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 Hebert, Delphine Chrétien, Marie-Christine Demazoin, Michael Falguières, et al.

TITLE

Broad range papillomavirus transcriptome as a biomarker of papillomavirus-associated cervical high-grade cytology

AUTHORS

Philippe Pérot[1], Anne Biton[2], Jacques Marchetta[3], Anne-Gaelle Pourcelot[4], André Nazac[4], Henri Marret[5], Thomas Hébert[5], Delphine Chrétien[1], Marie-Christine Demazoin[6], Michaël Falguières[6], Laurence Arowas[6], Hélène Laude[6], Isabelle Heard[6] and Marc Eloit[1,7*]

[1] Pathogen Discovery Laboratory, Biology of Infection Unit, Institut Pasteur, Paris, France

[2] Centre de bioinformatique, biostatistique et biologie intégrative (C3BI), Institut Pasteur, Paris, France

[3] Centre Hospitalier Universitaire, Angers, France

[4] Hôpital Le Kremlin-Bicêtre, France

[5] Centre Olympe de Gouges, CHU Bretonneau, Tours, France

[6] Centre National de Reference des Papillomavirus, Institut Pasteur, Paris, France

[7] National Veterinary School of Alfort, Paris-Est University, Maisons-Alfort, 94704 Cedex, France

[*] Corresponding author: Marc Eloit marc.eloit@pasteur.fr +33(0)144389216

SHORT RUNNING HEAD: HPV RNA-Seq

FUNDING: This work was funded by Institut Pasteur (grants InnovDiag and Pasteur Innov), Paris, France.

COMPETING INTERESTS: PP, AB, IH and ME have submitted patent applications covering the findings of this work.

This manuscript contains 33 pages (MS Word), 3 figures, 7 tables, 3 supplemental figures and 4 supplemental tables.

ABSTRACT

Human Papillomaviruses (HPV) are responsible for over 99% of cervical cancers. Molecular diagnostic tests based on the detection of viral DNA or RNA have low Positive Predictive Values (PPV) 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 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 HPV in cervical lesions (HPV16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66, 68, 73, and 82). The purpose of this novel molecular assay 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 (i) HPV RNA-Seq can detect papillomaviruses with performances comparable to a widely used HPV reference molecular diagnostic kit, and (ii) a combination of the number of sequencing reads at specific early *vs* late HPV transcripts can be used as a marker of high-grade cytology, with encouraging diagnostic performances as a triage test.

INTRODUCTION

Human papillomaviruses (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]. HPV are the etiologic agents responsible for over 99% of all cervical cancers[3]. HPV are small, non-enveloped DNA viruses commonly transmitted through sexual contact, which infect basal cells and replicate in the nucleus of squamous epithelial cells. HPV include more than 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), towards non-productive 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 lesion, with highly productive infection and high rate of spontaneous regression. In contrast, high-risk persistent HPV infection is responsible for high-grade lesion, the true precancerous lesion.

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 (Pap) 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 since 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 to 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 can not only 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 messenger RNA as done with former techniques such as quantitative RT-PCR used in HPV RNA tests. To do so, we have developed a multiplexed amplification system targeting the virus splice junctions coupled with NGS analysis, tentatively named HPV RNA-Seq (based on the AmpliSeq technology), that allows to describe fine equilibrium among transcript species of 13 high-risk HPV (HPV16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66) plus 3 putative high-risk HPV (HPV68, 73, 82), in a single reaction. This molecular approach makes, in particular, possible to take a snapshot of the early *vs* 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 (Pap 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 HSIL, 28 LSIL), starting from cervical smears conserved at room temperature, we have shown that (i) HPV RNA-Seq can detect papillomaviruses with performances comparable to a HPV DNA-based reference diagnostic kit[17], and (ii) a combination of the number of sequencing reads at specific early *vs* 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 Comité de Protection des Personnes Ile de France 1 (CPP IDF1) and by the Agence Nationale de Sécurité des Médicaments et des Produits de Santé (ANSM). The data processing was authorized by the Commission Nationale de l'Informatique et des Libertés (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 $7x10^4$ cells/mL in four transport medium: PreservCyt Solution (Hologic, USA), NovaPrep HQ+ Solution (Novaprep, France), RNA Protect Cell Reagent (Qiagen, Germany) and NucliSens Lysis Buffer (BioMérieux, France). The mixtures were aliquoted in 1mL tubes and kept at room temperature for 2 hours (D0), 48 hours (D2), 168 hours (D7), 336 hours (D14) and 504 hours (D21). In parallel, $7x10^4$ cells pellets without transport medium were kept frozen -80°C for 2 hours, 48

hours, 168 hours, 336 hours and 504 hours as a control. At D0, D2, D7, D14 and D21, room temperature aliquots were centrifuged, the medium removed, and the pellets were frozen -80°C for a short time (<1h) before proceeding with RNA extraction. In the particular case of the NucliSens Lysis Buffer since the cells were lysed, the entire 1mL aliquot was frozen -80°C for a short time without prior centrifugation. For each sample, RNA was extracted using the PicoPure RNA Isolation kit (Thermo Fisher Scientific, USA), together with the corresponding (time match) frozen control, so that all samples have undergone one freezing cycle. RT-qPCR was performed to quantify the expression of the two human genes *G6PD* (forward primer: TGCAGATGCTGTGTCTGG; reverse primer: CGTACTGGCCCAGGACC) and *GAPDH* (forward primer: GAAGGTGAAGGTCGGAGTC; reverse primer: GAAGATGGTGATGGGATTTC) and the expression of the two viral genes HPV16 *E6* (forward primer: ATGCACCAAAAGAGAACTGC; reverse primer: TTACAGCTGGGTTTCTCTAC) and *E7* (forward primer: GTAACCTTTTGTTGCAAGTGTGACT; reverse primer: GATTATGGTTTCTGAGAACAGATGG) (Supplemental Figure 1). RNA integrity was assessed on a Bioanalyzer instrument (Agilent, USA) (Supplemental Figure 2).

HPV selection and splice sites 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: 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). Multiple alignment of HPV genomes was done with ClustalW v2.1 using Geneious[19] v10. Previously known splice donor (SD) and splice acceptor (SA) sites for HPV16[20] and HPV18[21] were reported on the alignment, and predictions of unknown SD and SA sites were done manually for the other genotypes by sequence analogy (Figure 1A and 1B).

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 follow:

HPV spliced junctions (**sp**): a set of target sequences, which are specific HPV splice events, involving a pair of splice donor (SD) and splice acceptor (SA) sites. The nomenclature includes a "sp" tag. For example, "31_sp_1296_3295_J43-46" stands for HPV31 (31), splice junction (sp), SD at position 1296 on HPV31 genome, SA at position 3295 on HPV31 genome, and junction (J) at position 43-46 on amplicon. The junction coordinates are given in a 4-bases interval, where the first 2 bases correspond to the donor part (or left part) and the last 2 bases to the acceptor part (or right part) of the sequence.

HPV unspliced junctions (**unsp**): a set of target sequences which are specific HPV genomic regions spanning either SD or SA sites, in the absence of any splice event. The nomenclature includes an "unsp" tag. For example, "31_unsp_1296_1297_J43-46" stands for HPV31 (31), unspliced (unsp), last base of the left part of the amplicon at position 1296 on HPV31 genome, first base of the right part of the amplicon at position 1297 on HPV31 genome, junction (J) at position 43-46 on 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 section "Sequencing data processing".

HPV genome away from spliced junctions (**gen**): a set of target sequences which are specific HPV genomic regions, away from any SD or SA sites. The nomenclature includes a "gen" tag. For example, "45_gen_1664_1794_NoJ" stands for HPV45 (45), HPV genomic region (gen), amplicon coordinates from position 1664 to position 1794 on HPV45 genome.

HPV-human fusion sequences (**fus**): 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 (put. bkpt, see **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 (18), candidate fusion transcript (fus),

break/fusion at position 929 on HPV18 genome, fused with MYC mRNA isoform 001 exon 3, junction (J) at position 37-40 on amplicon.

Human sequences (**hg**): a set of 30 human sequences used as internal controls retrieved from publically available AmpliSeq projects and representing housekeeping genes (*ACTB*, *B2M*, *GAPDH*, *GUSB*, *RPLP0*), epithelial markers (*KRT10*, *KRT14*, *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*, *WNT1*). The nomenclature for these sequences includes an "hg" tag. For example, "hg_TOP2A_E21E22" stands for human topoisomerase 2A mRNA exon 21-22.

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). The average amplicon size of the panel (primers included) is 141bp (range: 81-204bp). 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 old referred for colposcopy consultation in French hospitals. The patients were referred for colposcopy in the context of a LSIL or a 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 20mL of PreservCyt Solution (Hologic), and sent at room temperature to the HPV National Reference Center (CNR) at Institut Pasteur, Paris, France. From July 2014 to April 2015, 84 patients were enrolled in the study, coming from 3 different French centers:

CHU Angers (n=66); CHU Kremlin-Bicêtre (n=10); CHU Tours (n=6). Samples were removed from the study because of technical reasons (sample leakage, n=1), legal issues (n=7) or because they were used for initial technical tests (RNA conservation, RNA extraction and amplification, n=4). The remaining 72 samples (HSIL=37; LSIL=35) were processed (Supplemental Table S2).

Data collection

The following bio-clinical data were collected: date and results of the cytology test, age at the time of the cytology test, date and results of all available histological results posterior to colposcopy. As colposcopy was performed in the context of routine healthcare, biopsies were not performed in case of normal colposcopy.

HPV DNA detection using the PapilloCheck Test Kit (HPV DNA)

Upon reception at CNR, 16 mL of cytological sample were transferred into a 50 mL Falcon tube and centrifuged at 4,500 g for 10 minutes. The supernatant was removed and the pellet washed with 1 mL of PBS. Sample was then centrifuged again at 5,000 g for 10 minutes and the supernatant removed. The pellet was frozen at -80°C before DNA extraction. Following DNA extraction (Macherey Nagel, Germany), HPV detection was done using the PapilloCheck Test Kit (Greiner Bio-One GmbH, Germany) according to manufacturer instructions (Supplemental Table S2).

RNA extraction and characterization

In parallel to the HPV DNA procedure, 3 x 1 mL aliquots of cytological specimen were centrifuged at 14,000 rpm for 7 minutes, the supernatant was removed and the pellet was washed with 1 mL of PBS. Sample was then centrifuged again at 14,000 rpm for 7 minutes and the supernatant removed. The pellet was frozen at -80°C before RNA extraction. RNA extractions were done using the PicoPure RNA Isolation kit (Thermo Fisher Scientific), including on-column DNAse treatment, with a final elution volume of 30 µL. Total RNA was quantified on a Nanodrop (Life Technologies, USA) and RNA integrity was evaluated on a Bioanalyzer RNA 6000 pico chip (Agilent) using the RIN (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: CATCGAGCACGGCATCGTCA; reverse primer: TAGCACAGCCTGGATAGCAAC; amplicon size = 210bp), and *GAPDH* (forward primer: GAAGGTGAAGGTCGGAGTC; reverse primer: GAAGATGGTGATGGGATTTC; amplicon size = 226bp) were done in a SYBR Green format with 45 cycles of amplification. RT-negative (RT-) PCR were run to evaluate the presence of residual DNA after RNA extraction (Supplemental Table S2).

Amplification and sequencing

Starting from RNA, cDNA were 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 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 4 Ion Proton runs. Raw data (.fastq files) are available on the NCBI SRA database under BioProject accession number PRJNA525642.

Sequencing data processing

Reads were aligned to the reference sequences of the amplicons using STAR[27] v2.5.3a 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 the 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 to that of the average library size across samples) and the 215 amplicons capturing splice junctions (sp) of the 16 high-risk or putative high-risk HPV 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 (e.g. "SD1-SA1", see **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. 33 out of the 55 clinical samples have been selected as presenting enough coverage of these specific amplicons (20 mono-infected and 13 multi-infected samples). The remaining 22 samples of the dataset 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 (here the grade of the cytology) as a function of the predictor variables (here 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:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = b_0 + b^t x$$

where $p = P(y=1|x)$ 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: $p = \frac{1}{1+e^{-(b_0+bx)}}$

Implementation of the logistic regression model

To limit overfitting, we used L2-norm (ridge) regularization, which allows shrinking the magnitudes of the regression coefficients such that they will better fit future data. We estimated the logistic model using the R (http://www.r-project.org/ ; last accessed on January 29, 2019) package *glmnet*[28]. Leave-one-out (LOO) 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 in 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 upon the clinical outcome LSIL or HSIL obtained from the cytological analysis, and estimated on a training set consisting of 20 mono-infected samples (5 LSIL and 15 HSIL) in order to avoid a confusion bias. It is indeed anticipated that, in the case of multi-infected samples, several HPV 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 mono-infected 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 we could discriminate which HPV was responsible of 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 (Hologic), the most widely used solution for gynecological specimen collection; NovaPrep HQ+ Solution (Novaprep), a competitor product used for cells and DNA recovery but never evaluated for RNA conservation; RNA Protect Cell Reagent (Qiagen), a popular solution for RNA stability; and NucliSens Lysis Buffer (BioMérieux), a lysis buffer part of the NucliSens automated acid nucleic procedure which has been described as a 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 48h 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 Figure 1). 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 2). We therefore decided to use the PreservCyt solution to collect the gynecological specimen 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 HPV31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66, 68, 73, and 82 (Figures 1A and 1B). 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 HPV and named HPV RNA-Seq. Putative breakpoints 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 & sequencing

72 gynecological samples (HSIL=37; LSIL=35) coming from 3 different French centers (Angers, Kremlin-Bicêtre and Tours) and collected in PreservCyt solution were processed with RNA extraction using a method designed to recover total RNA from as little as a single cell (PicoPure RNA Isolation kit, Thermo Fisher Scientific) (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 = 210bp) and GAPDH mRNA (amplicon size = 226bp) indicated that RNA quality was compatible with amplification of 200-250bp 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. qPCR performed after omitting the reverse transcription step (RT-) 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). Note, the presence of residual cellular DNA or HPV DNA in RNA preparation is not a major concern since 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 (i.e. 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).

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 usable reads per sample was reached (min=0.02M; max=8.36M), among which an average of 2.1 million reads mapped to the human sequences (hg) used as internal controls (min=0.01M; max=8.06M) (Supplemental Table S3). The detection of highly expressed human sequences in all samples, even though inter-sample variations were observed, contributed to validate the sequencing procedure, which is important especially for the interpretation of HPV-negative samples. Rare non-zero values were also observed for some of the numerous HPV-human fusion sequences (fus) that were hypothesized (Supplemental Table S3) but

were all false positives, 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 HPV targeted by the panel. The number of reads mapping to HPV-specific amplicons (i.e. the sum of categories "sp", "unsp" and "gen") was used to detect the presence of a given HPV genotype. To help determining a threshold for detection, we took as a reference a HPV DNA test validated for clinical use (PapilloCheck, Greiner Bio-One GmbH). The best sensitivity and specificity values between the two tests were obtained for threshold of 100-200 reads (Figure 2). For example, a threshold value of 150 reads resulted in a Sensitivity ($Se_{(HPV\text{-}DNA)}$) of 97.3%, a Specificity ($Sp_{(HPV\text{-}DNA)}$) of 83.3%, leading to a Positive Predictive Value ($PPV_{(HPV\text{-}DNA)}$) 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 around 50% of HSIL and 50% of LSIL (Table 2, raw data in the Supplemental Table S2 and Supplemental Table S3). A more detailed view of the genotypes identified by both techniques is given in Figure 3. The number of mono-infected, multi-infected, or HPV-negative samples identified by the two tests is summarized in Table 3. Note that, because the HPV DNA test can detect the 16 high-risk or putative high-risk HPV captured by HPV RNA-Seq (HPV16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66, 68, 73, and 82) plus 8 additional low-risk HPV (HPV6, 11, 40, 42, 43, 44/45, 53 and 70), the comparison was based only on the 16 HPV 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 *vs* 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 *vs* n=10 multi-infected samples, Table 3). Globally, HPV16 was found at a slightly weaker occurrence by HPV RNA-Seq (n=18 *vs* n=19) in favor of other genotypes such as HPV31, 33, 45, 52, 56, 58 or 66 which were less commonly found by the HPV DNA test (HPV31 n=5 *vs* n=4; HPV33 n=3 *vs* n=1; HPV45 n=3 *vs* n=2; HPV52 n=5 *vs* n=3; HPV56 n=4 *vs* n=2;

HPV58 n=5 *vs* n=4; HPV66 n=2 *vs* n=1, Figure 3). Apart from HPV16, only HPV51 was less frequently found by HPV RNA-Seq than by HPV DNA (n=2 *vs* 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

We conducted an exploratory analysis on 20 of the mono-infected samples in which we showed that HPV RNA spliced junctions could be used to predict high-grade cytology. We focused the analysis on amplicons capturing splice junctions (category "sp") to be sure to detect HPV transcripts. However, the number of mono-infected samples (n=20) used as training set was small, in particular the number of samples of LSIL (n=5). In this configuration, we were not able to perform a fully accurate variable selection, i.e. to select the strict minimum number of amplicons that were necessary for HSIL versus LSIL prediction and set the others to zero coefficient. In addition, we were not able to avoid over-fitting, as using only 5 LSIL and 15 HSIL samples did not allow capturing the diversity of the whole population. LOO cross-validation was used to pick the lambda giving the minimum cross-validated error using ridge regularization. Lambda = 0.08 gave a mean cross-validated error of 15%. We also computed a 20% prediction error using nested cross-validation. This error rate can be seen as an indicator of how the model could fit future datasets. We used the corresponding parameter 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 mono-infected samples was classified correctly, except for one observation (Table 5). It is interesting to note that this unique misclassified sample (IonXpress_019_2613), which was classified LSIL by the cytological analysis, was further found as containing a mixture of LSIL and HSIL lesions after histological examination performed more than one year after the sampling done for HPV RNA-Seq/cytology.

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 a HSIL prediction, the sample was considered to be HSIL. We calculated

performances for HSIL prediction for all samples, considering as not being of high-grade both the six 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 (mono-infected, multi-infected and HPV-negative) were $Se_{(cyto)}$=66.7%, $Sp_{(cyto)}$=85.7%, $PPV_{(cyto)}$=81.8% and $NPV_{(cyto)}$=72.7% (**Table 6A**). The performances were also calculated for the subset of 39 samples having at least one 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 6B**). Note that the ratio HSIL to LSIL remained similar between these two populations (around 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 *vs* HPV DNA) and high-grade cytology prediction (HPV RNA-Seq *vs* cytology), including posterior histological 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 histological examination were, however, not available for all patients. The time interval separating HPV RNA-Seq/cytology tests from histological analysis, varying between 0 and 780 days, was another limitation in this study. To try to overcome these drawbacks, we compared the performances of HPV RNA-Seq *vs* histology to the performances of cytology *vs* histology, considering either all available samples (**Supplemental Table S4 A**), or only samples for which histology was done less than 3 months after HPV RNA-Seq/cytology (**Supplemental Table S4 B**), or only samples for which histology was done less than 6 months after HPV RNA-Seq/cytology (**Supplemental Table S4 C**). In addition and for each category, we made the distinction between the performances obtained when HPV RNA-Seq HPV-positive and HPV-negative patients were grouped together (**Supplemental Table S4 1&1'**), or when only HPV-positive patients were considered (**Supplemental Table S4 2&2'**). Calculation of the PPV as a function of HSIL prevalence in the population was also done (**Supplemental Figure 3** and **Supplemental Table S4**). All these results are given as Supplemental Data.

DISCUSSION

We have developed a highly-multiplexed RT-PCR assay coupled with Next-Generation Sequencing (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 a HPV detection and genotyping assay were evaluated in comparison to the HPV DNA PapilloCheck kit (HPV DNA), 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 non-homogeneous distribution of infected cells. It is also important to note that 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 to HPV species in these three potential false HPV-positive patients was close to the limit of detection for some of them (<= 400 reads) but not for all (e.g. 39527 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 to older references.

Effective cervical cancer screening requires high Se and NPV for high-risk HPV detection, as women with a negative HPV test are usually tested again only after several years. The positive threshold for

458 HPV genotyping was set at 150 reads in this study because it optimized both Se and Sp values, but
459 lowering this threshold in order to maximize the sensitivity remains possible. Such adjustments will be
460 possible following the study of larger cohorts.
461
462 As a second application of HPV RNA-Seq, as a triage test, a logistic regression model for the prediction
463 of high-grade cytology was built based on a combination of the number of reads captured at specific
464 HPV RNA spliced junctions, using the grade found by cytology as a reference. This evaluation was
465 conducted in a population of women with LSIL or HSIL cytology results. Where at least one HPV was
466 given a high-grade signature, the patient's prediction was set as "HSIL". Conversely, "Not HSIL" was
467 used when either no HPV was detected in the sample (threshold of 150 reads) or when none of the
468 genotypes detected by HPV RNA-Seq were given high-grade prediction (absence of detectable
469 transcripts). We preferred to use the terminology "Not HSIL" rather than "LSIL" because the protocol
470 did not allow the comparison of the HPV-DNA positive samples evaluated as LSIL with the ones
471 evaluated as normal in cytology. Also, because there is a possibility that cervical lesions could in some
472 rare cases originate from causes other than HPV infections, the use of "Not HSIL" instead of "LSIL" in
473 the case of HPV-negative samples seemed more appropriate.
474
475 As far as the comparison with cytology could be used as a benchmark, when the 55 patients were
476 considered (including mono-infected, multi-infected and HPV-negative), the number of HSIL predicted
477 by HPV RNA-Seq (n=22) was lower than the number of HSIL identified by cytology (n=27), resulting in
478 $Se_{(cyto)}$=66.7%, $Sp_{(cyto)}$=85.7%, $PPV_{(cyto)}$=81.8% and $NPV_{(cyto)}$=72.7% (Table 6A). Interestingly, when only
479 HPV RNA-Seq HPV-positive samples were considered, the **PPV**$_{(cyto/HR+)}$ for the detection of high grade
480 lesions remained unchanged but the **Se**$_{(cyto/HR+)}$ and the **NPV**$_{(cyto/HR+)}$ increased to 94.7% (+28.0) and
481 94.1% (+21.4) respectively, with the number of HSIL predicted by HPV RNA-Seq (n=22) becoming
482 superior to the number of HSIL identified by cytology (n=19) (Table 6B). In this case the only one
483 patient identified HSIL by cytology but not predicted HSIL by HPV RNA-Seq (sample
484 IonXpress_020_3137) was found HSIL by the histological examination done 350 days later, which

485 opens the possibility that this sample might be positive if the patient would be tested again by HPV
486 RNA-Seq at a date closer to the histological examination.
487
488 In clinical use, after primary screening for high-risk HPV in the general population, a triage test with
489 high Sp and PPV is needed for the triage of women at risk of transforming infection before colposcopy.
490 In countries that have adopted HPV DNA as a screening test, cytological analysis can be used for the
491 triage of women at risk because cytology has better Sp and PPV than HPV DNA tests[31,32]. In line with
492 that, the $PPV_{(cyto)}$=81.8% of HPV RNA-Seq outperformed HPV DNA and other RNA tests, whose PPV as
493 triage assays never exceed 50% in a population of women referred for colposcopy composed with a
494 similar 1:1 ratio of HSIL to LSIL[10].
495
496 We thus sought to evaluate the added value of HPV RNA-Seq over cytology for the triage of women at
497 risk of developing cervical cancer. To do so, histology was used as the gold standard for the diagnosis
498 of cervical lesions. However, an inherent limitation of this work was that histology was not
499 concomitant with the sampling carried out for HPV RNA-Seq/cytology tests, which means that by the
500 time histology was performed (between 0 and 780 days after initial sampling), the lesion could have
501 evolved spontaneously in one direction (LSIL to HSIL) or another (HSIL to LSIL). To help clarify this point
502 we compared side to side the performances of HPV RNA-Seq *vs* histology to the performances of
503 cytology *vs* histology, considering different categories of samples (**Supplemental Table S4**).
504 Remarkably, whatever the category considered, the Sp of HPV RNA-Seq *vs* histology was always
505 greater than or equal to the Sp of cytology *vs* histology (+0.0 to +11.1), and the resulting PPV of HPV
506 RNA-Seq *vs* histology was always greater than the PPV of cytology *vs* histology (+2.4 to +7.4 in this
507 population reflecting others studies[9]) (**Supplemental Table S4**). Calculation of PPV as a function of HSIL
508 prevalence allowed anticipating a delta PPV between HPV RNA-Seq *vs* histology and cytology *vs*
509 histology. For example, in the case where the ratio of HSIL to LSIL would tend to 1:2 as seen
510 elsewhere[10], the delta PPV could be up to +10.4 in favor of HPV RNA-Seq *vs* histology (range: +4.4 to
511 +10.4, (**Supplemental Figure 3** 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 *vs* histology was always higher on the subset of HPV-positive patients (+12.0 to +33.3, Supplemental Table S4), similarly to the evaluation done with cytology taken as reference (Table 6). Lastly, the Sp of HPV RNA-Seq *vs* histology increased on the subset of patients for whom histology was performed less than 3 months after sampling (+5.0 to +8.3, Supplemental Table S4) but decreased on the subset of patients for whom histology was performed less than 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 or less per sample is enough for performing HPV genotyping, but more would be needed for HPV transcripts detection. We do realize for example that 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, former product by Thermo Fisher Scientific developed for Ion Proton and PGM instruments, has been transferred in 2018 to Illumina and is now fully compatible with Illumina sequencers. In a decentralized laboratory model, it becomes possible that 4-6 samples could run on a benchtop iSeq100 sequencer for a cost per sample around 200 USD, with RNA extraction, quality controls and data analysis included (salaries and equipment excluded). In a more centralized view where 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 around 10-20 USD 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 (hg) 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, in order 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 had actually 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 (ASCUS) 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 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 in order 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: The authors want to thank the Clinical Core of the Center of Translational Science (CRT-CC) of Institut Pasteur for the management of all legal and ethical aspects of the study, PathoQuest (Paris, France) and the Center for Translational Science (CRT) / Cytometry and Biomarkers Unit of Technology and Service (CB UTechS) at Institut Pasteur for providing access to the sequencers.

DECLARATIONS

Availability of data and material: The datasets 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 NCBI SRA database under BioProject accession number PRJNA525642.

Authors' contribution: ME and PP designed the study. PP designed the AmpliSeq custom panel. JM, AGP, AN, HM and TH collected the gynecological samples. MF, MCD, LA, HL and IH managed the samples at the HPV National Reference Center. IH contributed to the clinical protocol. MCD and MF provided SiHa cells. PP and DC performed the sequencing experiments and analyzed the data. AB did the biostatistical analysis. PP, ME and AB wrote the manuscript. ME supervised the study. All authors read and approved the final manuscript.

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 Off J Am Soc Clin Oncol, 2011, 29:4294–301
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, 189:12–9

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, Bravo IG, Stoler M, Broker TR, Stanley MA. The biology and life-cycle of human papillomaviruses. Vaccine, 2012, 30 Suppl 5:F55-70
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, 22:345–53
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, 108:908–13
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, 96:61–8
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 Off Publ Pan Am Soc Clin Virol, 2017, 87:23–9
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 Off Publ Eur Soc Clin Microbiol Infect Dis, 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 Off Publ Eur Soc Clin Microbiol Infect Dis, 2018, 24:436–7
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 Off Publ Eur Soc Clin Microbiol Infect Dis, 2018, 24:438–9
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 Off Publ Pan Am Soc 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. Bioinforma Oxf Engl, 2012, 28:1647–9
20. Zheng Z-M, Baker CC. Papillomavirus genome structure, expression, and post-transcriptional regulation. Front Biosci J Virtual Libr, 2006, 11:2286–302
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, 85:8080–92

641 22. Wentzensen N, Ridder R, Klaes R, Vinokurova S, Schaefer U, Doeberitz M von K. Characterization
642 of viral-cellular fusion transcripts in a large series of HPV16 and 18 positive anogenital lesions.
643 Oncogene, 2002, 21:419–26
644 23. Tang K-W, Alaei-Mahabadi B, Samuelsson T, Lindh M, Larsson E. The landscape of viral
645 expression and host gene fusion and adaptation in human cancer. Nat Commun, 2013, 4:2513
646 24. Peter M, Rosty C, Couturier J, Radvanyi F, Teshima H, Sastre-Garau X. MYC activation associated
647 with the integration of HPV DNA at the MYC locus in genital tumors. Oncogene, 2006, 25:5985–
648 93
649 25. Lu X, Lin Q, Lin M, Duan P, Ye L, Chen J, Chen X, Zhang L, Xue X. Multiple-integrations of HPV16
650 genome and altered transcription of viral oncogenes and cellular genes are associated with the
651 development of cervical cancer. PloS One, 2014, 9:e97588
652 26. Kraus I, Driesch C, Vinokurova S, Hovig E, Schneider A, von Knebel Doeberitz M, Dürst M. The
653 majority of viral-cellular fusion transcripts in cervical carcinomas cotranscribe cellular sequences
654 of known or predicted genes. Cancer Res, 2008, 68:2514–22
655 27. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR.
656 STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 2013, 29:15–21
657 28. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via
658 Coordinate Descent. J Stat Softw, 2010, 33:1–22
659 29. Dalstein V, Merlin S, Bali C, Saunier M, Dachez R, Ronsin C. Analytical evaluation of the
660 PapilloCheck test, a new commercial DNA chip for detection and genotyping of human
661 papillomavirus. J Virol Methods, 2009, 156:77–83
662 30. Klug SJ, Molijn A, Schopp B, Holz B, Iftner A, Quint W, J F Snijders P, Petry K-U, Krüger Kjaer S,
663 Munk C, Iftner T. Comparison of the performance of different HPV genotyping methods for
664 detecting genital HPV types. J Med Virol, 2008, 80:1264–74
665 31. Cuzick J, Szarewski A, Cubie H, Hulman G, Kitchener H, Luesley D, McGoogan E, Menon U, Terry
666 G, Edwards R, Brooks C, Desai M, Gie C, Ho L, Jacobs I, Pickles C, Sasieni P. Management of

667 women who test positive for high-risk types of human papillomavirus: the HART study. Lancet
668 Lond Engl, 2003, 362:1871–6
669 32. Schneider A, Hoyer H, Lotz B, Leistritza S, Kühne-Heid R, Nindl I, Müller B, Haerting J, Dürst M.
670 Screening for high-grade cervical intra-epithelial neoplasia and cancer by testing for high-risk
671 HPV, routine cytology or colposcopy. Int J Cancer, 2000, 89:529–34
672

673

674 **FIGURES AND TABLES**

675

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

684 **Figure 2**: Receiver Operating Characteristic (ROC) curve. HPV DNA (PapilloCheck) was used as a
685 reference to evaluate the performances of HPV RNA-Seq for the detection of at least one HPV
686 genotype in a sample. Data: 55 patients. AUC: Area Under Curve.

687 **Figure 3**: Comparison of the number of HPV genotypes identified by HPV RNA-Seq and HPV DNA.
688 Vertical bars represent the number of HPV genotypes identified by HPV RNA-Seq at threshold value
689 of 150 reads (red) *vs* HPV DNA (PapilloCheck, black). Putative high-risk HPV are indicated by a star (*).
690 Data: 55 patients.

691

692 **Table 1**: HPV RNA-Seq AmpliSeq custom panel contents.

	sp	unsp	gen	fus	hg	
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

693 The number of target amplicons is indicated for each category (sp, unsp, gen, fus, hg) and for each

694 viral and cellular origin. Putative high-risk HPV are indicated by a star (*).

695

696 **Table 2**: Performances of HPV RNA-Seq for HPV detection.

		HPV DNA			
		HPV+	HPV-	**Se**(HPV-DNA)	97.3%
HPV RNA-Seq	HPV+	36	3	**Sp**(HPV-DNA)	83.3%
	HPV -	1	15	**PPV**(HPV-DNA)	92.3%
				NPV(HPV-DNA)	93.8%

Performances of HPV RNA-Seq at threshold value of 150 reads *vs* HPV DNA (PapilloCheck) for HPV detection. Sensitivity (Se), Specificity (Sp), Positive Predictive Value (PPV) and Negative Predictive Value (NPV) are given. HPV+ means that at least one HPV genotype is identified in a patient. Data: 55 patients.

Table 3: Comparison of HPV RNA-Seq and HPV DNA for the classification of samples.

	HPV RNA-Seq	HPV DNA
Mono-infected samples	26	27
Multi-infected samples	13	10
HPV-negative samples	16	18

Number of mono-infected, multi-infected and HPV-negative samples by HPV RNA-Seq at threshold value of 150 reads, *vs* HPV DNA (PapilloCheck). Data: 55 patients.

Table 4: Coefficients of the (ridge) logistic regression.

junction	coefficient	name_transcript_category	name_transcript_contents
(Intercept)	0.468298365		
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 id 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 “early” transcript.

Table 5: Classification results of the (ridge) logistic regression.

sample	prediction_score	prediction_class	prediction_class	prediction_accuracy
IonXpress_039_115	0.115	-1	LSIL	TRUE
IonXpress_033_730	0.204	-1	LSIL	TRUE
IonXpress_038_114	0.259	-1	LSIL	TRUE
1492	0.425	-1	LSIL	TRUE
IonXpress_019_2613	0.562	1	LSIL	FALSE
IonXpress_027_598	0.653	1	HSIL	TRUE
729	0.716	1	HSIL	TRUE
567	0.718	1	HSIL	TRUE
IonXpress_018_2439	0.902	1	HSIL	TRUE

610	0.904	1	HSIL	TRUE
1066	0.911	1	HSIL	TRUE
IonXpress_034_758	0.919	1	HSIL	TRUE
1122	0.934	1	HSIL	TRUE
25	0.944	1	HSIL	TRUE
IonXpress_037_1267	0.947	1	HSIL	TRUE
IonXpress_024_26	0.965	1	HSIL	TRUE
IonXpress_025_538	0.97	1	HSIL	TRUE
752	0.976	1	HSIL	TRUE
IonXpress_021_443	0.984	1	HSIL	TRUE
2612	0.993	1	HSIL	TRUE

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

715

716 **Table 6:** Performances of HPV RNA-Seq for the prediction of high-grade cytology.

A		Cytology			
		HSIL	LSIL	**Se**(cyto)	66.7%
HPV RNA-Seq	HSIL	18	4	**Sp**(cyto)	85.7%
	Not HSIL	9	24	**PPV**(cyto)	81.8%
				NPV(cyto)	72.7%

B		Cytology			
		HSIL	LSIL	**Se**(cyto/HR+)	94.7%
HPV RNA-Seq	HSIL	18	4	**Sp**(cyto/HR+)	80.0%
HR+	Not HSIL	1	16	**PPV**(cyto/HR+)	81.8%
				NPV(cyto/HR+)	94.1%

Performances of HPV RNA-Seq *vs* cytology for HSIL detection (**A**) for the 55 patients and (**B**) for the subset of 39 patients having at least one HPV identified by HPV RNA-Seq. Sensitivity (Se), Specificity (Sp), Positive Predictive Value (PPV) and Negative Predictive Value (NPV) are given. "Not HSIL" means that either no HPV was detected in the sample by HPV RNA-Seq or that none of the HPV genotypes detected were given HSIL prediction.

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 (days) cyto-histo
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	n.d.
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	n.d.	n.d.
D-15-0041_567_BC9	16	16			16	HSIL	HSIL	HSIL	n.d.
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	neg	neg				Not HSIL	HSIL	LSIL	85
IonXpress_018_2439	58	58			58	HSIL	HSIL	LSIL	164
IonXpress_018_440	neg	neg				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	n.d.	130
IonXpress_021_443	58	33,58	33		58	HSIL	HSIL	LSIL	99
IonXpress_022_23	neg	neg				Not HSIL	HSIL	HSIL	n.d.
IonXpress_022_444	16,33	16,33		33	16	HSIL	HSIL	HSIL	69
IonXpress_023_24	(6),(11),(53)	neg				Not HSIL	HSIL	HSIL	[0-13]

IonXpress_023_536	neg	neg				Not HSIL	LSIL	LSIL	101
IonXpress_024_26	45	45			45	HSIL	HSIL	HSIL	106
IonXpress_024_537	neg	neg				Not HSIL	LSIL	LSIL	71
IonXpress_025_457	neg	neg				Not HSIL	LSIL	LSIL	278
IonXpress_025_538	35	31,35	31		35	HSIL	HSIL	HSIL	191
IonXpress_026_539	neg	neg				Not HSIL	LSIL	n.d.	n.d.
IonXpress_026_565	16	neg				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	neg	neg				Not HSIL	HSIL	n.d.	n.d.
IonXpress_030_612	neg	neg				Not HSIL	LSIL	LSIL	113
IonXpress_031_613	35,39,(44/55)	35,39		35,39		Not HSIL	LSIL	LSIL	83
IonXpress_032_728	neg	neg				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)	neg				Not HSIL	LSIL	n.d.	20
IonXpress_037_100	neg	neg				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	neg	neg				Not HSIL	HSIL	HSIL	85
IonXpress_039_115	56	56		56		Not HSIL	LSIL	LSIL	34
IonXpress_039_1598	neg	neg				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	neg	52,58	52,58			Not HSIL	LSIL	LSIL	160
IonXpress_046_2066	(6)	66	66			Not HSIL	LSIL	HSIL	99

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

730

731

732 **Supplemental Figure 1:** Evaluation of transport medium for RNA conservation: detection of cellular and
733 viral transcripts by RT-qPCR at D0, D2, D7, D14 and D21. (**A**) Cellular transcripts GAPDH and G6PD

734 (average values), (**B**) Viral transcripts E6 & E7 (average values). The relative abundance (y-axis, log
735 scale) is calculated in comparison to the higher expression value of the dataset, set to 1. Time scale (x-
736 axis, hours): 2 hours (D0), 48 hours (D2), 168 hours (D7), 336 hours (D14), 504 hours (D21). (JPEG
737 2233 Ko)

738 **Supplemental Figure 2:** Evaluation of transport medium for RNA conservation: RNA integrity at D0, D2,
739 D7, D14 and D21. Bioanalyzer gel-like visualization of RNA profiles, showing 18S and 28S rRNA
740 populations (two distinct bands) as a marker of RNA integrity. (JPEG 788 Ko)

741 **Supplemental Figure 3**: Positive Predictive Value (PPV) of HPV RNA-Seq vs histology (in blue) and of
742 cytology vs histology (in red) function of the prevalence of HSIL in the population (P HSIL). Calculations
743 are provided in the Supplemental Table S4. (JPEG 2780 Ko)