2612_BC8	16	16			16	HSIL	HSIL	ND	ND
D-15-0041_567_BC9	16	16			16	HSIL	HSIL	HSIL	ND
D-15-0041_610_BC2	16	16			16	HSIL	HSIL	HSIL	113
D-15-0041_729_BC3	16	16			16	HSIL	HSIL	HSIL	59
D-15-0041_752_BC10	16	16			16	HSIL	HSIL	HSIL	444
IonXpress_017_2437	(43), 51	51		51		Not HSIL	LSIL	LSIL	195
IonXpress_017_251	Negative	Negative				Not HSIL	HSIL	LSIL	85
IonXpress_018_2439	58	58			58	HSIL	HSIL	LSIL	164
IonXpress_018_440	Negative	Negative				Not HSIL	LSIL	LSIL	38
IonXpress_019_2613	16	16			16	HSIL	LSIL	HSIL	416-780
IonXpress_020_3137	(53)	56	56			Not HSIL	HSIL	HSIL	350
IonXpress_021_10	56, (44/55)	56		56		Not HSIL	LSIL	ND	130
IonXpress_021_443	58	33, 58	33		58	HSIL	HSIL	LSIL	99
IonXpress_022_23	Negative	Negative				Not HSIL	HSIL	HSIL	ND
IonXpress_022_444	16, 33	16, 33		33	16	HSIL	HSIL	HSIL	69
IonXpress_023_24	(6), (11), (53)	Negative				Not HSIL	HSIL	HSIL	0—13
IonXpress_023_536	Negative	Negative				Not HSIL	LSIL	LSIL	101
IonXpress_024_26	45	45			45	HSIL	HSIL	HSIL	106
IonXpress_024_537	Negative	Negative				Not HSIL	LSIL	LSIL	71
IonXpress_025_457	Negative	Negative				Not HSIL	LSIL	LSIL	278
IonXpress_025_538	35	31, 35	31		35	HSIL	HSIL	HSIL	191
IonXpress_026_539	Negative	Negative				Not HSIL	LSIL	ND	ND
IonXpress_026_565	16	Negative				Not HSIL	HSIL	HSIL	65
IonXpress_027_598	31	31			31	HSIL	HSIL	HSIL	52
IonXpress_028_609	35, 52	52			52	HSIL	LSIL	HSIL	83
IonXpress_029_611	Negative	Negative				Not HSIL	HSIL	ND	ND
IonXpress_030_612	Negative	Negative				Not HSIL	LSIL	LSIL	113
IonXpress_031_613	35, 39, (44/55)	35, 39		35, 39		Not HSIL	LSIL	LSIL	83
IonXpress_032_728	Negative	Negative				Not HSIL	HSIL	HSIL	59
IonXpress_033_730	31	31		31		Not HSIL	LSIL	HSIL	211—575
IonXpress_034_758	58	58			58	HSIL	HSIL	HSIL	43
IonXpress_035_1150	16, 39, 52	16, 39, 52		52	16, 39	HSIL	HSIL	HSIL	125
IonXpress_036_1151	(11), 31	31			31	HSIL	HSIL	HSIL	125
IonXpress_036_98	(42)	Negative				Not HSIL	LSIL	ND	20
IonXpress_037_100	Negative	Negative				Not HSIL	LSIL	LSIL	57
IonXpress_037_1267	45	45			45	HSIL	HSIL	LSIL	71
IonXpress_038_114	31	31		31		Not HSIL	LSIL	HSIL	154
IonXpress_038_1597	Negative	Negative				Not HSIL	HSIL	HSIL	85
IonXpress_039_115	56	56		56		Not HSIL	LSIL	LSIL	34

(*table continues*)

Table 7 (*continued*)

		HPV RNA-Seq							
		Genotyping	Marker of HSIL						
		Per patient	Per HPV			Per patient			
Sample name	HPV DNA	Detection	Not enough coverage on splice junctions	Not HSIL	HSIL	Prediction	Cytology	Histology	Time between cytology-histology, days
IonXpress_039_1598	Negative	Negative				Not HSIL	HSIL	LSIL	115
IonXpress_041_1650	66, (70)	56, 66	56, 66			Not HSIL	LSIL	LSIL	115
IonXpress_043_1871	51, 58, 68, 73	33, 51, 58, 68	33	51, 58, 68		Not HSIL	LSIL	LSIL	101
IonXpress_044_2064	39, 51	45	45			Not HSIL	LSIL	HSIL	129
IonXpress_045_2065	Negative	52, 58	52, 58			Not HSIL	LSIL	LSIL	160
IonXpress_046_2066	(6)	66	66			Not HSIL	LSIL	HSIL	99

HPV genotypes included in the scope of the HPV DNA test (PapilloCheck) but not in HPV RNA-Seq are indicated in parentheses. For each genotype identified by HPV RNA-Seq, a classification is given: not enough coverage on splice junctions (no prediction was possible for the genotype), not HSIL, or HSIL. When at least one HPV was given a high-grade signature, the patient's prediction was set as HSIL. Conversely, a final not HSIL means that either no HPV was detected in the sample or none of the HPV genotypes detected were given HSIL prediction.

HSIL, high-grade squamous intraepithelial lesion; LSIL, low-grade squamous intraepithelial lesion; ND, not done.

negative HPV test result are usually tested again only after several years. The positive threshold for HPV genotyping was set at 150 reads in this study because it optimized both Se and Sp values, but lowering this threshold to maximize the sensitivity remains possible. Such adjustments will be possible after the study of larger cohorts.

As a second application of HPV RNA-Seq, as a triage test, a logistic regression model for the prediction of high-grade cytology was built on the basis of a combination of the number of reads captured at specific HPV RNA spliced junctions, using the grade found by cytology as a reference. This evaluation was conducted in a population of women with LSIL or HSIL cytology results. Where at least one HPV was given a high-grade signature, the patient's prediction was set as HSIL. Conversely, not HSIL was used when either no HPV was detected in the sample (threshold of 150 reads) or none of the genotypes detected by HPV RNA-Seq were given high-grade prediction (absence of detectable transcripts). Not HSIL rather than LSIL terminology was used because the protocol did not allow the comparison of the HPV-DNA positive samples evaluated as LSIL with the ones evaluated as normal in cytology. Also, because there is a possibility that cervical lesions could, in some rare cases, originate from causes other than HPV infections, the use of not HSIL instead of LSIL in the case of HPV-negative samples seemed more appropriate.

As far as the comparison with cytology could be used as a benchmark, when the 55 patients were considered (including monoinfected, multi-infected, and HPV-negative patients), the number of HSIL predicted by HPV RNA-Seq ($n = 22$) was lower than the number of HSIL identified by cytology ($n = 27$), resulting in $Se_{(cyto)} = 66.7\%$, $Sp_{(cyto)} = 85.7\%$, $PPV_{(cyto)} = 81.8\%$, and $NPV_{(cyto)} = 72.7\%$ (Table 6). Interestingly, when only HPV RNA-Seq HPV-positive samples were considered, the $PPV_{(cyto/HR+)}$ for the detection of high-grade lesions remained unchanged but the $Se_{(cyto/HR+)}$ and the $NPV_{(cyto/HR+)}$ increased to 94.7% (+28.0%) and 94.1% (+21.4%), respectively, with the number of HSIL predicted by HPV RNA-Seq ($n = 22$) becoming superior to the number of HSIL identified by cytology ($n = 19$) (Table 6). In this case, the only one patient identified HSIL by cytology but not predicted HSIL by HPV RNA-Seq (sample IonXpress_020_3137) was found HSIL by the histologic examination performed 350 days later, which opens the possibility that this sample might be positive if the patient would be tested again by HPV RNA-Seq at a date closer to the histologic examination.

In clinical use, after primary screening for high-risk HPV in the general population, a triage test with high Sp and PPV is needed for the triage of women at risk of transforming infection before colposcopy. In countries that have adopted HPV DNA as a screening test, cytologic analysis can be used for the triage of women at risk because cytology has better Sp and PPV than HPV DNA tests.[31,32] In line with that, the $PPV_{(cyto)} = 81.8\%$ of HPV RNA-Seq outperformed HPV DNA and other RNA tests, whose PPV as triage assays never exceeded 50% in a population of women referred for colposcopy composed with a similar 1:1 ratio of HSIL/LSIL.[10]

Therefore, the added value of HPV RNA-Seq was evaluated over cytology for the triage of women at risk of developing cervical cancer. Histology was used as the gold standard for the diagnosis of cervical lesions. However, an inherent limitation of this work was that histology was not concomitant with the sampling performed for HPV RNA-Seq/cytology tests, which means that by the time histology was performed (between 0 and 780 days after initial sampling), the lesion could have evolved spontaneously in one direction (LSIL to HSIL) or another (HSIL to LSIL). To

help clarify this point, the performances of HPV RNA-Seq versus histology were compared with the performances of cytology versus histology, considering different categories of samples (Supplemental Table S4). Remarkably, whatever the category considered, the Sp of HPV RNA-Seq versus histology was always greater than or equal to the Sp of cytology versus histology (0.0 to 11.1), and the resulting PPV of HPV RNA-Seq versus histology was always greater than the PPV of cytology versus histology (2.4 to 7.4 in this population that reflects other studies[9]) (Supplemental Table S4). Calculation of the PPV as a function of HSIL prevalence allowed anticipating a ΔPPV between HPV RNA-Seq versus histology and cytology versus histology. For example, in the case in which the ratio of HSIL/LSIL would tend to 1:2, as seen elsewhere,[10] the ΔPPV could be up to 10.4 in favor of HPV RNA-Seq versus histology (range, 4.4 to 10.4) (Supplemental Figure S3 and Supplemental Table S4). This observation constitutes a solid argument in favor of a potential added medical value of HPV RNA-Seq over cytology, although studies on larger cohorts are now required. Another observation is that the Se of HPV RNA-Seq versus histology was always higher on the subset of HPV-positive patients (12.0 to 33.3) (Supplemental Table S4), similarly to the evaluation performed with cytology taken as reference (Table 6). Last, the Sp of HPV RNA-Seq versus histology increased on the subset of patients for whom histology was performed <3 months after sampling (5.0 to 8.3) (Supplemental Table S4) but decreased on the subset of patients for whom histology was performed <6 months after sampling (−1.7 to −2.3) (Supplemental Table S4), which may indicate that some lesions have evolved in the meantime.

Although the minimum number of reads required for the assay was not evaluated, our observations tend to support that ≤1 million reads per sample is enough for performing HPV genotyping, but more would be needed for the detection of HPV transcripts. The absence of detectable transcripts for a given HPV was assimilated to the absence of HPV transcripts in the sample, which may not be true if sequencing depth was insufficient. Generally speaking, the questions of the format of the test and of the model of use are of importance in the perspective of deploying NGS-based *in vitro* diagnostic tests. AmpliSeq, a former product by Thermo Fisher Scientific developed for Ion Proton and PGM instruments, has been transferred in 2018 to Illumina (San Diego, CA) and is now fully compatible with Illumina sequencers. In a decentralized laboratory model, it becomes possible that four to six samples could run on a benchtop iSeq100 sequencer for a cost per sample of approximately $200, with RNA extraction, quality controls, and data analysis included (salaries and equipment excluded). In a more centralized view in which all regional samples would converge to one laboratory, the use of production-scale sequencers, such as the HiSeq or NovaSeq instruments, could allow multiplexing up to 381 samples per run, potentially reducing the cost per sample to approximately $10 to $20 and thus making NGS-based assays competitive over PCR-based tests. Another point is that not all sequences of HPV RNA-Seq contributed equally to the result, with some of them giving useless or redundant information, suggesting that the format of the test can also evolve to keep only the most informative target sequences, while potentially reducing the depth of sequencing required for analysis and the associated costs. For example, a reduction in the number of human sequences used as internal controls could be considered.

HPV RNA-Seq will be further developed and validated as a companion test in HPV DNA-positive patients or when the result of cytology is uncertain, to allow focusing the colposcopies to the most relevant patients. It has recently been shown that only one third of women recommended for colposcopy after primary HPV testing (DNA) and cytology actually had HSIL.[9] By increasing the positive predictive value in detecting HSIL, HPV RNA-Seq could significantly increase the medical benefit/cost ratio of colposcopies. The case of atypical squamous cells of undetermined significance would also constitute an important patient's category to demonstrate an added value of the assay. Once the performances and the medical benefit have been evaluated on large cohorts, such a broad-range genotype papillomavirus transcriptome assay could eventually replace first-line cytology and DNA-based tests, by providing in a single procedure both HPV detection and genotyping together with a molecular marker of high-grade lesions. Other diagnostic applications in HPV-associated anogenital or head and neck cancers can also be envisioned.

In conclusion, HPV RNA-Seq can provide a second-line test in HPV-positive patients to reduce unnecessary colposcopies and even be used as a two-in-one test combining HPV typing with triage capabilities. HPV RNA-Seq is minimally invasive and is convenient for sample conserved at room temperature. The assay will now require further clinical validation in larger cohorts.

Acknowledgments

We thank the Clinical Core of the Center of Translational Science (Institut Pasteur) for the management of all legal and ethical aspects of the study; PathoQuest (Paris, France) and the Center for Translational Science/Cytometry and Biomarkers Unit of Technology and Service (Institut Pasteur) for providing access to the sequencers; and Novaprep for kindly providing NovaPrep HQ+ Solution.

M.E. and P.P. designed the study; P.P. designed the AmpliSeq custom panel; J.M., A.G.P., A.N., H.M., and T.H. collected the gynecologic samples; M.F., M.C.D., L.A., H.L., and I.H. managed the samples at the HPV National Reference Center; I.H. contributed to the clinical protocol; M.C.D. and M.F. provided SiHa cells; P.P. and D.C. performed the sequencing experiments and analyzed the data; A.B. performed the biostatistical analysis; P.P., M.E., and A.B. wrote the manuscript; M.E. supervised the study; all authors read and approved the final manuscript.

Supplemental Data

Supplemental material for this article can be found at *http://doi.org/10.1016/j.jmoldx.2019.04.010*.

References

1. Lin C, Franceschi S, Clifford GM: Human papillomavirus types from infection to cancer in the anus, according to sex and HIV status: a systematic review and meta-analysis. Lancet Infect Dis 2018, 18: 198–206
2. Chaturvedi AK, Engels EA, Pfeiffer RM, Hernandez BY, Xiao W, Kim E, Jiang B, Goodman MT, Sibug-Saber M, Cozen W, Liu L, Lynch CF, Wentzensen N, Jordan RC, Altekruse S, Anderson WF, Rosenberg PS, Gillison ML: Human papillomavirus and rising oropharyngeal cancer incidence in the United States. J Clin Oncol 2011, 29:4294–4301
3. Walboomers JM, Jacobs MV, Manos MM, Bosch FX, Kummer JA, Shah KV, Snijders PJ, Peto J, Meijer CJ, Muñoz N: Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. J Pathol 1999, 189:12–19
4. Schiffman M, Doorbar J, Wentzensen N, de Sanjosé S, Fakhry C, Monk BJ, Stanley MA, Franceschi S: Carcinogenic human papillomavirus infection. Nat Rev Dis Primer 2016, 2:16086
5. Woodman CBJ, Collins SI, Young LS: The natural history of cervical HPV infection: unresolved issues. Nat Rev Cancer 2007, 7:11–22
6. Doorbar J, Quint W, Banks L, B IG, Stoler M, Broker TR, Stanley MA: The biology and life-cycle of human papillomaviruses. Vaccine 2012, 30 Suppl 5:F55–F70
7. Shulzhenko N, Lyng H, Sanson GF, Morgun A: Ménage à trois: an evolutionary interplay between human papillomavirus, a tumor, and a woman. Trends Microbiol 2014, 22:345–353
8. Tornesello ML, Buonaguro L, Giorgi-Rossi P, Buonaguro FM: Viral and cellular biomarkers in the diagnosis of cervical intraepithelial neoplasia and cancer. Biomed Res Int 2013, 2013:519619
9. Ogilvie GS, van Niekerk D, Krajden M, Smith LW, Cook D, Gondara L, Ceballos K, Quinlan D, Lee M, Martin RE, Gentile L, Peacock S, Stuart GCE, Franco EL, Coldman AJ: Effect of screening with primary cervical HPV testing vs cytology testing on high-grade cervical intraepithelial neoplasia at 48 months: the HPV FOCAL randomized clinical trial. JAMA 2018, 320:43–52
10. Cuzick J, Cadman L, Mesher D, Austin J, Ashdown-Barr L, Ho L, Terry G, Liddle S, Wright C, Lyons D, Szarewski A: Comparing the performance of six human papillomavirus tests in a screening population. Br J Cancer 2013, 108:908–913
11. Virtanen E, Kalliala I, Dyba T, Nieminen P, Auvinen E: Performance of mRNA- and DNA-based high-risk human papillomavirus assays in detection of high-grade cervical lesions. Acta Obstet Gynecol Scand 2017, 96:61–68
12. Cook DA, Smith LW, Law J, Mei W, van Niekerk DJ, Ceballos K, Gondara L, Franco EL, Coldman AJ, Ogilvie GS, Jang D, Chernesky M, Krajden M: Aptima HPV Assay versus Hybrid Capture® 2 HPV test for primary cervical cancer screening in the HPV FOCAL trial. J Clin Virol 2017, 87:23–29
13. Ge Y, Christensen P, Luna E, Armylagos D, Xu J, Schwartz MR, Mody DR: Aptima Human Papillomavirus E6/E7 mRNA test results strongly associated with risk for high-grade cervical lesions in follow-up biopsies. J Low Genit Tract Dis 2018, 22:195–200
14. de Thurah L, Bonde J, Lam JUH, Rebolj M: Concordant testing results between various human papillomavirus assays in primary cervical cancer screening: systematic review. Clin Microbiol Infect 2018, 24: 29–36
15. Hawkes D, Brotherton JML, Saville M: Not all HPV nucleic acid tests are equal: only those calibrated to detect high grade lesions matter for cervical screening. Clin Microbiol Infect 2018, 24:436–437
16. de Thurah L, Bonde J, Lam JUH, Rebolj M: Not all HPV nucleic acid tests are equal: only those calibrated to detect high grade lesions matter for cervical screening: response to "Concordant testing results between various human papillomavirus assays in primary cervical cancer screening: systematic review" by de Thurah, Bonde, Uyen, Lam and Rebolj: published 27 May, 2017. Clin Microbiol Infect 2018, 24: 438–439
17. Heard I, Cuschieri K, Geraets DT, Quint W, Arbyn M: Clinical and analytical performance of the PapilloCheck HPV-Screening assay using the VALGENT framework. J Clin Virol 2016, 81:6–11
18. Chen Z, Schiffman M, Herrero R, DeSalle R, Anastos K, Segondy M, Sahasrabuddhe VV, Gravitt PE, Hsing AW, Burk RD: Evolution and taxonomic classification of alphapapillomavirus 7 complete genomes: HPV18, HPV39, HPV45, HPV59, HPV68 and HPV70. PLoS One 2013, 8:e72565
19. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A: Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 2012, 28:1647–1649
20. Zheng Z-M, Baker CC: Papillomavirus genome structure, expression, and post-transcriptional regulation. Front Biosci 2006, 11:2286–2302
21. Wang X, Meyers C, Wang H-K, Chow LT, Zheng Z-M: Construction of a full transcription map of human papillomavirus type 18 during productive viral infection. J Virol 2011, 85:8080–8092
22. Wentzensen N, Ridder R, Klaes R, Vinokurova S, Schaefer U, von Knebel Doeberitz M: Characterization of viral-cellular fusion transcripts in a large series of HPV16 and 18 positive anogenital lesions. Oncogene 2002, 21:419–426
23. Tang K-W, Alaei-Mahabadi B, Samuelsson T, Lindh M, Larsson E: The landscape of viral expression and host gene fusion and adaptation in human cancer. Nat Commun 2013, 4:2513
24. Peter M, Rosty C, Couturier J, Radvanyi F, Teshima H, Sastre-Garau X: MYC activation associated with the integration of HPV DNA at the MYC locus in genital tumors. Oncogene 2006, 25:5985–5993
25. Lu X, Lin Q, Lin M, Duan P, Ye L, Chen J, Chen X, Zhang L, Xue X: Multiple-integrations of HPV16 genome and altered transcription of viral oncogenes and cellular genes are associated with the development of cervical cancer. PLoS One 2014, 9:e97588
26. Kraus I, Driesch C, Vinokurova S, Hovig E, Schneider A, von Knebel Doeberitz M, Dürst M: The majority of viral-cellular fusion transcripts in cervical carcinomas cotranscribe cellular sequences of known or predicted genes. Cancer Res 2008, 68:2514–2522
27. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013, 29:15–21
28. Friedman J, Hastie T, Tibshirani R: Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010, 33:1–22
29. Dalstein V, Merlin S, Bali C, Saunier M, Dachez R, Ronsin C: Analytical evaluation of the PapilloCheck test, a new commercial DNA chip for detection and genotyping of human papillomavirus. J Virol Methods 2009, 156:77–83
30. Klug SJ, Molijn A, Schopp B, Holz B, Iftner A, Quint W, J F Snijders P, Petry K-U, Krüger Kjaer S, Munk C, Iftner T: Comparison of the performance of different HPV genotyping methods for detecting genital HPV types. J Med Virol 2008, 80:1264–1274
31. Cuzick J, Szarewski A, Cubie H, Hulman G, Kitchener H, Luesley D, McGoogan E, Menon U, Terry G, Edwards R, Brooks C, Desai M, Gie C, Ho L, Jacobs I, Pickles C, Sasieni P: Management of women who test positive for high-risk types of human papillomavirus: the HART study. Lancet 2003, 362:1871–1876
32. Schneider A, Hoyer H, Lotz B, Leistritza S, Kühne-Heid R, Nindl I, Müller B, Haerting J, Dürst M: Screening for high-grade cervical intra-epithelial neoplasia and cancer by testing for high-risk HPV, routine cytology or colposcopy. Int J Cancer 2000, 89:529–534