


RESEARCH ARTICLE

Open Access



# Targeted RNA next generation sequencing analysis of cervical smears can predict the presence of hrHPV-induced cervical lesions

Karolina M. Andralojc<sup>1,2†</sup>, Duaa Elmelik<sup>1†</sup>, Menno Rasing<sup>3</sup>, Bernard Pater<sup>3</sup>, Albert G. Siebers<sup>4,5</sup>, Ruud Bekkers<sup>6,7</sup>, Martijn A. Huynen<sup>8</sup>, Johan Bulten<sup>4</sup>, Diede Loopik<sup>9</sup>, Willem J. G. Melchers<sup>2†</sup> and William P. J. Leenders<sup>1,3\*†</sup> 

## Abstract

**Background:** Because most cervical cancers are caused by high-risk human papillomaviruses (hrHPVs), cervical cancer prevention programs increasingly employ hrHPV testing as a primary test. The high sensitivity of HPV tests is accompanied by low specificity, resulting in high rates of overdiagnosis and overtreatment. Targeted circular probe-based RNA next generation sequencing (ciRNAseq) allows for the quantitative detection of RNAs of interest with high sequencing depth. Here, we examined the potential of ciRNAseq-testing on cervical scrapes to identify hrHPV-positive women at risk of having or developing high-grade cervical intraepithelial neoplasia (CIN).

**Methods:** We performed ciRNAseq on 610 cervical scrapes from the Dutch cervical cancer screening program to detect gene expression from 15 hrHPV genotypes and from 429 human genes. Differentially expressed hrHPV- and host genes in scrapes from women with outcome “no CIN” or “CIN2+” were identified and a model was built to distinguish these groups.

**Results:** Apart from increasing percentages of hrHPV oncogene expression from “no CIN” to high-grade cytology/histology, we identified genes involved in cell cycle regulation, tyrosine kinase signaling pathways, immune suppression, and DNA repair being expressed at significantly higher levels in scrapes with high-grade cytology and histology. Machine learning using random forest on all the expression data resulted in a model that detected ‘no CIN’ versus CIN2+ in an independent data set with sensitivity and specificity of respectively  $85 \pm 8\%$  and  $72 \pm 13\%$ .

**Conclusions:** CiRNAseq on exfoliated cells in cervical scrapes measures hrHPV-(onco)gene expression and host gene expression in one single assay and in the process identifies HPV genotype. By combining these data and applying machine learning protocols, the risk of CIN can be calculated. Because ciRNAseq can be performed in high-throughput, making it cost-effective, it can be a promising screening technology to stratify women at risk of CIN2+. Further increasing specificity by model improvement in larger cohorts is warranted.

**Keywords:** Cervical intraepithelial neoplasia, High risk human papilloma virus, Machine learning, Screening, Targeted RNA sequencing

<sup>†</sup>Karolina M. Andralojc, Duaa Elmelik, Willem J. G. Melchers and William P. J. Leenders contributed equally to this work.

\*Correspondence: William.leenders@radboudumc.nl; William.leenders@predicadx.com

<sup>1</sup> Department of Biochemistry, Radboudumc, Radboud Institute of Molecular Life Sciences, Geert Grooteplein 26, Nijmegen 6525 GA, The Netherlands  
Full list of author information is available at the end of the article

## Background

Annually, 570,000 new cases of cervical cancer (CC) are diagnosed worldwide, with 310,000 attributable deaths [1–3]. Over 99% of CCs are associated with sexually transmitted and highly infectious high risk papilloma viruses (hrHPV) [4]. Because hrHPVs are causally



involved in cervical dysplasia and cancer, CC-screening programs are increasingly based on molecular screening of cervical smears for the presence of hrHPV, followed by triage with cytology [5, 6].

In 2019, 9.8% of women tested positive for hrHPV in the Dutch CC screening program [7]. Of these hrHPV-positive women, 69% had normal cytology and were invited for a follow-up smear and cytology 6 months later. The remaining 31% hrHPV-positive women with cytological outcome ASC-US (atypical squamous cells of undetermined significance) or low- or high-grade squamous intraepithelial lesion (LSIL or HSIL) were referred for colposcopy, often accompanied by a biopsy. Of this group, only 37% was diagnosed with moderate or severe cervical intraepithelial neoplasia (CIN2 or CIN3) [7]. Whereas CIN3 lesions can be removed by a loop electrosurgical excision procedure (LEEP), the decision to treat CIN2 depends on factors like age and child wish, as LEEP can have serious side effects during pregnancy [8, 9]. Furthermore, in the majority of cases of low and medium grade HPV-induced CIN lesions (CIN1 and CIN2), these are spontaneously cleared within a year, allowing watchful waiting. In summary, the high sensitivity but low specificity of hrHPV testing for detecting CIN2+ results in high rates of overdiagnosis and overtreatment with an associated risk on adverse events. There is an unmet need for better triage tests to reduce the number of unnecessary referrals.

Productive hrHPV infections require the maintenance of HPV genomic DNA in an episomal state and repression of the immune response, conditions that are supported by the expression of, among others, the early HPV-E2 gene [10, 11]. This state is mostly associated with low-grade CIN (CIN1). Such infections are often transient and clear spontaneously [12]. Persistent infection may result in the integration of the viral genome in host DNA, which is frequently observed in CC [13] and high-grade CIN lesions [14]. Such integration is often associated with loss of expression of functional E2 and constitutive expression of the hrHPV-E6/7 gene [15–17]. E2/E6 RNA ratios are therefore lower in cancer than in CIN lesions [18]. Transcription from the E6/7 gene produces a bicistronic messenger RNA, encoding the E6 and E7 oncoproteins that are responsible for degradation of cell cycle regulator proteins P53 and pRB, and for altering transcription in infected cells [11, 19, 20]. As a result of functional loss of P53 and RB, uncontrolled proliferation accompanied by lack of functional DNA repair occurs, two principal requirements to start the oncogenic process. HPV E6/7 RNA assays have therefore higher specificity to detect CIN2+ (median 46%) than HPV DNA assays (38%) [21, 22]. This can be biologically explained

because HPV DNA assays cannot distinguish between dormant, productive and oncogenic HPV infections.

For at least some hrHPV genotypes, the E6/7 gene contains a pseudo-intron that can be spliced out, resulting in the E6\*I splice product. The E7 open reading frame is more efficiently translated from E6\*I than from E6/7 mRNA, and it has been suggested that E6\*I expression is associated with progression to higher grade CIN [23, 24]. There is debate in the literature if measuring hrHPV E6\*I mRNA improves specificity to detect CIN2+ [23, 24]. Currently, there are no commercial tests available that can measure hrHPVE6\*I RNA.

We previously reported on the potential of targeted RNA next generation sequencing (ciRNAseq) in several cancer types [25–27]. We also analyzed benign and malignant gynecological tissues with the technique [28] and showed its potential to concomitantly determine hrHPV oncogene activity and host gene activity. We showed that low ratios of hrHPVE2:E6/7 expression may indicate integration of the viral genome in the host genome [9].

Here, we used CC cell lines and cervical scrapes to test if simultaneous profiling of HPV oncogenes and of host genes that are implicated in hrHPV-oncogenesis can identify hrHPV-positive women who are at risk of having or developing CIN2+.

## Methods

### Cell lines and clinical material

HeLa cells and CaSki cells (a gift from Dr. A. Kaufmann, Charite University hospital, Germany) were cultured under standard conditions, trypsinized and fixated in PreservCyt solution (LBC, ThinPrep, Hologic Corp, Marlborough, MA, USA). Women participating in the Dutch CC screening program were informed that their residual cervical smear material in PreservCyt could be used for anonymized research and had the opportunity to opt out. Only left-over material from women who did not opt out was selected and analyzed after pseudonymization. Cytological classification of hrHPV-positive smears was performed according to the Bethesda system [29]. Cytological results and histological outcomes during follow-up were obtained from the nationwide network and registry of histo- and cytopathology in the Netherlands (PALGA; Houten, the Netherlands).

One cohort of hrHPV-DNA positive cervical smears (cohort A,  $n=356$ ) was randomly collected from the Dutch CC screening program. Another independent cohort (cohort B,  $n=204$ ) consisted of hrHPV-DNA-positive smears, selected for enrichment of specific cytological abnormalities. Furthermore, a cohort of 50 hrHPV-DNA negative scrapes (Cohort C) was analyzed. Cytology outcomes of cohorts A and B are summarized

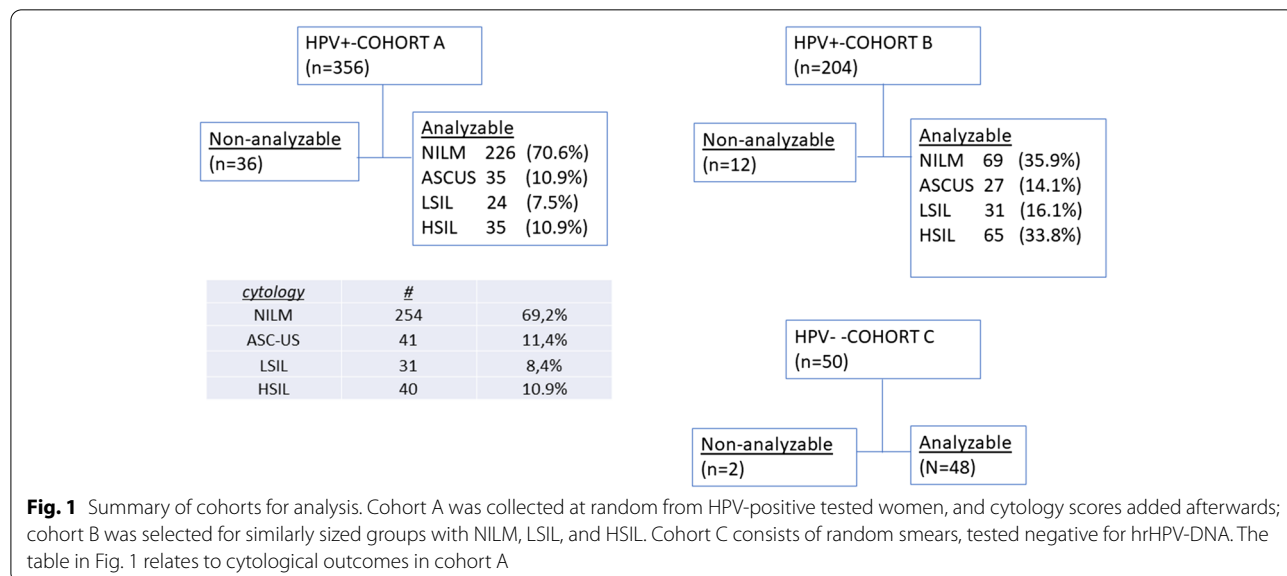
in Fig. 1. Clinical outcomes during follow-up of the combined cohorts A and B in relation to initial cytology scores are summarized in Table 1. For this study, women with cytology NILM (Negative for Intraepithelial Lesion and Malignancy) in both primary scrape and in return scrape during follow-up at 6 months were considered as “no CIN.” Median follow-up in this study was 7 months (range 6–709 days).

**CiRNAseq**

Five ml samples of residual scrape material in PreservCyt solution were centrifuged for 5 min at 2500×g. The pellet was lysed in 1 mL of Trizol reagent (Thermo Scientific), and RNA was isolated through standard procedures and dissolved in 20µl nuclease-free water. Sixteen microliters with a maximum of 2µg total RNA was treated with DNase, followed by cDNA generation using Superscript-II (ThermoScientific) [28, 30]. The same protocol was applied to CaSki cells and HeLa cells after fixation in PreservCyt. To validate the efficacy of hrHPV specific probes, hrHPV cDNA amplicons were generated by RT-PCR using RNA from scrapes that were previously

diagnosed as positive for RNA of different HPV genotypes. Amplicons were purified from agarose gel and equimolarly pooled. This pool was used as positive control in the assay.

The protocol for ciRNAseq was described before [28, 31]. In short, ~50 ng of cDNA or positive control was hybridized overnight with a set of 2394 single molecule molecular inversion probes (smMIPs), designed with MipGen software [32] to identify and quantitatively measure expression levels of a total of ~513 gene transcripts, including E2, E6/7, and E6\* from hrHPV genotypes HPV 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66, 68, and 73 and human transcripts encoding housekeeping proteins and enzymes involved in tyrosine kinase signaling, metabolism, DNA repair, oncogenes, tumor suppressor genes, and genes involved in immunity. Sequences were retrieved from hg38 ([https://www.ncbi.nlm.nih.gov/assembly/GCF\\_000001405.26/](https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/)) and from the PAVE database (<https://pave.niaid.nih.gov/>). The smMIP pool contained for each hrHPV genotype >5 smMIPs that target non-overlapping, independent ROIs. During the capture reaction, extension arms in smMIP-cDNA hybrids were



**Table 1** Follow-up of all HPV-positive women (cohorts A and B)

| 1 <sup>st</sup> CYTO | NO FOLLOW UP | No CIN/ASCUS/LSIL/CIN1 | CIN2+      |
|----------------------|--------------|------------------------|------------|
| NILM (N=321)         | 37 (11.5%)   | 264 (82.2%)            | 20 (6.2%)  |
| ASCUS (N=65)         | 7 (10.8%)    | 47 (72.3%)             | 11 (16.9%) |
| LSIL (N=61)          | 6 (9.8%)     | 39 (63.9%)             | 16 (26.2%) |
| HSIL (N=109)         | 5 (4.6%)     | 22 (20.2%)             | 82 (75.2%) |

extended with KlenTaq polymerase, and smMIPs were circularized by Ampligase (both from Epicentre, Madison, WI). After enzymatic removal of non-reacted, linear smMIPs and cDNAs by exonuclease treatment, purified circular smMIPs were PCR-amplified with barcoded Illumina primer sets. PCR products of the expected length of 266 bp were purified with Ampurebeads (Beckmann Coulter Genomics, High Wycombe, UK) measured on TapeStation and subjected to Illumina Novaseq sequencing on an SP flow cell (2 × 150 bp reads).

### Data processing

Illumina output was barcode-decomplexed to produce forward and reverse FASTQ files for each sample. FASTQ files were processed by Seqnext software (JSI systems, Ettingen, Germany) to count the total number of reads, generated by each smMIP. To eliminate PCR amplification bias, all smMIPs were designed to contain an 8 N unique molecule identifier (UMI). Reads that contain the same UMI and have identical sequences were collapsed to unique counts, reflecting the number of individual smMIPs that were circularized in the assay. Seqnext settings allowed 2% mismatches, to prevent missing counts from intratypic HPV variants. To prevent false identification of low-risk (lr)HPVs with high sequence homology, ROI sequences with higher sequence homology to lrHPV were discarded by filtering. A sample was annotated as hrHPV-RNA positive if more than one hrHPV-specific read for E2, E6/7, and/or E6\* was detected [28].

The total number of unique counts for each sample was used as a quality control for the efficiency of the capture reaction. Samples with an (arbitrary) total unique read count below 1000 did not match our quality requirements and were excluded from further analysis. Data were normalized by calculating

$$\frac{\text{unique\#reads for each smMIP}}{\text{total\#unique reads for all smMIPs}} * 10^6$$

and expressed as FPM (fragment per million). The mean FPM value was calculated from all smMIPs reactive against that transcript. Mean FPM values were considered as gene expression value.

### Computational biology

To investigate if there is value in the data, we selected all HPV-positive tested women with a repeated cytology diagnosis NILM (considered as “safe”) and with an outcome CIN2+. Matrices of mean FPM levels of these samples were subjected to unsupervised agglomerative clustering using Manhattan distance and Ward.D2 method [25, 28]. Clusters were visualized using Clust-Vis [33]. Fisher’s exact test was performed to calculate

significance of the asymmetric distribution of NILM and CIN2+ over the clusters. In a supervised analysis, differentially expressed genes between groups “safe” and “CIN2+” were identified by Mann Whitney U tests. Benjamini-Hochberg was used to calculate adjusted *P*-values, corrected for false discoveries. In parallel, mean FPM levels were log-transformed to achieve a normal distribution (after adding 0.1 to prevent log<sup>0</sup>), and differentially expressed genes were identified by a two-sided *T*-test. Genes that came out as differentially expressed by both tests with *P* < 0.05 and consistently over cohorts A and B were identified.

In the next step, decision tree models were built using the R package randomForest, version 4.6-14 [34] using ciRNAseq profiles of hrHPV-positive scrapes from women with outcome CIN2+ and women classified as safe (NILM in two consecutive scrapes). Data were randomly sampled into 5 training and validation sets (70/30) without allowing duplicates. For each pair, sets of 50 models were built. The models with the minimum number of false negatives were selected to form an aggregate set of five models. These models were then applied to ciRNAseq profiles from an independent external validation set of scrapes with known clinical outcome to calculate sensitivity and specificity.

## Results

### Gene expression profiling of PreServCyt fixed cell lines

We first investigated specificity and technical performance of ciRNAseq on HPV-positive cell lines CaSki and HeLa, and on a positive control sample, consisting of a mix of HPV cDNA amplicons. To mimic clinical scrapes as closely as possible, we fixated cultured cells in PreservCyt and stored aliquots of 10,000 cells for 1, 7, and 28 days as room temperature before proceeding to RNA isolation. Good quality data (as defined by total numbers of unique read counts) were obtained for samples even after 7 days of storage, whereas after 28 days, data quality was significantly less, though still interpretable (25-fold less unique read counts as compared to day 1, not shown). Table 2 shows a representative example of triplicate analysis of CaSki cells and HeLa cells, showing that in CaSki exclusively HPV16E2, E6\* and E6/7 are detected, whereas in HeLa, HPV18E6/7 and E6\* are detected, with relatively few reads from only a single HPV18E2-detecting smMIP. The lack of HPV18E2 reads in HELA cells was not a result of low-performance of the smMIPs because these performed well in the positive control (Tale IIB). Both cell lines were completely negative for all other hrHPV genotypes that are measured in this assay (data not shown).

**Table 2** (A) Fragment of raw output with unique read counts of HPV16 and HPV18 E2, E67, and E6\* with expression levels in CaSKI and HELA cell lines. (B) Output of the same probes on a positive control sample containing hrHPV amplicons

| A               |       |      |      | B     |       |       |                   |
|-----------------|-------|------|------|-------|-------|-------|-------------------|
| CELL LINE       | CaSKI |      |      | HELA  |       |       | amplicon controls |
| HPV16E2_smMIP1  | 237   | 197  | 286  | 0     | 0     | 0     | 868               |
| HPV16E2_smMIP2  | 584   | 443  | 705  | 0     | 0     | 0     | 1533              |
| HPV16E2_smMIP3  | 655   | 738  | 1056 | 0     | 0     | 0     | 1552              |
| HPV16E2_smMIP4  | 37    | 23   | 45   | 0     | 0     | 0     | 81                |
| HPV16E6*_smMIP5 | 871   | 842  | 1291 | 0     | 0     | 0     | 200               |
| HPV16E6_smMIP6  | 636   | 576  | 905  | 0     | 0     | 0     | 1886              |
| HPV16E6_smMIP7  | 25    | 19   | 41   | 0     | 0     | 0     | 926               |
| HPV16E7_smMIP8  | 1654  | 1661 | 2369 | 0     | 0     | 0     | 1446              |
| HPV18E2_smMIP1  | 0     | 0    | 0    | 57    | 44    | 45    | 426               |
| HPV18E2_smMIP2  | 0     | 0    | 0    | 0     | 0     | 0     | 197               |
| HPV18E2_smMIP3  | 0     | 0    | 0    | 0     | 0     | 0     | 2506              |
| HPV18E6*_smMIP4 | 0     | 0    | 0    | 11560 | 9834  | 9875  | 869               |
| HPV18E6_smMIP5  | 0     | 0    | 0    | 989   | 882   | 815   | 543               |
| HPV18E6_smMIP6  | 0     | 0    | 0    | 308   | 270   | 240   | 107               |
| HPV18E7_smMIP7  | 0     | 0    | 0    | 13202 | 11229 | 11451 | 7548              |

### Profiling of cervical scrapes

Having established that PreservCyt fixation of CaSKI and HeLa cells and storage at room temperature is compatible with *ciRNAseq* analysis, we proceeded with analysis of cervical scrapes that are routinely collected and stored in PreservCyt at room temperature. Storage time of samples was variable. We performed *ciRNAseq* analysis on 50 hrHPV-DNA negative scrapes and two independent cohorts of 356 and 204 hrHPV-DNA-positive scrapes from women, participating in the Dutch population-based screening program. Cytology characteristics of the randomly collected cohort A (Fig. 1) were in accordance with previously published national data [6], confirming that this cohort was representative for the hrHPV-positive Dutch population.

Quality of *ciRNAseq* data, expressed as total unique reads in a sample, varied from 0 to 1.12 million (mean 180,000, median 114,000). This high variability can be explained by differences in cellularity and RNA yield between samples. Dropout percentage (samples with less than 1000 total unique read counts) was ~8%, leaving 320, 192, and 48 analyzable samples in cohorts A, B, and C, respectively. HrHPV RNA was undetectable in all hrHPV-DNA negative scrapes that passed our quality control standards (Fig. 1 and data not shown).

### Associations of hrHPVE6/7 and HPV E6\* gene expression with cytology

We first investigated the randomly collected cohort A as a representation of hrHPV-positive women in the

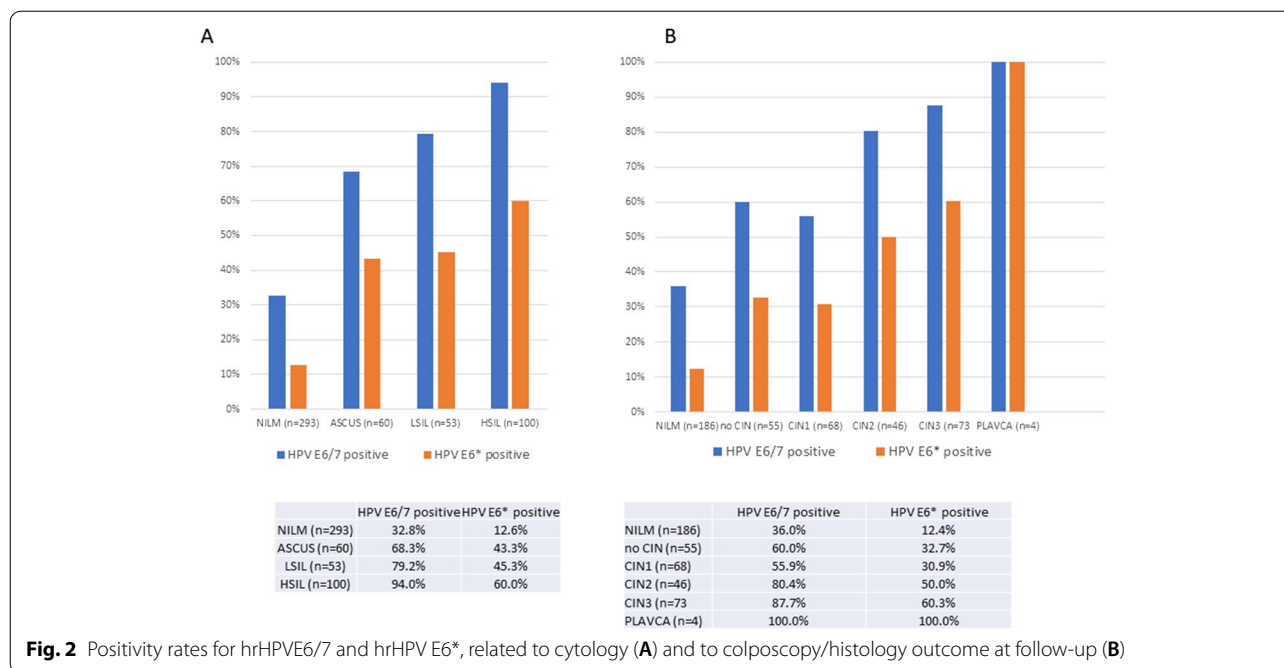
Dutch population. Results of hrHPV E6/7 RNA and E6\* RNA expression in scrapes with cytology NILM, ASC-US, LSIL, and HSIL are summarized in Fig. 2A. In 33% of hrHPV-positive scrapes with the cytologic outcome NILM, hrHPV E6/7 RNA was detected. The percentage of hrHPV-E6/7 RNA positivity was 68%, 79%, and 94% in groups with cytology score ASC-US, LSIL, and HSIL, respectively. In 28/226 scrapes with no detectable hrHPVE6/7, we detected hrHPV-E2 mRNA without expression of E6/7 (not shown), suggestive of a productive state of the virus [16]. Of these women, 25 had outcome NILM, no CIN, or CIN1 (89%), 1 had an outcome CIN2 (4%), and 2 had outcome CIN3 during follow-up (8%).

HrHPV-E6\* was expressed in 13% of all NILM-scrapes, 43% in ASC-US, 45% in LSIL, and 60% of HSIL. Similar patterns of hrHPV E6/7 and E6\* expression were observed when hrHPV expression profiles in the first scrape were correlated to histology outcome with median follow-up of 7 months (see Fig. 2B).

### Associations of HrHPVE6/7 and HPV E6\* gene expression with clinical outcome

In cohort B, similar frequencies of hrHPV RNA positivity were found as in cohort A (not shown). Therefore, cohorts A and B were combined in our further analyses. We first investigated to which extent the various hrHPV transcripts could predict the cytology diagnosis. Results are presented in Table 3 and show that hrHPV





**Table 3** A) distribution of HPV-RNA positivity over groups with cytology scores NILM and ASCUS+ (reason for referral to a gynecologist). B) distribution of HPV-RNA positivity in scrapes from women with an outcome <CIN2+ and >CIN2+ (median follow up 7 months)

| <b>A</b>      |      |        |           | <b>B</b>      |       |       |           |
|---------------|------|--------|-----------|---------------|-------|-------|-----------|
|               | NILM | ASCUS+ |           |               | <CIN2 | CIN2+ |           |
| hrHPVE6/7 neg | 197  | 36     | sens: 83% | hrHPVE6/7 neg | 217   | 37    | sens: 81% |
| hrHPVE6/7 pos | 96   | 177    | spec: 65% | hrHPVE6/7 pos | 165   | 155   | spec: 48% |
|               |      |        | NPV=85%   |               |       |       | NPV=85%   |
|               |      |        | PPV=66%   |               |       |       | PPV=48%   |
|               |      |        | sens:52%  |               |       |       | sens:56%  |
| hrHPVE6* neg  | 256  | 103    | spec: 75% | hrHPVE6* neg  | 318   | 85    | spec: 59% |
| hrHPVE6* pos  | 37   | 110    | NPV:71%   | hrHPVE6* pos  | 75    | 107   | NPV:79%   |
|               |      |        | PPV: 75%  |               |       |       | PPV: 59%  |

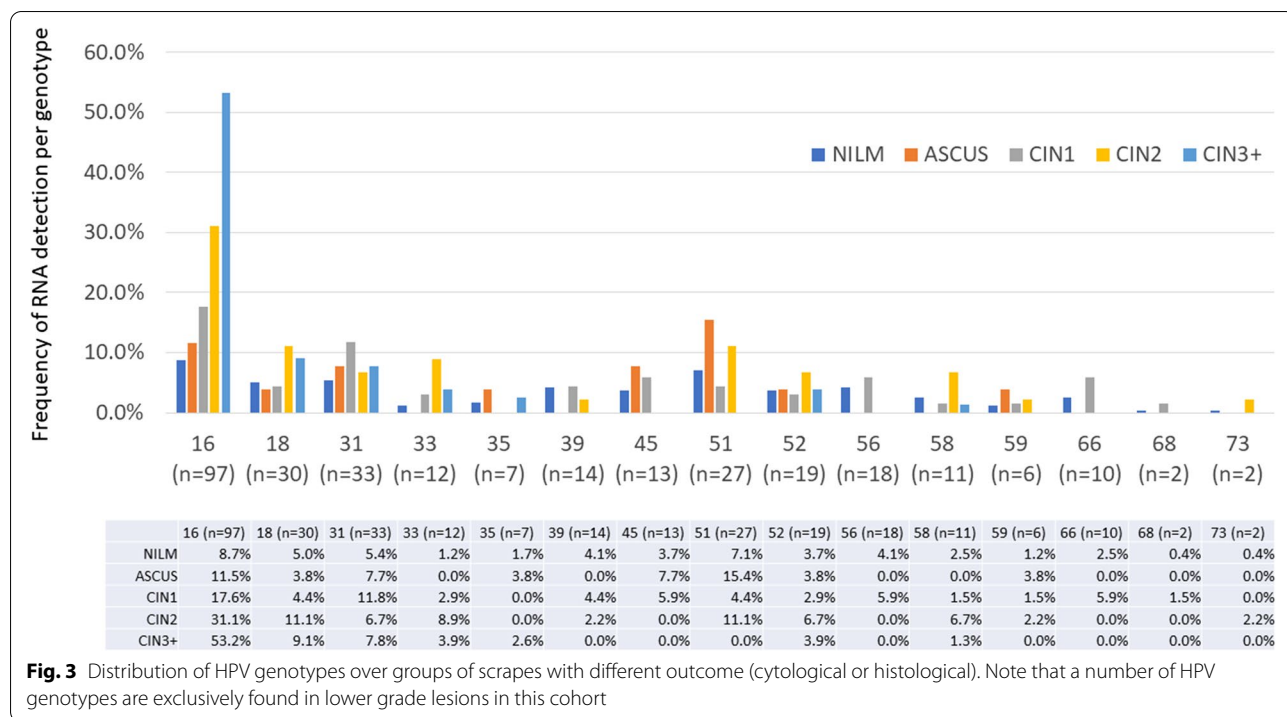
E6/7 RNA detection has the highest negative predictive value (85%), while hrHPVE6\* RNA detection has the highest positive predictive value of detecting ASCUS or higher (75%, see Table 2 (B)). Also, for detecting outcome <CIN2, HPV E6/7 RNA had the better negative predictive value (85%) while hrHPVE6\* RNA had the higher positive predictive value for detecting CIN2+ (59%, Table 3 (C)).

**Associations of hrHPV genotypes with clinical outcome**

Sequence information from ciRNAseq can be directly used for hrHPV genotyping. Overall distribution of hrHPV genotypes in the two combined cohorts is

presented in Fig. 3. As expected, HPV16 constitutes the majority of infections and was associated most with high-grade CIN. In this cohort, hrHPV genotypes 45/56/66/68 were underrepresented in CIN2+ relative to groups with no CIN/CIN1. In 72/298 hrHPV RNA positives (24%), multiple hrHPV genotypes were detected (not shown).

These data confirm that hrHPV-RNA detection has higher negative and positive predictive value than HPV-DNA testing but also show that hrHPV-RNA testing is not sufficient to stratify women who are at risk of CIN [35]. Therefore, additional biomarkers are needed.



**Additional value of host gene profiling**

In the process of hrHPV-induced oncogenesis, the transcriptional activity of the infected host cell changes [11, 19, 20]. To investigate if and how host gene expression levels correlate with histological outcome, we performed unsupervised hierarchal clustering of host gene expression data on all samples from women from with a cytology score NILM in first and second scrape, or with histology outcome no CIN, from here on referred to as “safe” (n=195) and women with an outcome CIN2+ (n=105) as described before [28]. We omitted scrapes from women with no follow-up and from women with an initial cytology ASC-US or higher, who had no CIN during follow-up, to prevent contamination of the group with underdiagnosed cases. Results in Fig. 4A show that the analysis yielded two main clusters. Distribution of CIN2+ and “safe” over the clusters was asymmetric with high significance (Fishers’ exact test, P < 0.0001). Even with this unsupervised clustering method, negative and positive predictive values of 76% and 56% were obtained for predicting CIN2+ from host gene expression data. Thus, these data show that host gene expression levels in scrapes can discriminate women who are safe from women with CIN2+.

We proceeded to identify the most prominent differentially expressed genes between groups “safe” and CIN2+ (Mann-Whitney U test, FDR < 0.00002, > 2-fold change in gene expression). This resulted in a set of 117 genes, a selection of which is shown in Fig. 4B.

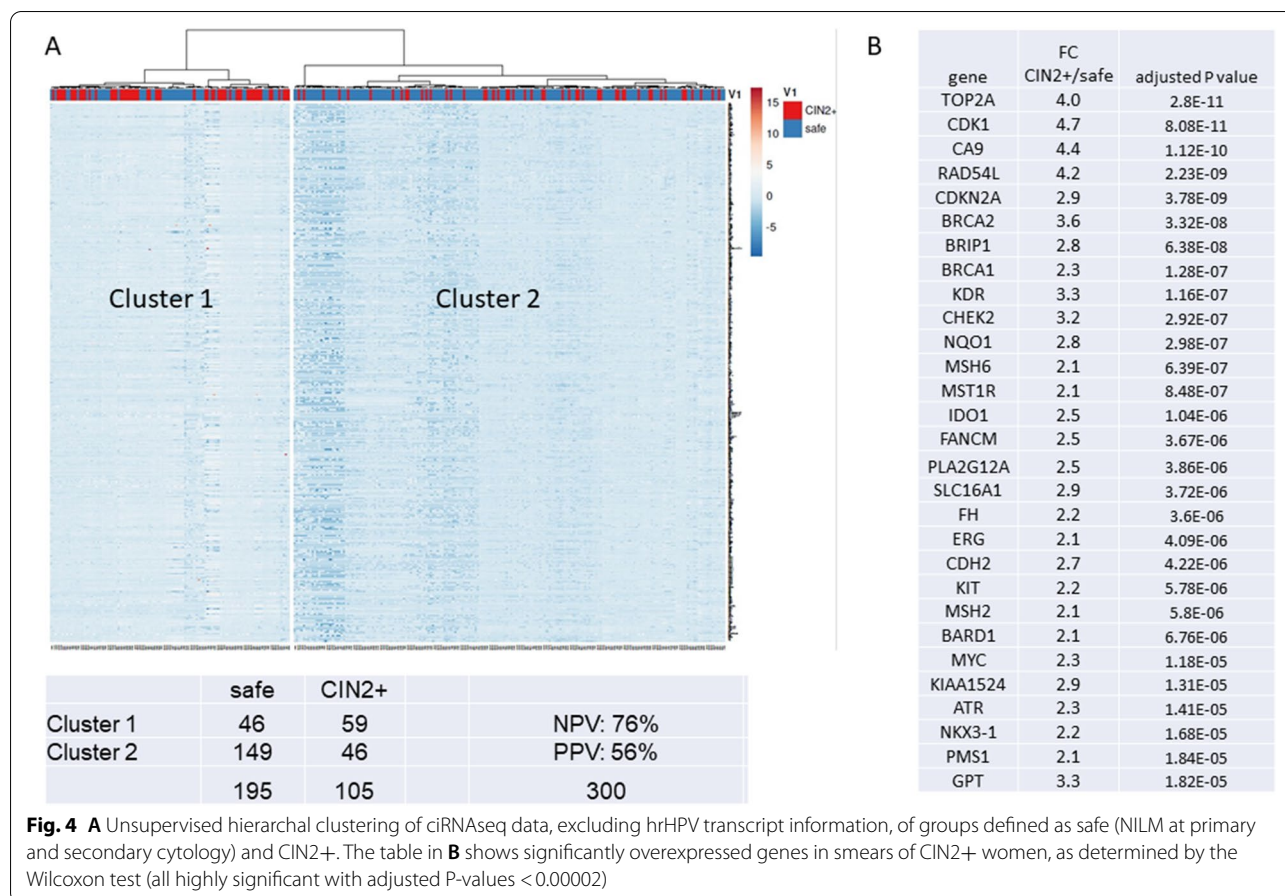
Relatively highly expressed genes in scrapes from women with outcome CIN2+ include genes involved in cell cycle regulation (e.g., CDK1, CDKN2A), DNA synthesis and repair (ATR, BRCA1, BRCA2, BRIP1, FANCM, MSH2, MSH6, RAD54L, TOP2A), kinase signaling (KIT, NTRK1, PTPRZ), metabolism (CA9, GRIK5, NQO1, SLC16A1), transcription factors (MYC), and immunity (IDO1).

To explore the potential of predicting CIN2+ based on the CiRNAseq data, we applied machine learning-based algorithms. Figure 5 shows one of the 5 independent cohorts that were analyzed by our random-forest based algorithm and shows that with a risk score cutoff of 0.7 (established during the building of the model) CIN2+ were identified by the model with a sensitivity of 85 ± 1% and a specificity of 72 ± 13%.

**Discussion**

The introduction of primary hrHPV screening in prevention programs has led to increased numbers of referrals for colposcopy and biopsy of which more than 70% are in retrospect unnecessary. To reduce these numbers, there is an unmet need for reliable risk assessment tests [19, 36] that we addressed here using gene expression values measured with CiRNAseq.

CiRNAseq is a high-throughput technology of multiplexed targeted RNA sequencing that quantitatively measures RNAs of interest. Because smMIP probes can be selected to have exon-exon boundaries in their

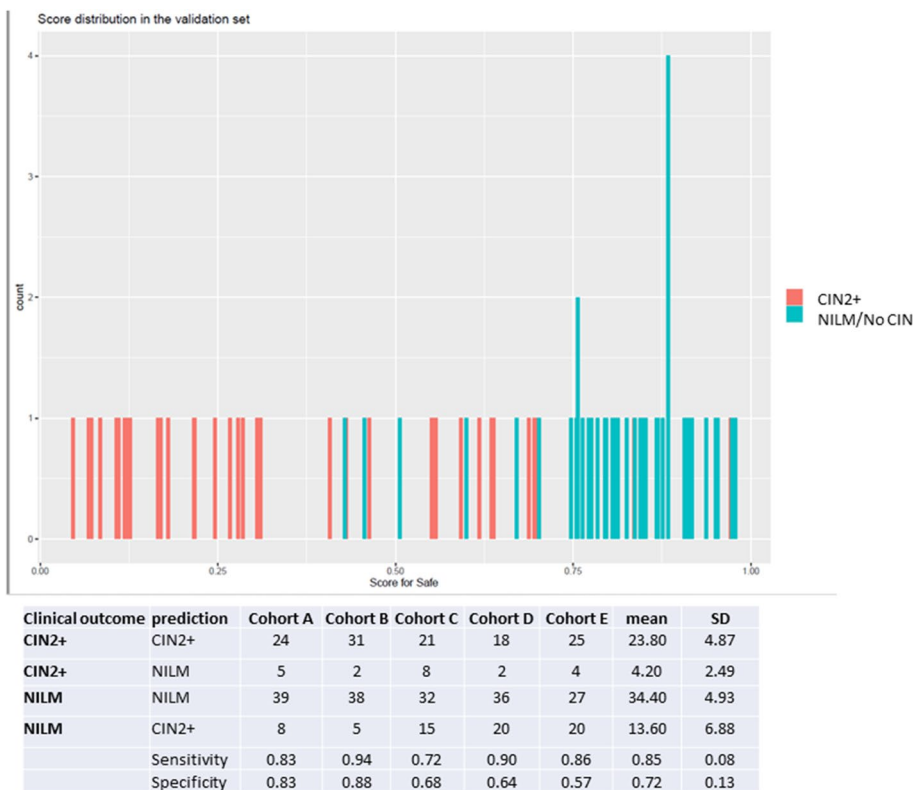


region of interest, the technique can be used for quantitative detection of splice variants of interest, such as E6\*. In previous work, we demonstrated that in a subgroup of cervical cancers, HPV E6/7 RNA was found without expression of the E2 early gene [28]. To investigate this further and to confirm that PreservCyt fixation does not decrease RNA quality, we profiled PreservCyt-fixed cervical cancer cell lines CaSki and HeLa. For HeLa cells, only one of three E2-selective smMIPs recognizing the 5'-region of the E2 transcript was reactive against HeLa RNA. Since all E2-selective smMIPs were effective in control samples (Table 2 (B)), this confirms that the E2 gene is disrupted in this cell line [37], consistent with integration of the HPV gene in the HeLa genome with a break point in E2 [38]. Whereas E6/7:E2 ratios are elevated in cancers compared to CIN [18], it remains to be investigated if these ratios are reflected in cervical scrapes from women with high-grade CIN and cancer. This requires thorough investigation because expression of the E2 open reading frame on the transcript level is not necessarily reflected in the protein level, such as observed in CASKI cells [39, 40].

Our analysis of 560 cervical scrapes shows that ciRNAseq analysis of hrHPV-positive cervical scrapes can identify women who are at risk of high-grade CIN, by combining data on hrHPV gene- and host gene expression data. In the process, the technique also identifies HPV genotype. Analysis of the data confirms that HPV types 35/39/56/59 and 66 are mostly associated with low grade lesions [41]. Additionally, we found that in this cohort, HPV45, HPV58, and HPV68 were less frequently detected in HSIL than in lower grade cytology. This finding needs confirmation in larger independent cohorts.

In most bicistronic hrHPV E6/7 RNAs, the start codon for the E7 open reading frame (ORF) is close to the stop codon of the E6 ORF, leading to inefficient translation initiation of the E7 ORF. Removal of the E6 intron leads to a shortened E6 product, placing the E7 start codon in a context allowing efficient translation initiation. It has been suggested that expression of E6\* is associated with higher grade CIN [23, 24], but we could not confirm this in this study. Additional studies in larger cohorts are required to investigate if E6\* of specific genotypes have added value in risk predictions.





**Fig. 5** Outcome of the application of a random forest model, generated with *ciRNAseq* data from 360 smears, on an independent dataset of 63 smears. With a preset cutoff score of 0.8, all samples regarded safe (NILM at first scrape and repeat scrape) were correctly identified

In this study, we found that ~70% of women with hrHPV-DNA positive, but hrHPV-RNA negative scrapes were diagnosed as NILM, suggesting that these cases concern latent infections, contributing to a higher negative-predictive value of *ciRNAseq*. On the other hand, for 7.5% of scrapes in which we could not detect hrHPV-E6/7 RNA in the first scrape, women were diagnosed as CIN2+ in follow-up. Without exception, these cases were diagnosed at least 7 months after the first HPV-DNA positive scrape. Whether these cases concern newly acquired infections, or activation of latent infections diagnosed in the first scrape, is not known and requires that diagnosis of CIN lesions is accompanied by an additional scrape analysis with *ciRNAseq*. In this context, it is important to note that, whereas cervical cancers are considered to be caused by hrHPV, in a subgroup of HPV-DNA-positive cervical cancers, no HPV transcripts could be detected [42]. Also, the cancer genome atlas (TCGA) includes several HPV-negative cervix carcinomas [15]. Studies have shown that gene expression profiles of these cancers are distinct from those of HPV-positive cervical cancers, providing evidence that HPV-negative cervical cancers comprise a separate entity [43]. These cancers will be missed in screening programs that use HPV

screening as a primary test. Additional clinical studies are required to test whether RNA profiling of cervical scrapes can identify HPV-negative gynecological cancers and which host cell RNA biomarkers should be measured for this purpose [35].

We argued that early-stage transcriptional alterations in clinically significant hrHPV infections would be readily detectable with *ciRNAseq*, giving additional predictive and prognostic value to HPV gene expression profiles [19, 36]. This hypothesis was confirmed in unsupervised cluster analysis of *ciRNAseq* data and machine learning which separated “safe” women from women with CIN2+ during follow-up.

We identified a set of human genes that are significantly higher expressed in scrapes from women with follow-up diagnosis CIN2+ compared to women with normal histology. Our results seem in part to be related to hrHPV biology: expression of hrHPV-E6/7 oncogenes results in degradation of cell cycle gatekeepers TP53 and RB, leading to accumulation of DNA damage in cells with active DNA replication [44]. This may explain the upregulation of DNA damage sensor and repair proteins in scrapes with high-grade cytology. A well-known consequence of hrHPV E6/7 expression is upregulation of

the cell cycle inhibitor gene *CDKN2A*, the gene encoding the p16<sup>INK4a</sup> protein. Under physiological conditions, *CDKN2A* expression is mutually exclusive with expression of cell proliferation markers. However, transforming hrHPV infection results in a lack of RB, evoking expression of the *CDKN2A* product P16<sup>INK4A</sup>. The lack of RB also leads to continuous cell cycling. The co-expression of *CDKN2A* with proliferation markers such as CDK1 and proliferating cell nuclear antigen (PCNA) in CIN2+ [45–51] is a unique feature of malignant HPV-biology and was recapitulated in our *ciRNAseq* data. Interestingly, we also identified elevated expression levels of actionable genes. One example is *IDO1*, a protein involved in immune suppression and a possible target for therapy [52].

To investigate the value of *ciRNAseq* as a triage test on hrHPV-positive tested scrapes, we built prediction models using the random forest method on *ciRNAseq* profiles from scrapes with outcome “safe” or CIN2+ and tested the models on an independent cohort of scrapes with extreme outcomes (safe and CIN2+). The sensitivity and specificity of the models to predict CIN2+ in this group was respectively  $85 \pm 8\%$  and  $72 \pm 13\%$ . To improve specificity, large prospective clinical studies with sufficiently sized groups per hrHPV genotype and per cytology and long clinical follow-up are needed. Because of the small group sizes of low-prevalent hrHPVs, for this study, we grouped all hrHPVs. Our study shows that certain hrHPV genotypes are restricted to low-grade dysplasia only. If this can be confirmed in large studies, this biological knowledge can be implemented in the models. The same is true for detection of the HPV-E6\* splice variant of different HPV genotypes. Other options to improve specificity of detecting CIN2+ could be the additional profiling of genes involved in immunity and inflammation, simply by predicting immune-mediated HPV clearance. If specificity can be raised, the technique could replace cytology as a triage test and has the potential to reduce overtreatment of healthy women that now receive a false-positive cytology diagnosis. The application of *ciRNAseq* can be seen in several scenarios, with advantages and disadvantages. It could be used as a primary screening test, which, according to our data, would immediately lead to a 70% reduction of false positive results (DNA-positive but RNA negative). Because in this case information on latent HPV-infection would be missed, effective screening would probably require retesting after 2 years. In a more realistic scenario, the test can be performed on HPV-DNA positive scrapes, substituting PAP tests, or can be performed as an addition to the PAP test. More research is required to determine what the most efficient scenario is with respect to cost-benefit.

## Conclusions

We here show the potential of *ciRNAseq* on cervical scrapes to detect expression of HPV oncogene RNAs from high-risk HPV genotypes, concomitant with genotyping and detection of expression levels from human host genes that are associated with CIN2+. Apart from hrHPV oncogenes, we identify a set of genes that are upregulated in scrapes from women with high-grade lesions. The combined hrHPV gene expression- and host gene expression data can be used to build decision-tree based models for more specific classification of women who need treatment because of increased risk of CIN.

## Abbreviations

CC: Cervical cancer; CIN: Cervical intraepithelial neoplasia; FPM: Fragment per million; hrHPV: High risk human papilloma virus; NILM: Negative for Intraepithelial Lesion and Malignancy; smMIP: Single molecule molecular inversion probe.

## Acknowledgements

We thank Dr. Hans van Leeuwen and Dr. Wynand Alkema (TenWise BV) for help with biostatistics. Dr. A. Kaufmann is acknowledged for kindly providing us with CaSki cells.

## Authors' contributions

KMA, DE, and MR performed the experiments; BP, DE, and MAH contributed to the statistical analyses; RB and DL contributed to the collection of clinical materials; AG contributed to the retrieving of clinical metadata; JB performed the cytology; KMA, WJGM, and WPJL contributed to the conceptual design. All authors read and approved the final manuscript.

## Funding

This work was financially supported by a grant from NWO (Take-Off I feasibility study), Radboudumc (Innovation grant), Ruby&Rose foundation, and RedMed-Tech Ventures.

## Availability of data and materials

Data will be made available on request.

## Declarations

### Ethics approval and consent to participate

The regional institutional review board and the National Institute for Public Health and Environment granted approval before start of the study (No. 2014-1295).

### Consent for publication

All authors have given their consent for publication.

### Competing interests

William Leenders is shareholder and part-time employee of the Radboudumc spin-off company Predica Diagnostics. MR and BP are employees of Predica Diagnostics. The other authors declare that they have no competing interests.

### Author details

<sup>1</sup>Department of Biochemistry, Radboudumc, Radboud Institute of Molecular Life Sciences, Geert Grooteplein 26, Nijmegen 6525 GA, The Netherlands. <sup>2</sup>Department of Medical Microbiology, Radboudumc, PO Box 9101, Nijmegen 6500 HB, The Netherlands. <sup>3</sup>Predica Diagnostics, Toernooiveld 1, Nijmegen 6525 ED, The Netherlands. <sup>4</sup>Department of Pathology, Radboudumc, PO Box 9101, 6500 HB Nijmegen, The Netherlands. <sup>5</sup>PALGA, De Bouw 123, Houten 3991 SZ, The Netherlands. <sup>6</sup>Department of Obstetrics and Gynecology, Catharina Hospital Eindhoven, Michelangelolaan 2, Eindhoven 5623 EJ, The Netherlands. <sup>7</sup>GROW, School for Oncology and Reproductive Biology, Maastricht University, Maastricht, The Netherlands. <sup>8</sup>Center for Molecular and Biomolecular Informatics, Radboud Institute of Molecular Life Sciences, PO Box 9101, Nijmegen 6500 HB, The Netherlands. <sup>9</sup>Department of Gynecology and Obstetrics, Radboudumc, PO Box 9101, Nijmegen 6500 HB, The Netherlands.

Received: 11 February 2022 Accepted: 26 April 2022  
Published online: 09 June 2022

## References

- Chan CK, Aimagambetova G, Ukybassova T, Kongrtay K, Azizan A. Human papillomavirus infection and cervical cancer: epidemiology, screening, and vaccination-review of current perspectives. *J Oncol*. 2019;2019:3257939.
- Schiffman M, Solomon D. Clinical practice. Cervical-cancer screening with human papillomavirus and cytologic cotesting. *N Engl J Med*. 2013;369(24):2324–31.
- Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394–424.
- Walboomers JM, Jacobs MV, Manos MM, Bosch FX, Kummer JA, Shah KV, et al. Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *J Pathol*. 1999;189(1):12–9.
- Ronco G, Dillner J, Elfstrom KM, Tunesi S, Snijders PJ, Arbyn M, et al. Efficacy of HPV-based screening for prevention of invasive cervical cancer: follow-up of four European randomised controlled trials. *Lancet*. 2014;383(9916):524–32.
- Aitken CA, van Agt HME, Siebers AG, van Kemenade FJ, Niesters HGM, Melchers WJG, et al. Introduction of primary screening using high-risk HPV DNA detection in the Dutch cervical cancer screening programme: a population-based cohort study. *BMC Med*. 2019;17(1):228.
- Monitor Dutch Population based cervical screening program. <https://www.rivm.nl/en/cervical-cancer-screening-programme>. 2019.
- Loopik DL, van Drongelen J, Bekkers RLM, Voorham QJM, Melchers WJG, Massuger L, et al. Cervical intraepithelial neoplasia and the risk of spontaneous preterm birth: a Dutch population-based cohort study with 45,259 pregnancy outcomes. *PLoS Med*. 2021;18(6):e1003665.
- Aitken CA, Siebers AG, Matthijssse SM, Jansen EEL, Bekkers RLM, Becker JH, et al. Management and treatment of cervical intraepithelial neoplasia in the Netherlands after referral for colposcopy. *Acta Obstet Gynecol Scand*. 2019;98(6):737–46.
- Scott ML, Woodby BL, Ulicny J, Raikhy G, Orr AW, Songcock WK, et al. Human papillomavirus 16 E5 inhibits interferon signaling and supports episomal viral maintenance. *J Virol*. 2020;94(2):e01582–19.
- Songcock WK, Kim SM, Bodily JM. The human papillomavirus E7 oncoprotein as a regulator of transcription. *Virus Res*. 2017;231:56–75.
- Sveen CW, Kagie MJ, Nagelkerke NJ, Veldhuizen RW, Trimbos JB. Can viral load, semi-quantitatively evaluated, of human papillomavirus predict cytological or histological outcome in women with atypical squamous or glandular cells of undetermined significance cytology? *Eur J Gynaecol Oncol*. 2005;26(4):393–7.
- Hu Z, Zhu D, Wang W, Li W, Jia W, Zeng X, et al. Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism. *Nat Genet*. 2015;47(2):158–63.
- Liu Y, Zhang C, Gao W, Wang L, Pan Y, Gao Y, et al. Genome-wide profiling of the human papillomavirus DNA integration in cervical intraepithelial neoplasia and normal cervical epithelium by HPV capture technology. *Sci Rep*. 2016;6:35427.
- Cancer Genome Atlas Research N, Albert Einstein College of M, Analytical Biological S, Barretos Cancer H, Baylor College of M, Beckman Research Institute of City of H, Buck Institute for Research on A, Canada's Michael Smith Genome Sciences C, Harvard Medical S, Helen FGCC, et al. Integrated genomic and molecular characterization of cervical cancer. *Nature*. 2017;543(7645):378–84.
- McBride AA, Warburton A. The role of integration in oncogenic progression of HPV-associated cancers. *PLoS Pathog*. 2017;13(4):e1006211.
- Liu CY, Li F, Zeng Y, Tang MZ, Huang Y, Li JT, et al. Infection and integration of high-risk human papillomavirus in HPV-associated cancer cells. *Med Oncol*. 2015;32(4):109.
- Choi YJ, Lee A, Kim JT, Jin HT, Seo YB, Park JS, et al. E2/E6 ratio and L1 immunoreactivity as biomarkers to determine HPV16-positive high-grade squamous intraepithelial lesions (CIN2 and 3) and cervical squamous cell carcinoma. *J Gynecol Oncol*. 2018;29(3):e38.
- Ebisch RM, Siebers AG, Bosgraaf RP, Massuger LF, Bekkers RL, Melchers WJ. Triage of high-risk HPV positive women in cervical cancer screening. *Expert Rev Anticancer Ther*. 2016;16(10):1073–85.
- Diaz D, Santander MA, Chavez JA. HPV-16 E6 and E7 oncogene expression is downregulated as a result of Mdm2 knockdown. *Int J Oncol*. 2012;41(1):141–6.
- Ratnam S, Coutlee F, Fontaine D, Bentley J, Escott N, Ghatage P, et al. Aptima HPV E6/E7 mRNA test is as sensitive as Hybrid Capture 2 Assay but more specific at detecting cervical precancer and cancer. *J Clin Microbiol*. 2011;49(2):557–64.
- Derbie A, Mekonnen D, Woldeamanuel Y, Van Ostade X, Abebe T. HPV E6/E7 mRNA test for the detection of high grade cervical intraepithelial neoplasia (CIN2+): a systematic review. *Infect Agent Cancer*. 2020;15:9.
- Liu S, Minaguchi T, Lachkar B, Zhang S, Xu C, Tenjimbayashi Y, et al. Separate analysis of human papillomavirus E6 and E7 messenger RNAs to predict cervical neoplasia progression. *PLoS One*. 2018;13(2):e0193061.
- Tang S, Tao M, McCoy JP Jr, Zheng ZM. The E7 oncoprotein is translated from spliced E6\*1 transcripts in high-risk human papillomavirus type 16- or type 18-positive cervical cancer cell lines via translation reinitiation. *J Virol*. 2006;80(9):4249–63.
- Lenting K, van den Heuvel C, van Ewijk A, Elmelik D, de Boer R, Tindall E, et al. Mapping actionable pathways and mutations in brain tumours using targeted RNA next generation sequencing. *Acta Neuropathol Commun*. 2019;7(1):185.
- van den Heuvel C, van Ewijk A, Zeelen C, de Bitter T, Huynen M, Mulders P, et al. Molecular profiling of druggable targets in clear cell renal cell carcinoma through targeted RNA sequencing. *Front Oncol*. 2019;9:117.
- Gottgens EL, van den Heuvel CN, de Jong MC, Kaanders JH, Leenders WP, Ansems M, et al. ACLY (ATP Citrate Lyase) mediates radioresistance in head and neck squamous cell carcinomas and is a novel predictive radiotherapy biomarker. *Cancers*. 2019;11(12):1971.
- van den Heuvel C, Loopik DL, Ebisch RMF, Elmelik D, Andralojc KM, Huynen M, et al. RNA-based high-risk HPV genotyping and identification of high-risk HPV transcriptional activity in cervical tissues. *Mod Pathol*. 2020;33(4):748–57.
- The 1988 Bethesda System for reporting cervical/vaginal cytological diagnoses. National Cancer Institute Workshop. *JAMA*. 1989;262(7):931–4.
- de Bitter T, van de Water C, van den Heuvel C, Zeelen C, Eijkelenboom A, Tops B, et al. Profiling of the metabolic transcriptome via single molecule molecular inversion probes. *Sci Rep*. 2017;7(1):11402.
- Arts P, van der Raadt J, van Gestel SHC, Steehouwer M, Shendure J, Hoischen A, et al. Quantification of differential gene expression by multiplexed targeted resequencing of cDNA. *Nat Commun*. 2017;8:15190.
- Boyle EA, O'Roak BJ, Martin BK, Kumar A, Shendure J. MIPgen: optimized modeling and design of molecular inversion probes for targeted resequencing. *Bioinformatics*. 2014;30(18):2670–2.
- Metsalu T, Vilo J. ClustVis: a web tool for visualizing clustering of multivariate data using principal component analysis and heatmap. *Nucleic Acids Res*. 2015;43(W1):W566–70.
- Liaw A, Wiener M. Classification and regression by randomForest. *Rnews*. 2002;2:18–22.
- Blatt AJ, Kennedy R, Luff RD, Austin RM, Rabin DS. Comparison of cervical cancer screening results among 256,648 women in multiple clinical practices. *Cancer Cytopathol*. 2015;123(5):282–8.
- Van Ostade X, Dom M, Tjalma W, Van Raemdonck G. Candidate biomarkers in the cervical vaginal fluid for the (self-)diagnosis of cervical precancer. *Arch Gynecol Obstet*. 2018;297(2):295–311.
- DeFilippis RA, Goodwin EC, Wu L, DiMaio D. Endogenous human papillomavirus E6 and E7 proteins differentially regulate proliferation, senescence, and apoptosis in HeLa cervical carcinoma cells. *J Virol*. 2003;77(2):1551–63.
- Adey A, Burton JN, Kitzman JO, Hiatt JB, Lewis AP, Martin BK, et al. The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature*. 2013;500(7461):207–11.
- Xue Y, Lim D, Zhi L, He P, Abastado JP, Thierry F. Loss of HPV16 E2 protein expression without disruption of the E2 ORF correlates with carcinogenic progression. *Open Virol J*. 2012;6:163–72.
- Schmitt M, Pawlita M. The HPV transcriptome in HPV16 positive cell lines. *Mol Cell Probes*. 2011;25(2-3):108–13.

41. Schiffman M, Hyun N, Raine-Bennett TR, Katki H, Fetterman B, Gage JC, et al. A cohort study of cervical screening using partial HPV typing and cytology triage. *Int J Cancer*. 2016;139(11):2606–15.
42. Banister CE, Liu C, Pirisi L, Creek KE, Buckhaults PJ. Identification and characterization of HPV-independent cervical cancers. *Oncotarget*. 2017;8(8):13375–86.
43. Liu Y, Xu Y, Jiang W, Ji H, Wang ZW, Zhu X. Discovery of key genes as novel biomarkers specifically associated with HPV-negative cervical cancer. *Mol Ther Methods Clin Dev*. 2021;21:492–506.
44. Kim YT, Zhao M. Aberrant cell cycle regulation in cervical carcinoma. *Yonsei Med J*. 2005;46(5):597–613.
45. Ebisch RM, van der Horst J, Hermsen M, Rijstenberg LL, Vedder JE, Bulten J, et al. Evaluation of p16/Ki-67 dual-stained cytology as triage test for high-risk human papillomavirus-positive women. *Mod Pathol*. 2017;30(7):1021–31.
46. Loghavi S, Walts AE, Bose S. CIntec(R) PLUS dual immunostain: a triage tool for cervical pap smears with atypical squamous cells of undetermined significance and low grade squamous intraepithelial lesion. *Diagn Cytopathol*. 2013;41(7):582–7.
47. Zhu Y, Ren C, Yang L, Zhang X, Liu L, Wang Z. Performance of p16/Ki67 immunostaining, HPV E6/E7 mRNA testing, and HPV DNA assay to detect high-grade cervical dysplasia in women with ASCUS. *BMC Cancer*. 2019;19(1):271.
48. Rajkumar T, Sabitha K, Vijayalakshmi N, Shirley S, Bose MV, Gopal G, et al. Identification and validation of genes involved in cervical tumorigenesis. *BMC Cancer*. 2011;11:80.
49. Peres AL, Paz ESKM, de Araujo RF, de Lima Filho JL, de Melo Junior MR, Martins DB, et al. Immunocytochemical study of TOP2A and Ki-67 in cervical smears from women under routine gynecological care. *J Biomed Sci*. 2016;23(1):42.
50. Tosuner Z, Turkmen I, Arici S, Sonmez C, Turna S, Onaran O. Immunocyto-expression profile of ProExC in smears interpreted as ASC-US, ASC-H, and cervical intraepithelial lesion. *J Cytol*. 2017;34(1):34–8.
51. Branca M, Ciotti M, Giorgi C, Santini D, Di Bonito L, Costa S, et al. Up-regulation of proliferating cell nuclear antigen (PCNA) is closely associated with high-risk human papillomavirus (HPV) and progression of cervical intraepithelial neoplasia (CIN), but does not predict disease outcome in cervical cancer. *Eur J Obstet Gynecol Reprod Biol*. 2007;130(2):223–31.
52. Hascitha J, Priya R, Jayavelu S, Dhandapani H, Selvaluxmy G, Sunder Singh S, et al. Analysis of Kynurenine/Tryptophan ratio and expression of IDO1 and 2 mRNA in tumour tissue of cervical cancer patients. *Clin Biochem*. 2016;49(12):919–24.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

