Universal Human Papillomavirus Typing Assay: Whole Genome Sequencing Following Target Enrichment

Tengguo Li[1], Elizabeth R. Unger[1], Dhwani Batra[2], Mili Sheth[2], Martin Steinau[1*], Jean Jasinski[3], Jennifer Jones[3], Mangalathu S. Rajeeven[1]#

[1]Division of High-Consequence Pathogens & Pathology, Centers for Disease Control and Prevention, 1600 Clifton Road, Atlanta, GA 30333; [2]Division of Scientific Resources, Centers for Disease Control and Prevention, 1600 Clifton Road, Atlanta, GA 30333; [3]Agilent Technologies, Inc., Santa Clara, CA

[*]Current affiliation: Division of Global HIV and TB, Centers for Disease Control and Prevention, 1600 Clifton Road, Atlanta, GA 30333

#Author for Correspondence

Mangalathu S. Rajeevan, PhD;
Division of High-Consequence Pathogens & Pathology, Centers for Disease Control and Prevention,
1600 Clifton Road, Atlanta, GA 30329

E-mail: mor4@cdc.gov

Running title: Universal HPV typing by whole genome sequencing

Disclaimer: The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

ABSTRACT

We designed a universal HPV typing assay based on target enrichment and whole genome sequencing (eWGS). The RNA bait included 23,941 probes targeting 191 HPV types and 12 targeting beta-globin as control. We used Agilent SureSelect XT2 protocol for library preparation, Illumina HiSeq 2500 for sequencing and CLC genomics workbench for sequence analysis. Mapping stringency for type assignment was determined based on 8 (6 HPV-positive and 2 HPV negative) control samples. Using the optimal mapping conditions, types were assigned to 24 blinded samples. eWGS results were 100% concordant with Linear Array (LA) genotyping results for 9 plasmid samples and fully or partially concordant for 9 of the 15 cervical-vaginal samples, with 95.83% overall-type specific concordance for LA genotyping. eWGS identified 7 HPV types not included in the LA genotyping. Since this method does not involve degenerate primers targeting HPV genomic regions, PCR bias in genotype detection is minimized. With further refinements aimed at reducing cost and increasing throughput, this first application of eWGS for universal HPV typing could be a useful method to elucidate HPV epidemiology.

Key words: HPV typing, broad-spectrum assay, whole genome sequencing, target enrichment

INTRODUCTION

Human Papillomaviruses (HPV) are double-stranded DNA viruses in the family *Papillomaviridae*. More than 200 genotypes are recognized based on the sequence of the approximately 8 (k)basepairs [(k)bp] circular genome, with variants in each genotype based on sequence relatedness (1,2). Detection and typing of HPV has clinical and public health significance because of the association of some genotypes in the alpha-papillomavirus genus with anogenital and oropharyngeal cancers. As a result, most assays used in HPV epidemiology and natural history studies are directed to these genotypes, and many PCR-based assays use degenerate primers directed to the L1 region of the HPV genome (3,4). Studies examining disease associations of an increasingly broad spectrum of HPV genotypes have been hampered by the need for multiple assays to detect genotypes in alpha, beta and gamma genera (5,6).

Next generation sequencing (NGS) based methods have the promise to improve the spectrum of HPV genotypes detected. Some NGS methods used PCR with consensus primers targeting the L1 region for library preparation, restricting regions of the genome that could be analyzed (7-13). HPV genome sequencing to single or limited HPV genotypes was achieved using the standard protocol for library preparation without any target enrichment (14), or after whole genome amplification by rolling circle amplification (15) or highly multiplexed degenerate primers (16). There are a few reports of NGS using target enrichment to study the mechanistic signatures of integration of a limited number of HPV genotypes in cervical carcinomas (17-19). The goal of this study was to capitalize on the availability of new target enrichment technology to facilitate whole genome sequencing to detect, genotype and potentially characterize variant and integration status of all known HPVs belonging to alpha, beta and gamma genera using a single assay. Target enrichment technology uses hybridization to purify genomic fragments of interest with complementary DNA or RNA baits. DNA/RNA baits were initially used and proved to be highly effective to enrich human exomes for variant calling by deep sequencing (20,21). We developed custom RNA baits specific to all known HPV genomes to enrich the sample for sequencing of whole viral genomes. Here we describe and provide an initial evaluation of this whole genome sequence based approach for broad-spectrum HPV genotype determination in samples with single and multiple HPV infection using the Agilent SureSelect target enrichment technology.

Journal of Clinical Microbiology

MATERIALS AND METHODS

Samples

DNA extracted from three cells lines (ATCC, Manassas, VA) known to include HPV 16 (Caski: ~500 copies/cell, SiHa: 1-2 copies/cell) and HPV 18 (HeLa: ~50 copies/cell) served as HPV positive controls, and water and human placental DNA (Sigma-Aldrich Corporation, St. Louis, MO) were included as HPV negative controls. A panel of coded samples was prepared so that testing and analysis could be performed in a blinded fashion. The blinded set included 9 residual samples from past WHO validation panels [whole HPV genome DNA plasmids diluted to 50,000 copies/sample in human placental DNA (100 ng/50 µL)] and 15 residual DNA extracts from anonymized cervical/vaginal samples previously typed using the Linear Array (LA; Roche Molecular Diagnostics, Indianapolis, IN). Two of the cervical/vaginal extracts were negative for the 37 genotypes targeted by LA and the other 13 samples were selected to include multiple HPV genotypes (median 5; range 2-12). DNA was quantified by fluorescence-based Picogreen assay (Molecular Probes, Inc., Eugene, OR) and volume was adjusted with a buffer (pH 8.0) composed of 10 mM Tris-HCl and 1 mM EDTA (ethylenediaminetetraacetic acid) so that a final concentration of 100 ng/50 µL was used for library preparation. The exception was the water blank (0 ng DNA), and cell line samples that were also prepared at 10 ng DNA to test the effect of reduced DNA input on sequencing results. The final number of samples sequenced was 32: 8 control samples [3 cell line controls (100 ng), 3 cell line controls (10 ng), human placental DNA (100 ng), water (0 ng)] and 24 blinded samples [9 plasmid samples and 15 cervical/vaginal extracts (100 ng)].

HPV whole genome reference sequences

We retrieved the whole genome HPV reference sequences for all 188 HPV genotypes that were available in the PapillomaVirus Episteme (PaVE) database (http://pave.niaid.nih.gov/) as of September 2014 (22). We also retrieved whole genome sequences for three HPV subtypes detected by LA (HPV55, a HPV44 subtype; HPV82A/IS39; and HPV68b) from the National Center for Biotechnology (NCBI) database (http://www.ncbi.nlm.nih.gov/). The 191 HPV reference genomes included in this study (Table S1) ranged in size from 7095bp – 8104bp and were distributed phylogenetically into 5 genera (67 alpha, 47 beta, 74 Gamma, 2 Mu and 1 Nu). The complete reference sequences were used for RNA bait design as well as for sequence mapping.

Design and synthesis of RNA baits

Agilent Technologies Inc. (Santa Clara, CA,) designed a custom RNA bait library using eArray software comprised of 120 bases RNAs complementary to one strand of the whole genomes of all 191 HPV genotypes at two fold coverage (60 bases overlapping). The resulting 23,941 RNA sequences comprised a bait library size of 1.442 (M)bp. The bait library sequences were subjected to BLAST (Basic Local Alignment Search Tool) search against NCBI human genome database, and 21 fragments were identified to have homology to human sequences. Removing these 21 fragments from RNA bait synthesis resulted in 100% coverage for 182 HPV genotypes and 99.17%-99.25% coverage for 9 HPV genotypes (HPV 6, 19, 34, 71, 82A/IS39, 142, 171, 172, 175), for overall design coverage of 99.96%. A BLAST search of the final synthesized bait sequences against the full NCBI database identified 15, 669 unique hits, 99.94% specific to HPV sequences. The other 10 hits (0.064%) shared identity to chimpanzee, feline and macaque papillomaviruses. As an internal control for the quality of samples, we included 12 RNA fragments covering nt2041 to 3480 of the coding sequence of human beta-globin (HBG) gene (1439bp, GenBank #GU324922.1). The custom biotinylated RNA bait library (HPV and HBG) was synthesized at Agilent Technologies. Upon request to the authors, the custom design ID may be shared with those who want to synthesize these probes through Agilent Technologies.

Target enrichment library preparation and sequencing

The work-flow for the preparation of the sequencing library followed Agilent Technologies' SureSelect^{XT2} target enrichment protocol (Version D3) optimized for 100 ng DNA (Figure 1). Briefly, DNA samples (50 µL in 96 well microTUBE plate) were sheared using Covaris LE220 Focused-ultrasonicator with the software SonoLab 7.3.2.4 (Covaris, Inc., Woburn, MA). The conditions used for shearing (duty factor, 15%; peak incident power, 450; treatment time, 490 seconds) generated DNA fragments with a peak around 150bp as determined by the high sensitivity DNA kit and Bioanalyzer 2100 (Agilent Technologies). DNA fragments were end-repaired, 3' A-tailed, and underwent pre-capture indexing whereby each sample was barcoded with an index of 8bp sequence. Indexed libraries were amplified by PCR with limited number of cycles (8 cycles), followed by purification and assessment of the quality and quantity of each library by Bioanalyzer 2100 (Agilent Technologies). Indexed libraries with unique barcodes were pooled (16 samples/pool) for overnight hybridization

with custom RNA bait, followed by capture of hybridized fragments and 14 cycles of PCR to amplify indexed libraries. Following further purification, the quality and quantity of post-capture HPV enriched pooled libraries were assessed by Bioanalyzer 2100 and qPCR using KAPA DNA library quantification kit (KAPA Biosystems, Wilmington, MA) and LightCycler 480 (Roche Diagnostics, Indianapolis, IN). Pooled libraries were paired-end sequenced on a two-lane flow cell on Illumina HiSeq 2500 at the Centers for Disease Control (CDC) Core Facility using TruSeq Rapid SBS Kit HS (200 cycle) (Illumina, San Diego, CA,) with Rapid run mode according to the following settings: Read1, 100 cycles; index (i7), 9 cycles; read 2, 100 cycles. For each DNA library, a seeding concentration of 5.33pM was applied for cluster generation with 5% PhiX virus genome library added as sequencing control.

Bioinformatics

The raw sequence data were de-multiplexed, and the adaptors and barcodes were removed using Illumina BCl2fastq V1.8.4. Reads with base quality Q score were exported as fastq files for batch mapping to reference sequences using CLC genomics workbench 7.5 (CLCbio, Waltham, MA). The reference sequences of 191 HPV types/subtypes and HBG that were used for bait design were imported to CLC genomics for mapping.

The analysis was directed to HPV detection and typing so duplicate reads were not removed. We used reads with 0 mismatches in the index sequence for analysis in this report. Two parameters, read lengths (L0.5 or L1) and similarity scores (S0.8 or S1.0), were used to adjust mapping stringency. BAM (Binary Alignment/Map) files were analyzed to generate mapping statistics by CLC genomics. Three additional parameters were evaluated in the process of differentiating signal from noise: number of reads mapped to the HPV type-specific reference sequence, average depth of coverage for the mapped sequences and fraction of HPV genome covered by mapped reads. HPV types in 8 control samples detected by enriched whole genome sequencing (eWGS) with different mapping stringency and acceptance parameters were recorded prior to un-blinding expected HPV results from 24 blinded samples. Subsequent matching of HPV genotype calls by WGS to expected results was used to evaluate performance of the assay and analysis.

RNA bait performance was evaluated based on results from single HPV plasmids: percentage HPV reference genome covered by sequenced reads, percentage of reads that mapped to predicted types (compared to

182 total HPV reads), and uniformity of coverage. To evaluate the uniformity, the reference sequences were divided
183 into bins consisting of 300bp, and the average mapping depth within each bin calculated by Strand NGS software
184 (http://www.strand-ngs.com/) using BAM files generated by CLC genomics under L1-S1 mapping stringency.
185 The mean and the standard deviation (SD) of the average mapping depths among the bins were calculated. The
186 uniformity of coverage was calculated as the percentage of bins with coverage within the average read depth
187 ±2SD for all bins. The target enrichment factor for a sample was calculated using formula: (Total HPV mapped
188 reads/Total HPV and human mapped reads)/(HPV genome size/human diploid genome size) (23). HPV and
189 human diploid genome sizes were considered 8 (k)bp and 6.6 (G)bp respectively for this calculation.
190
191 **RESULTS**
192 **Assessment of sequence data quality**
193 DNA sheared approximately to 150bp is expected to increase in size to around 300bp in the indexed
194 library after ligation of adaptors that enable limited PCR amplification and sequencing of the library. As expected,
195 bioanalyzer analysis of the indexed pooled libraries prepared for sequencing indicated the size distribution with
196 peak around 300bp. The seed concentration of 5.3pM generated cluster densities of 1173K and 1183K/mm^2 in
197 pool 1 (samples 1-16) and 2 (samples 17-32), respectively. All raw reads from both pools passed the default
198 filtering of the Illumina BCl2fastq V1.8.4 software (pool 1- 268,096,524 and pool 2 -226,336,668). The average
199 number of reads per sample in pool 1 was 8,538,564 and those from Caski 100 ng (sample 11) with ~500 HPV16
200 copies/cell dominated, comprising 21.9% of the total (Figure 2A). The average number of reads per sample in
201 pool 2 was 11,224,892. The mean base quality Q score for each sample (excluding sample 9, water control)
202 ranged from 34.57 to 36.87 (mean=Q35.6), and 88% of the bases had quality scores greater than 30 (Figure 2B).
203 The water control generated 3,692 reads with Q scores ranging from 2 to 40 (mean=Q25) with only 47 % of bases
204 having Q score greater than 30.
205 **Mapping results for internal control HBG**
206 The fraction of the globin reference sequence mapped in all samples with genomic DNA (all except water
207 control, sample 9) ranged from 93-100% (mean 98%). Samples with 100 ng genomic DNA (samples 1-8, 10-11,

Journal of Clinical Microbiology

13, 15, and 17-32) generated a mean of 7406 reads (range 1903–11225) that, using L1-S1 stringency, mapped to the globin reference (HBG nt2041-3480) with an average depth of coverage ranging from 132.2–779.5 (mean coverage 514.3) (Figure 3). The number of mapped reads (range 902-3327, mean 2000.3) and coverage (range 62.6 - 231, mean 138.9) were reduced in samples with 10 ng genomic DNA (samples 12, 14 and 16). Only 2 reads from the water control (sample 9) mapped to beta globin (depth of coverage 0.1 and fraction of reference sequence mapped 12%).

Evaluating cut-off to improve signal/noise ratio for HPV genotyping from whole genome sequence data

We evaluated mapping results for the 8 control samples (Table 1) in terms of the number of reads, average coverage and the fraction of genome covered using different stringencies to differentiate signal from noise for determination of HPV genotype. Caski (100ng) generated a total of 50,890,604 reads mapped to any HPV reference sequence, of which 99.94% of reads (50,861,070 reads) mapped specifically to HPV16 under the less stringent L0.5-S0.8 mapping condition. WGS also detected HPV 16 in SiHa (134,664 out of 164,138 total reads, 82.04%) and HPV 18 in HeLa (314,969 out of 342,887 total reads, 91.86%) as the most dominant types. Without any cut-offs for non-specific signal, all positive controls detected additional HPV types under all three mapping stringencies. The presence of non-specific signal was also seen in water and placental DNA negative controls. For example, WGS detected HPV 16 with 1800 reads in the negative placental DNA using L0.5-S0.8. With increased mapping stringency (L1-S1), placental DNA still detected HPV 16 with 861 reads and 10.9 average coverage. Based on this, we selected a cut-off of ≥1000 mapped reads and average coverage ≥20 for reliable sequence assignment. We also added the fraction of reference genome covered ≥0.5 (to indicate that at least 50% of viral genome is retained, allowing for loss due to potential integration events) to the cut-off parameters. With L1-S1 mapping stringency and the selected cut-offs (number of mapped reads ≥1000, average coverage ≥20 and fraction of genome covered ≥0.5), HPV genotyping results for the 8 controls were concordant with the expected results (Table 1). The control cell line samples, Caski (~500 HPV16 copies/cell), Hela (~50 HPV 18 copies/cell), and SiHa (1-2 HPV16 copies/cell), vary in known copy numbers of HPV and were analyzed at two concentrations of input DNA (100 ng, 10 ng). In each case, the number of mapped reads under stringent conditions roughly correlated with copy number (Table 1) and type assignment could be made with input of 10 ng

DNA. The fraction of genome covered for HPV18 in HeLa cell line was only 63%. No reads mapped to a 2.6 (k)bp central region (nt3100-5730), compatible with a deletion (Figure 4).

Determination of HPV types from whole genome sequence data

eWGS data from the remaining 24 samples were analyzed using criteria set in the control samples to assign WGS HPV results without knowledge of prior HPV data. Comparing eWGS typing results with prior LA results, type-specific HPV detection was at least partially concordant in 18 of 24 samples (75%) (Table 2). Genotype determinations for the 9 HPV plasmids (HPV types 45, 58, 31, 33, 52, 6, 18, 11 and 16) were 100% concordant. The average depth of coverage for the 9 WHO plasmids ranged from 1,210 to 3,751 with a mean of 2,837. Among the 15 cervico-vaginal swab samples, typing results for 9 samples were fully or partially concordant. Of these 9, full concordance was found in 5 samples, 2 negative for HPV and 3 samples with multiple types (3, 4 and 8 types). Among the 6 discordant samples, 4 were negative by eWGS but positive for multiple types by LA whereas 2 samples were HPV positive by both methods but for different types. The overall type-specific concordance for types included in the LA assay was 95.83%, and for LA targeted types, there were no instances of samples positive by eWGS and negative by LA. In six samples, eWGS identified HPV types not included in the LA assay (HPV 30, 43, 68a, 87, 90, 91, and 114). In cervico-vaginal samples with multiple HPV types, the average depth of coverage varied greatly from as low as 27 to 129,998 (Table 2), probably a result of varying viral loads.

While some LA types were not detected by WGS in this small sample set (HPV 26, 35, 53, 72, 81, and 84), other LA types were detected in some, but not all samples (HPV 18 and 66). In some instances, eWGS detected types but did not meet the criteria for number of reads or depth of coverage. For example, in sample 25, eWGS failed to call HPV 66 and 18 which were detected by LA, but 1478 reads mapped to HPV66 with 99% of genome covered but failed because average depth of coverage (18.9) was just below cutoff (20); and HPV 18 had a subthreshold 234 reads mapped with 71% genome coverage. In sample 20, eWGS failed to call HPV 72, although 1412 reads mapped to HPV 72 with 87% genome coverage but average depth of coverage 17.7 was below cutoff 20.

In sample 29, LA detected HPV84 but eWGS detected HPV 114. HPV84 and HPV114 have 84% identity over the whole genome, suggesting the possibility of misclassification based on genomic fragments. To explore this, sample 29 reads were mapped with HPV84 as the only reference sequence. While no reads mapped to HPV 84 under stringent conditions, under L0.5-S0.8 conditions 62,203 reads mapped to HPV84 with 94% of the genome covered. HPV 53 was not identified in any of the 4 samples that were LA positive for this type, however HPV 53 sequences were found in one of four samples (number of reads, 1438; average coverage, 18; and genome coverage, 99%) if mapping was done under less stringent conditions (L0.5-S0.8).

LA includes probes for HPV 64 which has been reclassified as a subtype of HPV 34. For sample 17, 16,120 reads mapped to HPV34 with 16% coverage (compared to whole genome), whereas 1736 reads mapped to 458 bp HPV64 L1 partial sequence (GenBank# AJ81226.1) with 85% coverage. As the full length reference sequence for HPV 64 is not available and HPV 64 is reclassified as a subtype of HPV34, the eWGS results were assigned to HPV 64 based on post-hoc assessment, giving partial concordance with LA results

Evaluation of Custom HPV Bait

The performance of the HPV custom RNA bait pool for the 9 HPV types included as individual plasmids is shown in Table 3. One HPV 6 bait sequence had to be removed due to homology with the human genome, resulting in 99.2% design coverage for that type; design coverage was 100% for other 8 types. The mean coverage of mapped reads to reference sequences was 99.8% (range 99.2% – 100%) and 99.2% to 100% of HPV reads mapped to predicted type (mean 99.60%). The mapped reads gave 95.9% uniform mean coverage (range 92.6 %–96.3%) suggesting that the bait resulted in unbiased enrichment of the whole HPV genome. The eWGS method averaged 184,483-fold enrichment for HPV sequences in HPV positive samples (range 3,294–914,377).

DISCUSSION

This is the first application of NGS for whole genome identification of essentially all known HPV types (191 HPV types in the alpha, beta and gamma genera) using RNA baits of the Agilent SureSelect target enrichment technology. The target enrichment method avoids the limitations of targeting only limited areas of the genome, minimizes the potential for PCR bias and significantly increases the potential to increase the number of

Journal of Clinical Microbiology

types identified in a single assay. A recent report of using Roche NimbleGen DNA bait based target enrichment largely focused on understanding the mechanistic signatures of integration of HPV types (87 types, 63% alpha) in cervical carcinomas (17) with limited data on NGS-based HPV type determination and concordance with a current HPV typing assay. In comparison, we mainly focused on the development of a universal HPV typing assay to detect all known HPV types in epidemiologic studies that increasingly examines a broad-spectrum of HPV in alpha, beta and gamma genera in both mucosal and cutaneous specimens for their disease association (5,6,11,24). Towards this goal, we evaluated our method in terms of the performance of RNA baits for whole genome identification, level of enrichment, and a number of read mapping metrics for determination of HPV types under single and multiple infection.

The custom RNA bait library exhibited excellent performance in terms of the fraction of genome coverage, percentage of on-target reads mapped to predicted HPV types, uniformity of coverage for the types that we evaluated, and by the observed average level of target enrichment (184,483) by a factor of nearly 5 log. Our results support the concept of specific capture and whole genome sequencing of viruses from clinical samples through target enrichment technologies (18,25,26). Sequencing the target enriched libraries generated high quality reads with more than 88% of bases having Q scores greater than 30. Mapping results revealed that this method could detect single and multiple infections of HPV along with HBG as an internal control for cellularity.

The mapping stringencies and threshold for the number of mapped reads required to avoid "bleed through" of non-specific HPV signals was established using 8 control samples ([3 cell line controls (100 ng), 3 cell line controls (10 ng), human placental DNA (100 ng), water (0 ng)]. The most stringent L1-S1 mapping condition along with a cut-off of ≥1000 mapped reads, average coverage ≥20 and fraction of genome covered greater than 50% allowed reliable sequence assignment. The control cell lines, Caski (~500 HPV16 copies/cell), Hela (~50 HPV 18 copies/cell), and SiHa (1-2 HPV16 copies/cell), vary in copy numbers of HPV and were analyzed at two concentrations of input DNA (100 ng, 10 ng in 50 µL). In each case, the number of mapped reads under stringent conditions roughly correlated with copy number, and type assignment could be made with an input of 10 ng DNA. In agreement with previous reports by other methods (27,28), the fraction of genome

Journal of Clinical Microbiology

covered for HPV18 in HeLa was only 63% since no reads mapped to a 2.6 (k)bp central region (nt3100-5730), compatible with a deletion due to integration.

The bleed-through of reads from dominant sequences in adjacent clusters has been reported in multiplexed samples sequenced using the Illumina platform (29,30). This might happen more frequently under relatively high seeding density for cluster generation and when one sequence dominates. These two conditions seemed to have taken place in this study. At the seeding density of 5.3 pM used in this study, pool 1 and pool 2 libraries generated cluster densities of 1173K and 1183K respectively which is slightly higher than the recommended cluster density of 850-1000K. Additionally, over 31 million reads of HPV 16 from Caski 100 ng dominated sequences and corresponded to the HPV16 bleed-through found in other samples in the same pool. The HPV 16 sequences in the placental DNA negative control in pool 1 had the unique 29 nucleotide substitutions identical to Caski HPV16 sequences (data not shown). It is unlikely that misidentification of barcode sequences contributed to bleed-through since we restricted the analysis to reads with 0 mismatches in the index reads.

We applied the mapping stringency and cut-off determined with the control samples for determination of HPV types in 24 samples blinded to their HPV status. The HPV types determined by the eWGS were 100% concordant with LA in the 9 plasmid samples but concordance was lower in the cervi-covaginal samples (Table 2). Nine of 15 samples (60%) showed complete or partial concordance and type-specific concordance for the 37 LA types was 95.83%. It could be anticipated that biologic samples from exfoliated cells would present greater challenges to HPV detection than purified plasmid DNA. These samples vary significantly in the amount of cellular and viral material. Given the relation between viral copy number and number of reads noted in the plasmid and cell line results, it could be suspected that low copy numbers may contribute to failure to detect types found by LA. Additionally, LA used 10 µl of extract regardless the DNA concentration, while the eWGS method used 100 ng. Of the 4 samples that were LA positive but eWGS negative, three had 6-10 times (600 ng – 1030 ng) more DNA input for LA. On the other hand, even with only 15 epidemiologic samples, eWGS detected 7 HPV types (HPV 30, 43, 68a, 87, 90, 91, and 114) not targeted by LA. The ability of this eWGS method to detect HPV types not targeted by current assays indicates it could be useful in evaluation of HPV negative cervical

lesions (31) and in epidemiologic studies of HPV in geographic regions that may have uncommon types in circulation (32).

Some HPV types detected by LA were not detected by eWGS in this small sample set (HPV 26, 35, 53, 72, 81, 84). HPV 84 was also reported discordant between NGS and LA in earlier studies (10,15). Further evaluation is needed to determine explanations for the failure to detect these types. Reduced mapping stringency did detect HPV 53 and 84 in some instances, and there could be unreported variants affecting mapping, and/or these types may be present at low copy numbers. We were able to confirm the specificity of RNA baits for HPV53 at the highest stringency of reference mapping by detecting HPV53 plasmid DNA (data from subsequent study not shown).

In conclusion, we developed and provided results from the initial evaluation of a non-PCR whole genome sequence based approach which is as possibly close to a "gold standard" for broad-spectrum HPV genotype determination. The method relies on a custom-made RNA bait library that showed excellent performance in terms of genome coverage, percentage of reads mapped to predicted HPV type, uniformity of coverage, and the level of target enrichment. Further optimization and evaluation of this eWGS method for HPV detection is required in terms of cost/sample and throughput. Additional studies with samples including more types and optimization of sequencing conditions to minimize the bleed-through are needed. Studies evaluating the reproducibility and lower limits of detection are planned. Finally, it should be noted that despite enrichment with RNA bait by a factor of 5 log, 75% of the total reads (average) were human (off-target), with the exception of the sample with the highest viral load (100 ng Caski) in which 36% of total reads were off-target (data not shown). This suggests adjustment in the conditions for on-target enrichment could enhance results. Future laboratory and bioinformatics efforts are focused on determining the sensitivity and reproducibility of this HPV genotyping method, along with expanding its application for HPV detection and genotyping in samples for a variety of anatomical sites, and development of a bioinformatics pipeline for automatic determination of HPV types and variants. This methodology is conceptually suitable for detecting both known and unknown HPV types through adjusting stringencies at the level of hybridization to RNA bait and reference mapping parameters. The current reference mapping stringency used to detect known HPV genotypes would likely miss unknown types.

ACKNOWLEDGMENTS

Support for Tengguo Li was provided by the research participation program at the Centers for Disease Control and Prevention (CDC), National Center for Emerging and Zoonotic Infectious Diseases, Division of High-Consequence Pathogens and Pathology, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the CDC. The authors wish to acknowledge the support of CDC HPV laboratory team members with samples, and the CDC core facility with sequencing related activities and advice. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

1. **Burk RD, Harari A, Chen Z**. 2013. Human papillomavirus genome variants. Virology **445**:232-243.
2. **de Villiers EM, Fauquet C, Broker TR, Bernard HU, zur HH**. 2004. Classification of papillomaviruses. Virology **324**:17-27.
3. **Cubie HA, Cuschieri K**. 2013. Understanding HPV tests and their appropriate applications. Cytopathology **24**:289-308.
4. **Poljak M, Kocjan BJ**. 2010. Commercially available assays for multiplex detection of alpha human papillomaviruses. Expert.Rev.Anti.Infect.Ther. **8**:1139-1162.
5. **Chouhy D, Gorosito M, Sanchez A, Serra EC, Bergero A, Fernandez BR, Giri AA**. 2010. New generic primer system targeting mucosal/genital and cutaneous human papillomaviruses leads to the characterization of HPV 115, a novel Beta-papillomavirus species 3. Virology **397**:205-216.
6. **Pierce Campbell CM, Messina JL, Stoler MH, Jukic DM, Tommasino M, Gheit T, Rollison DE, Sichero L, Sirak BA, Ingles DJ, Abrahamsen M, Lu B, Villa LL, Lazcano-Ponce E, Giuliano AR**. 2013. Cutaneous human papillomavirus types detected on the surface of male external genital lesions: a case series within the HPV Infection in Men Study. J.Clin.Virol. **58**:652-659.
7. **Arroyo LS, Smelov V, Bzhalava D, Eklund C, Hultin E, Dillner J**. 2013. Next generation sequencing for human papillomavirus genotyping. J.Clin.Virol. **58**:437-442.
8. **Militello V, Lavezzo E, Costanzi G, Franchin E, Di CB, Toppo S, Palu G, Barzon L**. 2013. Accurate human papillomavirus genotyping by 454 pyrosequencing. Clin.Microbiol.Infect. **19**:E428-E434.
9. **Yi X, Zou J, Xu J, Liu T, Liu T, Hua S, Xi F, Nie X, Ye L, Luo Y, Xu L, Du H, Wu R, Yang L, Liu R, Yang B, Wang J, Belinson JL**. 2014. Development and validation of a new HPV genotyping assay based on next-generation sequencing. Am.J.Clin.Pathol. **141**:796-804.

10. **Yin L, Yao J, Chang K, Gardner BP, Yu F, Giuliano AR, Goodenow MM**. 2016. HPV Population Profiling in Healthy Men by Next-Generation Deep Sequencing Coupled with HPV-QUEST. Viruses. **8**:28.

11. **Agalliu I, Gapstur S, Chen Z, Wang T, Anderson RL, Teras L, Kreimer AR, Hayes RB, Freedman ND, Burk RD**. 2016. Associations of Oral alpha-, beta-, and gamma-Human Papillomavirus Types With Risk of Incident Head and Neck Cancer. JAMA Oncol. **2**:599-606.

12. **Ambulos NP, Jr., Schumaker LM, Mathias TJ, White R, Troyer J, Wells D, Cullen KJ**. 2016. Next-Generation Sequencing-Based HPV Genotyping Assay Validated in Formalin-Fixed, Paraffin-Embedded Oropharyngeal and Cervical Cancer Specimens. J.Biomol.Tech. **27**:46-52.

13. **da Fonseca AJ, Galvao RS, Miranda AE, Ferreira LC, Chen Z**. 2016. Comparison of three human papillomavirus DNA detection methods: Next generation sequencing, multiplex-PCR and nested-PCR followed by Sanger based sequencing. J.Med.Virol. **88**:888-894.

14. **Conway C, Chalkley R, High A, Maclennan K, Berri S, Chengot P, Alsop M, Egan P, Morgan J, Taylor GR, Chester J, Sen M, Rabbitts P, Wood HM**. 2012. Next-generation sequencing for simultaneous determination of human papillomavirus load, subtype, and associated genomic copy number changes in tumors. J.Mol.Diagn. **14**:104-111.

15. **Meiring TL, Salimo AT, Coetzee B, Maree HJ, Moodley J, Hitzeroth II, Freeborough MJ, Rybicki EP, Williamson AL**. 2012. Next-generation sequencing of cervical DNA detects human papillomavirus types not detected by commercial kits. Virol.J. **9**:164.

16. **Cullen M, Boland JF, Schiffman M, Zhang X, Wentzensen N, Yang Q, Chen Z, Yu K, Mitchell J, Roberson D, Bass S, Burdette L, Machado M, Ravichandran S, Luke B, Machiela MJ, Andersen M, Osentoski M, Laptewicz M, Wacholder S, Feldman A, Raine-Bennett T, Lorey T, Castle PE, Yeager M, Burk RD, Mirabello L**. 2015. Deep sequencing of HPV16 genomes: A new high-throughput tool for exploring the carcinogenicity and natural history of HPV16 infection. Papillomavirus.Res. **1**:3-11.

17. **Holmes A, Lameiras S, Jeannot E, Marie Y, Castera L, Sastre-Garau X, Nicolas A**. 2016. Mechanistic signatures of HPV insertions in cervical carcinomas. Npj Genomic Medicine **1**:16004-16019.

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

19. **Liu Y, Lu Z, Xu R, Ke Y**. 2016. Comprehensive mapping of the human papillomavirus (HPV) DNA integration sites in cervical carcinomas by HPV capture technology. Oncotarget. **7**:5852-5864.

20. **Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C**. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat.Biotechnol. **27**:182-189.

21. **Varela I, Tarpey P, Raine K, Huang D, Ong CK, Stephens P, Davies H, Jones D, Lin ML, Teague J, Bignell G, Butler A, Cho J, Dalgliesh GL, Galappaththige D, Greenman C, Hardy C, Jia M, Latimer C, Lau KW, Marshall J, McLaren S, Menzies A, Mudie L, Stebbings L, Largaespada DA, Wessels LF, Richard S, Kahnoski RJ, Anema J, Tuveson DA, Perez-Mancera PA, Mustonen V, Fischer A, Adams DJ, Rust A, Chan-on W, Subimerb C, Dykema K, Furge K, Campbell PJ, Teh BT, Stratton MR, Futreal PA**. 2011. Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. Nature **469**:539-542.

Journal of Clinical Microbiology

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

23. **Ware JS, John S, Roberts AM, Buchan R, Gong S, Peters NS, Robinson DO, Lucassen A, Behr ER, Cook SA**. 2013. Next generation diagnostics in inherited arrhythmia syndromes : a comparison of two approaches. J.Cardiovasc.Transl.Res. **6**:94-103.

24. **Chockalingam R, Downing C, Tyring SK**. 2015. Cutaneous Squamous Cell Carcinomas in Organ Transplant Recipients. J.Clin.Med. **4**:1229-1239.

25. **Depledge DP, Palser AL, Watson SJ, Lai IY, Gray ER, Grant P, Kanda RK, Leproust E, Kellam P, Breuer J**. 2011. Specific capture and whole-genome sequencing of viruses from clinical samples. PLoS.One. **6**:e27805.

26. **Brown JR, Roy S, Ruis C, Yara RE, Shah D, Williams R, Breuer J**. 2016. Norovirus whole genome sequencing by SureSelect target enrichment: a robust and sensitive method. J.Clin.Microbiol. **54**:2530-2537.

27. **Sun H, Chen C, Lian B, Zhang M, Wang X, Zhang B, Li Y, Yang P, Xie L**. 2015. Identification of HPV integration and gene mutation in HeLa cell line by integrated analysis of RNA-Seq and MS/MS data. J.Proteome.Res. **14**:1678-1686.

28. **Adey A, Burton JN, Kitzman JO, Hiatt JB, Lewis AP, Martin BK, Qiu R, Lee C, Shendure J**. 2013. The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. Nature **500**:207-211.

29. **Kircher M, Sawyer S, Meyer M**. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. **40**:e3.

30. **Mitra A, Skrzypczak M, Ginalski K, Rowicka M**. 2015. Strategies for achieving high sequencing accuracy for low diversity samples and avoiding sample bleeding using illumina platform. PLoS.One. **10**:e0120520.

31. **Petry KU, Cox JT, Johnson K, Quint W, Ridder R, Sideri M, Wright TC, Jr., Behrens CM**. 2016. Evaluating HPV-negative CIN2+ in the ATHENA trial. Int.J.Cancer **138**:2932-2939.

32. **Flores-Miramontes MG, Torres-Reyes LA, Alvarado-Ruiz L, Romero-Martinez SA, Ramirez-Rodriguez V, Balderas-Pena LM, Vallejo-Ruiz V, Pina-Sanchez P, Cortes-Gutierrez EI, Jave-Suarez LF, Aguilar-Lemarroy A**. 2015. Human papillomavirus genotyping by Linear Array and Next-Generation Sequencing in cervical samples from Western Mexico. Virol.J. **12**:161.

473 **Table 1. Determination of cut-off to differentiate signal from noise in HPV whole genome sequence data**
474 **based using control samples**

Control samples	L0.5-S0.8*			L0.5-S1*			L1-S1*			HPV type	
	No of reads mapped	Average coverage	Fraction of genome covered	No of reads mapped	Average coverage	Fraction of genome covered	No of reads mapped	Average coverage	Fraction of genome covered	Without Cut-off	With Cut-off**
Caski (100 ng)	50,861,070	637,921.0	1.00	49,687,563	624,209.5	1.00	31,807,177	402,316.9	0.99	16	16
	187	2.1	0.63	139	1.7	0.57	116	1.5	0.50	33	
Caski (10 ng)	9,020,079	113,103.0	1.00	8,807,582	110,605.2	1.00	5,582,647	70,612.8	0.99	16	16
	37	0.4	0.16	14	0.2	0.12	10	0.1	0.10	45	
SiHa (100 ng)	134,664	1,679.0	1.00	130,955	1,638.0	1.00	87,797	1,110.5	0.98	16	16
	75	0.9	0.22	65	0.8	0.19	47	0.6	0.12	18	
SiHa (10 ng)	68,146	850.0	1.00	66,413	831.1	1.00	44,921	568.2	0.98	16	16
	94	1.0	0.33	66	0.8	0.30	59	0.7	0.29	31	
Hela (100 ng)	314,969	3,953.7	0.67	298,994	3,760.7	0.67	183,032	2,329.5	0.64	18	18
	1,339	16.0	1.00	1,100	13.8	1.00	659	8.3	0.90	16	
Hela (10 ng)	83,489	1,049.0	66.49	78,847	992.9	66.49	48,305	614.8	0.62	18	18
	610	7.3	0.99	409	5.1	0.98	225	2.8	0.77	16	
H_2O	115	1.5	0.29	110	1.4	0.27	69	0.9	0.18	18	Negative
	12	0.1	0.13	2	0.0	0.02	1	0.0	0.01	16	
Placenta	1,800	21.3	1.00	1378	17.3	1.00	861	10.9	0.88	16	Negative
	46	0.5	0.26	37	0.5	0.24	23	0.3	0.16	18	

475

476

477 * Three mapping stringencies (L0.5-S0.8, L0.5-S1 and L1-S1) were evaluated using combination of parameters L
478 and S that represent read length (0.5 or 1 and similarity score (0.8 or 1), respectively.

479 **HPV types determined following cut-off on number of mapped reads ≥1000, average coverage ≥20 and the
480 fraction of genome covered ≥0.5 under L1-S1 mapping stringency.

481

482

483

484

485 **Table 2. HPV genotype determination in blinded samples based on target enrichment and whole genome**
486 **sequencing**

Sample No[$]	Sample type	No of reads mapped	Average coverage	Fraction of genome covered (≥0.5)	HPV type (WGS result)[*]	HPV type (LA result)	Concordance[&]
1	Plasmid	282773	3598.5	0.99	**45**	**45**	Yes
2	Plasmid	293528	3751.6	1.00	**58**	**58**	Yes
3	Plasmid	210381	2659.0	1.00	**31**	**31**	Yes
4	Plasmid	190903	2413.7	0.99	**33**	**33**	Yes
5	Plasmid	237806	2994.3	1.00	**52**	**52**	Yes
6	Plasmid	268460	3357.4	1.00	**6**	**6**	Yes
7	Plasmid	95100	1210.4	1.00	**18**	**18**	Yes
8	Plasmid	169355	2135.4	1.00	**11**	**11**	Yes
17[#]	Cervicovaginal	16120	208.7	0.16	(34)		Yes, partial
		1736	379	0.85	**64 (34 subtype)**	**64 (reclassified as 34 subtype)**	
						81	
18	Cervicovaginal	7873	100.9	0.99	**67**	**67**	Yes
		10820	139.5	0.97	**54**	**54**	
		8012	101.4	0.91	**70**	**70**	
		3967	49.4	0.99	*90*[*]		
19	Cervicovaginal	667344	8554.6	0.90	**67**	**67**	Yes
		675774	8535.7	0.97	**42**	**42**	
		680615	8425.5	0.91	**89**	**89**	
		345249	4372.5	0.82	**59**	**59**	
		170609	2105.2	0.53	**83**	**83**	
		36981	472.7	1.00	**66**	**66**	
		32800	419.2	0.72	**58**	**58**	
		13844	176.5	0.91	**56**	**56**	
20	Cervicovaginal				HPV-	72, 35, 52, 53, 54, 62, 81	No
21	Cervicovaginal				**HPV-**	**HPV-**	Yes
22	Cervicovaginal				HPV-	6, 53, 56, 62, 70	No
23	Cervicovaginal	135138	1711.5	0.73	**59**	**59**	Yes, partial
		110107	1392.2	0.99	**40**	**40**	
		95442	1193.2	0.99	*87*[*]		
		73386	940.7	0.84	**67**	**67**	
		49125	638.0	0.87	**73**	**73**	
		36101	452.7	0.98	*43*[*]		
		29721	375.9	0.97	**16**	**16**	
		13891	171.4	0.73	**83**	**83**	
		9380	119.9	0.99	**66**	**66**	
		3091	39.5	0.99	*68a**		
		2121	27.0	0.85	**56**	**56**	
						45, 51, 61, 89	
24	Cervicovaginal	2503879	31916.9	0.99	**56**	**56**	Yes, partial
		2221727	27974.4	1.00	**52**	**52**	
		1054024	13048.1	0.99	**89**	**89**	
		598662	7506.7	0.90	*43*[*]		
		266588	3318.7	0.98	*90*[*]		
		213130	2692.1	0.95	**42**	**42**	
		50171	628	0.66	**61**	**61**	
		40905	523.9	0.99	**51**	**51**	
		31221	390.3	0.98	*87*[*]		
						53, 68b	
25	Cervicovaginal				HPV-	66, 18, 31	No

Journal of Clinical Microbiology

26	Cervicovaginal	10277696	129998.7	1.00	**16**	**16**	Yes, partial
		428009	5411.7	0.97	**40**	**40**	
		184266	2310.5	0.86	*43*[*]		
		127862	1609.9	0.99	**52**	**52**	
		118695	1490.0	0.83	*91*[*]		
		90888	1148.0	0.99	**42**	**42**	
		3722	46.0	0.66	**62**	**62**	
		2610	33.3	0.96	**39**	**39**	
						72, 83	
27	Cervicovaginal				HPV-	26, 42, 58, 83, 84	No
28	Cervicovaginal	9041	115.1	0.99	*30*[*]	35, 53	No
29	Cervicovaginal	23813	295.1	0.76	*114*[*]	40, 54, 66, 84, 89	No
30	Cervicovaginal	2102063	26836.0	1.00	**39**	**39**	Yes
		1743872	22288.8	0.99	**66**	**66**	
		81139	1039.2	0.85	**51**	**51**	
		49403	617.8	0.62	**6**	**6**	
		12102	152.6	0.96	**11**	**11**	
31	Cervicovaginal				**HPV-**	**HPV-**	Yes
32	Plasmid	270183	3417.4	1.00	**16**	**16**	Yes

Bold font = types detected by both assays; Italics font = type not included in LA assay

[$]Sample numbers 1-8 multiplexed with control samples in pool 1 and samples 17-32 multiplexed in pool 2. Sequence information is given for only those passed the signal/noise cut-off for HPV type determination

[*]HPV types not included in LA but detected by WGS

[&]HPV concordance based on WGS and LA results.

[#]Note, HPV 64 has been reclassified as subtype of HPV 34. The only reference sequence available for HPV 64 is an L1 fragment. On post-hoc assessment, results for this sample are assigned to HPV 64 based on reads mapped to L1 fragment with good coverage, and reads mapping to HPV 34 with low genome coverage.

Table 3: Performance of different RNA bait evaluation metrics based on WGS mapping results under L1-S1 from plasmid samples

Sample No	HPV type (eWGS result)*	Bait design coverage (%)@	HPV genome coverage by mapped reads$ (%)	HPV reads mapped to predicted type (%)#	Uniformity of coverage (%)&
1	45	100	99.0	99.8	96.3
2	58	100	100	99.7	96.3
3	31	100	100	99.7	96.3
4	33	100	99.2	99.2	96.3
5	52	100	100	99.7	96.3
6	6	99.2	100	99.8	96.3
7	18	100	100	100	96.3
8	11	100	100	99.5	92.6
32	16	100	100	99.8	96.3
Mean		99.9	99.8	99.6	95.9

*HPV type determined by eWGS for the corresponding samples numbers (see Table 2)

@Proportion of the HPV genome covered by the bait design criteria

$Proportion of the HPV reference genome covered by the mapped reads

#Proportion of reads that mapped to HPV predicted type compared to the total HPV reads

&Uniformity of coverage across the genome was calculated as the percentage of bins with coverage within the average read depth ±2SD for all bins (see methods for details).

Figure legends

Figure 1: Laboratory work-flow for HPV genotyping following RNA bait-based target enrichment and whole genome sequencing.

Figure 2: Number of reads (A) and mean base quality (Q) score of reads (B) passing the default filtering of Illumina BCl12fastq V1.8.4 software. Reads restricted to 0 mismatches in 8bp index reads.

Figure 3. Performance of RNA baits for internal control human beta-globin gene based on number of reads (A), average coverage (B) and fraction of reference sequence covered by the reads (C).

Figure 4. Mapping results showing high specificity of RNA baits restricted to the HPV 18 genome integrated to the HeLa genome. Reduced fraction of HPV genome covered by the sequenced reads due to deletion of central region of HPV 18 genome (the central region indicated within the two dashed vertical lines), compatible with a deletion due to integration, is shown reproducibly with 100 ng (A) and 10 ng (B) input DNA. The stringent L1-S1 mapping conditions result in small gaps in the consensus sequence due to mismatched reads. Mapping results schematically aligned to HPV genome with the location of "early" and "late" region genes (C).