Contents lists available at ScienceDirect

Journal of Clinical Virology

journal homepage: www.elsevier.com/locate/jcv

HPV16 whole genome minority variants in persistent infections from young Dutch women

Sonja Lagström[a,b,c,1], Pascal van der Weele[d,e,1], Trine Ballestad Rounge[b], Irene Kraus Christiansen[a,f], Audrey J. King[d,*], Ole Herman Ambur[g,*]

[a] Department of Microbiology and Infection Control, Akershus University Hospital, Lørenskog, Norway
[b] Department of Research, Cancer Registry of Norway, Oslo, Norway
[c] Institute of Clinical Medicine, University of Oslo, Oslo, Norway
[d] National Institute for Public Health and the Environment (RIVM), Centre for Infectious Disease Research, Diagnostics and Screening, Bilthoven, the Netherlands
[e] Vrije Universiteit-University Medical Center (VUmc), Department of Pathology, Amsterdam, the Netherlands
[f] Department of Clinical Molecular Biology (EpiGen), Division of Medicine, Akershus University Hospital and Institute of Clinical Medicine, University of Oslo, Oslo, Norway
[g] Faculty of Health Sciences, OsloMet - Oslo Metropolitan University, Oslo, Norway

ARTICLE INFO

Keywords:
Human papillomavirus
Persistent infection
Genome variation
Mutational signature

ABSTRACT

Background: Chronic infections by one of the oncogenic human papillomaviruses (HPVs) are responsible for near 5% of the global cancer burden and HPV16 is the type most often found in cancers. HPV genomes display unexpected levels of variation when deep-sequenced. Minor nucleotide variations (MNVs) may reveal HPV genomic instability and HPV-related carcinogenic transformation of host cells.
Objectives: The objective of this study was to investigate HPV16 genome variation at the minor variant level on persisting HPV16 cervical infections from a population of young Dutch women.
Study design: 15 HPV16 infections were sequenced using a whole-HPV genome deep sequencing protocol (TaME-seq). One infection was followed over a three-year period, eight were followed over a two-year period, three were followed over a one-year period and three infections had a single sampling point.
Results and conclusions: Using a 1% variant frequency cutoff, we find on average 48 MNVs per HPV16 genome and 1717 MNVs in total when sequencing coverage was > 100 × . We find the transition mutation T > C to be the most common, in contrast to other studies detecting APOBEC-related C > T mutation profiles in pre-cancerous and cancer samples. Our results suggest that the relative mutagenic footprint of HPV16 genomes may differ between the infections in this study and transforming lesions. In addition, we identify a number of MNVs that have previously been associated with higher incidence of high-grade lesions (CIN3+) in a population study. These findings may provide a starting point for future studies exploring causality between emerging HPV minor genomic variants and cancer development.

1. Introduction

Human papillomavirus (HPV) is the most common sexually transmitted infection worldwide [1] and persistent infection with an oncogenic HPV type is required, but not sufficient, for the development of cervical cancer [2]. Although most HPV infections clear naturally within 12-18 months [3], a subset may persist, potentially progressing to cervical intraepithelial lesions of varying degrees (CIN1-3) and invasive cervical cancer. HPV is a double stranded DNA virus that uses host replication machinery and has co-diverged with humans to constitute highly conserved genotypes [4,5]. Within HPV types, distinction is made between lineages (1–10% whole genome genetic difference), sublineages (0.5–1.0%) and variants (< 0.5%) [6]. Lineages and sublineages of HPV have been associated with differential risks for disease outcomes [7,8]. In addition, recent studies have shown that HPV exhibits large variation, both at the population level and within its human host despite the strongly conserved genome [9–14]. Currently, limited information is available explaining the origin of this diversity.

Deep sequencing of HPV genomes has revealed the presence of minor nucleotide variations (MNVs). These polymorphic sites show one or multiple different nucleotides in addition to the consensus or majority nucleotide [15]. Such MNVs can only be reliably detected by

[*] Corresponding authors.
E-mail addresses: audrey.king@rivm.nl (A.J. King), olam@oslomet.no (O.H. Ambur).
[1] S. L. and P. v. d. W. have contributed equally to this work.

https://doi.org/10.1016/j.jcv.2019.08.003
Received 7 February 2019; Received in revised form 6 August 2019; Accepted 9 August 2019

means of high-resolution sequencing. HPV is considered to evolve slowly due to the high fidelity of its human host replication machinery [4,16]. However, humans also encode several low-fidelity polymerases, some of which are upregulated in early stages of HPV16 infection [17]. These polymerases are often recruited for DNA repair by means of non-homologous end-joining [18]. The HPV life cycle involves two separate rounds of replication. An initial round in proliferating cells at the basal layer of the stratified epithelium yielding 10–100 copies of the viral genome per cell, and another productive round, in differentiated cells at the suprabasal level, resulting in thousands of viral copies per cell [19]. Several DNA repair pathways are required for productive HPV replication, yet information relating to their influence on generating mutations at the minor variant level is lacking. In addition, viral mutation rates can be affected by, sequence context, template secondary structure, the cellular microenvironment and several other factors relating to replication, post-replicative corrections and DNA repair [20]. A known source of mutations in HPV genomes is apolipoprotein B mRNA editing enzyme (APOBEC) activity, which is part of the host innate immune response against viruses [21]. APOBEC enzymes induce genetic change by converting cytidine to uridine, which may base pair with adenosine, causing C > T substitution mutants after replication. APOBEC-related changes have been identified in cervical cancer patient genomes [22]. Additionally, the HPV genome is itself susceptible to APOBEC editing [12,23]. HPV oncoproteins E6 and E7 upregulate the expression of APOBEC3A and APOBEC3B [24,25]. In turn, APOBEC3B activity is upregulated in cancer tissues [26,27]. Interestingly, conservation of the HPV *E7* gene, through a lack of APOBEC-related editing, was shown to be essential for the development of cervical cancer in a population study [12]. Despite these findings, APOBEC activity in HPV infections in young women remains largely uncharacterized.

In this study, we aim to identify intra-sample MNVs in HPV16 infections from young women and monitor changes over time. To this end, we use TaME-seq for sequencing [28]. TaME-seq adapts tagmentation-assisted (enzymatic cleaving and tagging of double-stranded DNA) library preparation by replacing one of the generic sequencing primers with a cocktail of 52 HPV specific primers. Reactions are performed separately for forward and reverse sequencing products, replacing the forward generic primer with a HPV specific one and vice versa. This multiplex PCR enrichment approach results in a higher yield of HPV specific sequencing data. Here, we apply TaME-seq, to a longitudinal retrospective cohort study [29].

2. Materials and methods

2.1. Sample selection

Vaginal self-swabs were obtained from the *Chlamydia trachomatis* Screening and Implementation (CSI) study. Recruitment criteria, methods and additional consent for HPV testing have been described previously [29–31]. Cytology was not performed on these samples, but considering the age of study participants (16–29 years old), the identified infections are likely benign. Participants supplied up to four samples over time. For this study, the median interval between sampling moments was 48 weeks (95% CI: 46–51 weeks; min: 17, max: 63 weeks). Total DNA from 200 μL of sample was isolated using the MagnaPure96 platform (Total Nucleic Acid Isolation Kit, Roche Diagnostics) according to the manufacturer's protocol. Isolated material was eluted in 100 μL and subsequently genotyped via the SPF10-DEIA-LiPA25 platform (DDL Diagnostics) [32,33]. Viral load of HPV16 positive samples was quantified via type-specific qPCR [34]. Infections were selected if they were HPV16 positive during at least three subsequent follow-up moments, preferably with no other HPV genotypes present (Fig. 2).

2.2. Library preparation and sequencing

Library preparation was performed using TaME-seq [28]. Briefly, each sample was tagmented using the Nextera DNA library prep kit (Illumina, Inc., San Diego, CA) and subsequently amplified in two separate reactions. Amplification occurred by multiplex PCR using pools of 27 forward (F) and 25 reverse (R) HPV16 primers in combination with i7 and i5 index primers [35] from the Nextera index kit (Illumina, Inc., San Diego, CA). Libraries from all samples were sequenced on the Illumina MiSeq and HiSeq2500 platform as 151 bp paired-end reads with two 8 bp index reads.

2.3. Sequence alignment and nucleotide variant calling

Sequence data was analyzed using an in-house bioinformatics pipeline [28]. Reads were mapped to the human genome (GRCh38/hg38) and HPV16 reference genome (GI:333,031 HPV16REF.1) [36], using HISAT2 (v2.1.0) [37]. Consensus sequences were extracted using samtools (v1.8) mpileup (-E -d 200,000 -L 200,000), bcftools (v1.6) (call -c –ploidy 1) and vcfutils.pl. Consensus sequences were compared to Sanger data from a previous study [10] using MUSCLE (v.3.8.1551) to align sequences, IQtree (v1.5.5) to infer maximum likelihood phylogeny and FigTree (v1.4.3) to visualize the alignment. Mapped nucleotide counts over HPV reference genomes and average mapping quality values of each nucleotide were retrieved from BAM files. Variant calling was performed using an in-house R (v3.4.4) script (Fig. 1). In each sample, nucleotides called ≤2 times in each genomic position or with mean Phred score of < 30 were removed. From either reaction, results with coverage < 100× were filtered out. F and R nucleotide counts were pooled per sample and major and minor variant frequencies were calculated per position. Samples were excluded if < 45% of the genome was covered ≥100 × . Variants were called if variant frequency was > 1%. If F and R reactions from the same sample showed discordant variants, the reaction with higher coverage was chosen for total variant calling. Genomic locations of MNVs were mapped and major to minor variant mutations were classified as synonymous or non-synonymous in each infection. In addition, MNVs appearing consecutively in follow-up samples from the same infection were identified. Selected samples with a high read count (> 1,000,000) mapped to HPV16, were downsampled randomly to 100.000 reads to rule out possible effects of excessively high sequencing coverage on variant calling.

2.4. Mutational signature analysis

All observed nucleotide substitutions were classified into the six base substitutions, C > A (G > T), C > G (G > C), C > T (G > A),

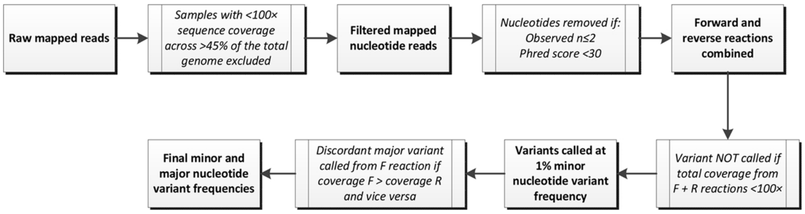
Raw mapped reads
Samples with <100× sequence coverage across >45% of the total genome excluded
Filtered mapped nucleotide reads
Nucleotides removed if: Observed n≤2 Phred score <30
Forward and reverse reactions combined
Variant NOT called if total coverage from F + R reactions <100×
Variants called at 1% minor nucleotide variant frequency
Discordant major variant called from F reaction if coverage F > coverage R and vice versa
Final minor and major nucleotide variant frequencies

Fig. 1. Schematic representation of the nucleotide variant calling.

T > A (A > T), T > C (A > G), and T > G (A > C) substitutions, and then into 96 trinucleotide substitution types that include information on the bases immediately 5′ and 3′ of the mutated base. Analysis was performed using an in-house R (v3.4.4) script. A region frequently subject to insertions / deletions (indels) was identified in the non-coding region (NCR) at positions 4184 and 4185. At these positions, small indels in the sequenced genomes often resulted in mapping errors. Consequentially, 18 T > A and 2 T > G mutations in these two positions have been removed from the present analysis.

2.5. Data availability

The data obtained in this study was deposited in ENA under project number (will be added when available).

3. Results

3.1. Mean sequencing coverage and viral load

In total, 59 samples from persistent HPV16 infections were processed using TaME-seq and 61% (36/59) had $> 45\%$ genome covered by minimum 100× (Table S1), which was the criterion for further analysis. The remaining 36 samples originated from 15 infections (Fig. 2). The mean sequencing coverage per sample ranged from 653 to 399,653 reads (Table S1). Samples had varying HPV16 viral load, which correlated strongly with the per sample mean sequencing coverage (Fig. 3, Pearson correlation coefficient 0.89).

Of the samples with a high viral load (> 1500 copies/μL; n = 30), 29 could be included in downstream analyses. Of the samples with a lower viral load (< 1500 copies/μL; n = 29), only seven could be included, bringing the total sample number included in downstream analyses to 36.

3.2. Comparison of NGS data to previous Sanger results

Sanger data from a previous study was available for 29 out of 36 samples [10]. Consensus sequences obtained in the present study were compared to those previously described [10]. The alignment of Sanger and NGS results overlaps, suggesting high concordance between datasets (Fig. S1).

3.3. HPV16 minor nucleotide variations

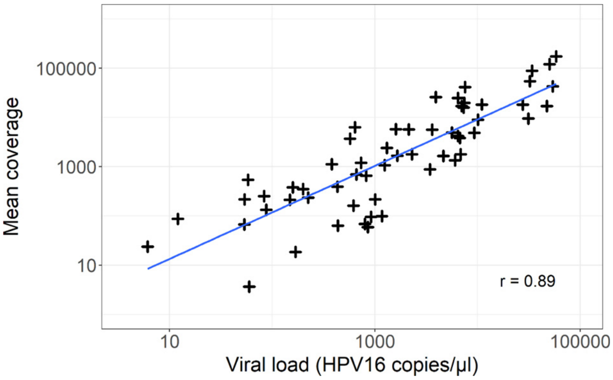

Fig. 3. Correlation between mean sequencing coverage and viral load (HPV16 copies/μL) in each sample.

A total of 1717 HPV16 MNVs (variant frequency $> 1\%$ and coverage $\geq 100\times$) were detected in the 36 samples (Table 1; Fig. 4; Table S2), with 15 to 82 different variants per sample (average 48.3 variants per genome, Table S2). Variant frequency ranged from 1% to 49.6%. No significant correlation was found between the number of variable sites detected and mean or median sequencing coverage (Pearson correlation coefficient: r = -0.41). We note however that the sample (545351-3) with the by far highest mean coverage (399,653) and viral load reports the lowest number of variable sites (n = 15) (Table S1 and S2). The two samples (340223-1 and 407612-1) with the lowest mean coverage, report the mean (48) or below (32) number of variable sites (Table S1 and S2). Of all variants, 85.3% (1465/1717) had a frequency of $< 5\%$. Non-synonymous and synonymous MNVs were analyzed and are summarized in Table 2.

In order to explore unusual mutational patterns in any gene region, the number of synonymous and non-synonymous MNVs was mapped against the consensus sequence of each infection (Table 2). On average there were 167 times (STDEV ± 019) more non-synonymous than synonymous mutations. No genomic region could be singled out as notably different from other regions.

The total number of MNVs observed in each gene region varied considerably (Table 2), but correlated well with gene length (Pearson correlation coefficient: 0.98; Fig. 5). The *L2* gene showed a lower than expected amount of variation, although sequence coverage was low around genome positions 4800–5000 bp. Overall, the majority of MNVs found (90%, n = 1550/1717), were caused by transition events (Table 1). Transversion mutations were detected in 10% of cases. The most common MNV was T > C (A > G; 67%, n = 1146/1717) followed by C > T (G > A; 24%, n = 404/1717) (Table 1; Fig. 6). The overall T > C mutation ratio was 67%. In comparison, the T > C

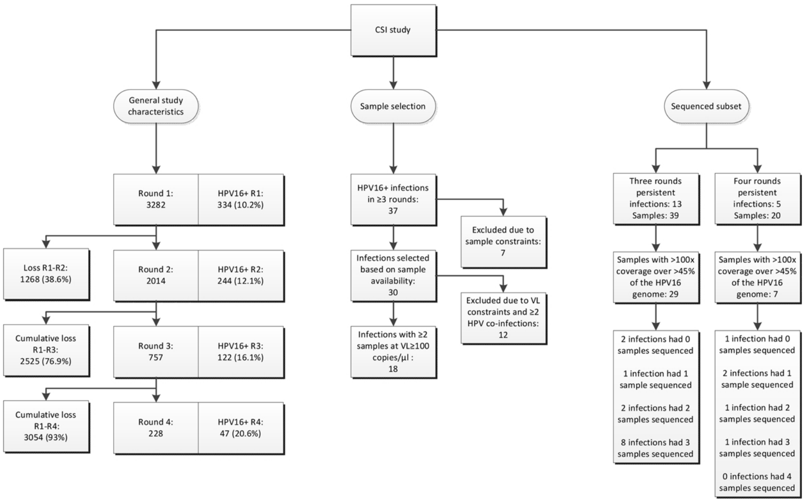
CSI study
General study characteristics
Sample selection
Sequenced subset
Round 1: 3282
HPV16+ R1: 334 (10.2%)
Loss R1-R2: 1268 (38.6%)
Round 2: 2014
HPV16+ R2: 244 (12.1%)
Cumulative loss R1-R3: 2525 (76.9%)
Round 3: 757
HPV16+ R3: 122 (16.1%)
Cumulative loss R1-R4: 3054 (93%)
Round 4: 228
HPV16+ R4: 47 (20.6%)
HPV16+ infections in ≥3 rounds: 37
Excluded due to sample constraints: 7
Infections selected based on sample availability: 30
Excluded due to VL constraints and ≥2 HPV co-infections: 12
Infections with ≥2 samples at VL≥100 copies/µl : 18
Three rounds persistent infections: 13 Samples: 39
Four rounds persistent infections: 5 Samples: 20
Samples with >100x coverage over >45% of the HPV16 genome: 29
Samples with >100x coverage over >45% of the HPV16 genome: 7
2 infections had 0 samples sequenced
1 infection had 1 sample sequenced
2 infections had 2 samples sequenced
8 infections had 3 samples sequenced
1 infection had 0 samples sequenced
2 infections had 1 sample sequenced
1 infection had 2 samples sequenced
1 infection had 3 samples sequenced
0 infections had 4 samples sequenced

Fig. 2. Study flowchart describing selected samples and sequencing outcome. VL = viral load.

Table 1

Composition of minor nucleotide variations (MNVs) identified in this study. Percentages of total (%) MNVs were identified from a single reaction of either the forward (F) or reverse (R) sequencing reaction (only F or R, coverage > 100x) or from both sequencing reactions (F and R, coverage > 100×). MNVs identified in regions where F and R overlapped were compared to major nucleotides and scored if they matched (same major, same MNVs from both F and R) or mismatched (same major, different MNV from both F and R). Finally the number of MNVs detected repeatedly in follow-up samples is shown.

MNV calling/mutation type	T > C	T > A	T > G	C > T	C > A	C > G	Total
Total MNVs	1146 (67%)	56	68	404 (23%)	26	17	1717
MNVs with coverage > 100× for either F or R	610 (68%)	36	38	190 (21%)	13	8	895
MNVs with coverage > 100× for both F and R	536 (65%)	20	30	214 (26%)	13	9	822
Same major and different MNV F and R	398 (63%)	16	22	175 (28%)	8	6	625
Same major and same MNV F and R	135 (73%)	2	6	35 (19%)	4	2	184
Consecutive detection of same MNV	31 (44%)	3	6	24 (34%)	4	2	70

ratios identified from either F or R sequencing reactions or both sequencing reactions together were 68% and 65% respectively (Table 1). When MNVs were detected in regions where F and R reactions overlapped, the T > C ratio was 63% when either the F/R reactions identified a MNV over the set threshold (same major different minor) and 73% when both F/R reactions made the same MNV call (same major and same MNV) (Table 1).

Consecutive samples collected at one-year intervals from the same infection generally showed different MNVs over time. However, 35 MNVs across the HPV16 genome were recaptured in one of the follow-up samples of eleven different infections (Table S3) amounting to 4% (70/1717) of the total MNVs. Furthermore, the T > C mutation ratio drops to 44% in this subset relative to the overall ratio. The T > C MNVs were the most prevalent in all but one sample collected at the third sampling point (444086-3), where the C > T minor variants were dominant (Fig. S2, S3).Moreover, 45 MNVs were detected at 21 polymorphic sites previously associated with CIN3+ (Table S4) [12]. Of these, the two polymorphisms most frequently found were seven in position 3410 in the E2/E4 gene and six in position 4042 in the E5 gene.

4. Discussion

Using the highly sensitive TaME-seq assay, we investigated consecutive HPV16 positive samples from the same infection. Our data suggests the presence of numerous HPV16 MNVs. Consensus sequences (major nucleotide variants) were conserved over time (up to two years follow-up), in line with previous results from this cohort [10]. The detection of MNVs correlated with depth of coverage, which in turn and as to be expected, correlated strongly with sample viral load. The distribution of synonymous and non-synonymous MNVs across the genome appeared uniform and therefore gave no grounds for interpreting selection. Furthermore, MNVs are generally greatly outnumbered by the consensus type, which would be available for transcription of functional proteins. At this MNV level we cannot therefore interpret any substitution rates. Further studies using samples with lesions of varying degrees are required to study the dynamics of and associations between specific MNVs and carcinogenesis.

From 15 HPV16 positive infections (36 samples), we identified a total of 1717 polymorphic positions. Per infection we found on average 48 MNVs/genome (range 15–82) using the 1% frequency cutoff. Our study coincides by magnitude with findings reported by de Oliveira et al. (5–125 MNVs/genome, 1% cutoff) [15], as well as a study investigating HPV16/52/58 MNVs in CIN1+ by Hirose et al. (0–85 MNVs/genome, 0.5% cutoff) [38,39]. Hirose and colleagues further found that the number of HPV16 variants negatively correlated with histological grade. On average, we observe more variants than Hirose et al., which may be in part due to methodological differences, but likely also due to the age group from which our samples were obtained.

Although the number of MNVs identified is comparable to other studies, the nature of the mutation profiles differs. We find that the overall majority (67%) of MNVs were T > C changes, whereas other studies point to a higher frequency of C > T mutations, which we find the second-most abundant (23%). The ratios of T > C mutations against all MNVs are very consistent in our data irrespective of how they were called. Although TaMe-seq is designed with high primer

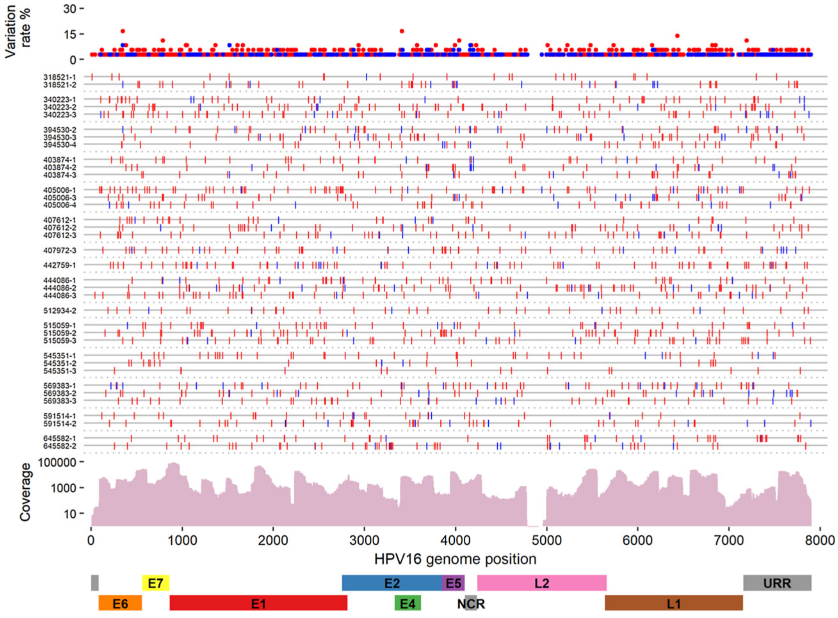

Fig. 4. Variable sites (n = 1717) and mean sequencing coverage in the 36 samples from 15 individuals. Variation rate (top) shows the amount of samples (in %) carrying a minor nucleotide variant in each position. Each horizontal line represents an individual sample, which is named according to case number and sample number (1–4) indicating the sample collection time point. Samples from the same infection are clustered and separated from others by dashed lines. Variable positions with variant frequency of ≤5% are marked with red and variable positions with variant frequency > 5% is marked with blue. Mean sequencing coverage is shown across the HPV16 genome. The location of early (*E1*, *E2*, *E4-7*), late (*L1*, *L2*) genes, URR and NCR is indicated below the HPV16 genomic positions (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

Table 2
Minor nucleotide variants per HPV16 gene/genome region in the 36 HPV16 samples included in the analysis. Where applicable, MNVs are sorted by effect on coding sequence relative to the major nucleotide variant of each infection. *Since certain genes overlap, 84 MNVs are reported more than once.

Gene	Length (bp)	Total number (n) of minor nucleotide variations			
		All (%)	Synonymous (%)	Non-synonymous (%)	Nonsense (%)
E6	477	116 (24.3)	46 (9.6)	66 (13.8)	4 (0.8)
E7	297	65 (21.9)	26 (8.8)	39 (13.1)	0
E1	1950	439 (22.5)	143 (7.3)	282 (14.5)	14 (0.7)
E2	1098	248 (22.6)	90 (8.2)	155 (14.1)	3 (0.3)
E4	288	65 (22.6)	25 (8.7)	40 (13.9)	0
E5	252	68 (27.0)	23 (9.1)	44 (17.5)	1 (0.4)
L2	1422	233 (16.4)	91 (6.4)	142 (10.0)	0
L1	1518	350 (23.1)	133 (8.8)	210 (13.8)	7 (0.5)
URR	832	180 (21.6)	–	–	–
SUM		1801*	577	978	29.

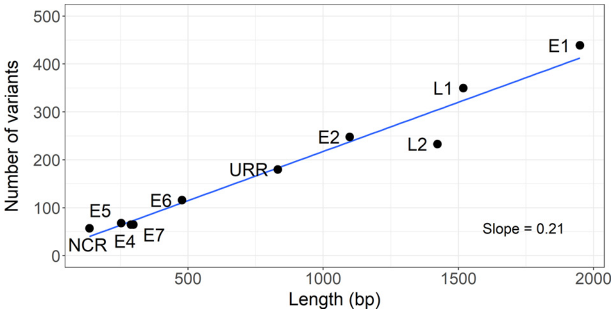

Fig. 5. Correlation between the total number of minor nucleotide variants (MNVs) and the length of viral gene regions including URR and NCR. Since certain genes overlap, MNVs can be counted more than once.

density to cover the entire HPV16 genome (52 in total), it is not designed to completely cover the genome with both the two F and R reactions separately. Despite this, nearly half (n = 809/1717) of the called MNVs are found in overlapping regions obtained from the two reactions independently. Most of these (n = 625) are called in either one of the F or R reactions suggesting that they are either below the 1% frequency cutoff in the other reaction, stochastically amplified from a variant pool by only one of the reactions, or noise. The T > C mutation ratio is lowest in these unpaired MNVs (63%) and highest (73%) in those that are called by both the F and R reactions (11% of total MNVs, n = 184/1717). This is the opposite of what could be expected if T > C mutations were erroneously called, assuming that MNVs independently detected by the F and R reactions confirm each other. The probability of falsely calling the same MNV in two separate reactions is extremely small. Therefore, the derived mutation ratios support the overall finding that T > C mutations dominate the MNVs in our samples. The origin of these T > C mutations remains to be explored, particularly with a focus on early infection events and influence from genome dynamics, DNA repair and viral replication.

The C > T mutation profile is associated with APOBEC activity [12,38]. Over time, APOBEC-related C > T changes accumulate in progressing infections, resulting in mutation patterns observed in CIN1+ materials [12,38]. Our findings imply that APOBEC activity could manifest at later developmental stages of infections than those included in this study. Interestingly, in our dataset we find one infection that shifts over time from a T > C heavy mutation profile, to a C > T heavy mutation profile (444086, Fig. S2), suggesting APOBEC activity. This is further supported by the observation that the C > T mutations in the last collected sample of this infection are almost exclusively in the 5´-TC dinucleotide context which is the preferred APOBEC3A and APOBEC3B motif [40]. MNVs were generally not recaptured in consecutive samples. This may be due to sampling of random fractions of the low frequency MNV for each sample and potentially changes in HPV genome dynamics over time. Despite this, 35 MNVs could be detected repeatedly in follow-up samples. Although the numbers are small, it is noticeable that the T > C ratio is lower (44%) and the C > T ratio higher (34%) in these persistent MNVs relative to the overall distribution of mutations (67 and 28%, respectively). Although this dataset is too small to make firm statements, it is tempting to speculate that an APOBEC footprint accumulates, and therefore becomes more easily detectable in the viral pool over time. This does not necessarily occur from selection but from persisting APOBEC activity. In this study, we repeatedly identified (1–7 times) 45 MNVs at 21 polymorphic sites. These sites overlap with a subset of HPV16 SNPs reported by Mirabello et al., which are significantly associated with disease outcome at the population level [12]. Here, they are identified at the minority level. Although, the biological relevance of low frequency variants is yet to be determined, changes in MNV frequency over time might be an indication of microevolution linking to disease progression. This study presents a first look at the development of MNVs over time. Since previous knowledge on this subject is scarce, a number of unknowns become apparent. Currently, we do not fully comprehend the origin or interplay of minority variants. Variants with similar fitness could be originating naturally over time within hosts, who could then transmit them during intercourse. The role of repeated exposure is also unknown and could

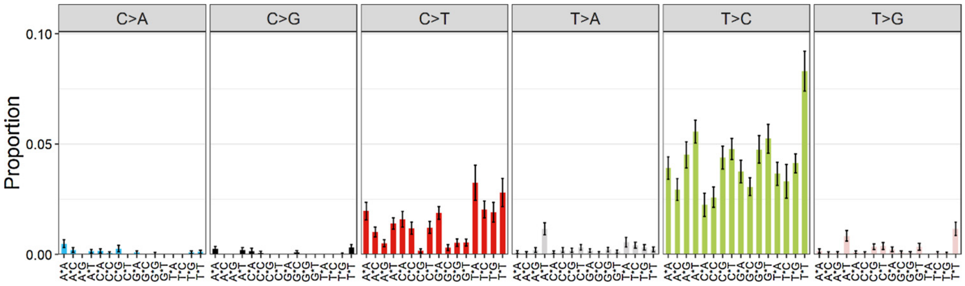

Fig. 6. Overall mutational signatures of 1717 minor nucleotide variants (MNVs) in 36 samples. Mutations are classified into six base substitutions and further into 96 trinucleotide substitution types. Mean proportion of each 96 mutational signature was calculated in the samples. Error bars represent the standard error of the mean.

lead to an increase of variant diversity for each exposure event. Importantly, the detection of abundant intra host MNVs does not challenge the well-established slow evolution of HPVs but rather increases our understanding of the variable HPV16 genome substrate that can be available for natural selection and evolution at the population level. Future research is required to unravel the fundamentals of HPV variant genesis and their role in transmission and establishment of new infections.

One of the strengths of this study is the use of TaME-seq for deep whole HPV genome sequencing. A comparison of consensus sequences obtained using TaME-seq with previously described Sanger sequencing data showed similar results [10]. In addition, the robustness and reliability of the bioinformatics pipeline, calling mutation profiles from raw sequence data, was controlled by reanalysis of the data from Hirose et al. [38], producing excellent compatibility. Finally, our method enabled us to detect 11% of the called MNVs independently in overlapping reads obtained from the two amplification reactions (F and R). Using these, we compared mutational profiles to the whole dataset and similar distributions of mutations were observed.

The design and method used in this study carry some limitations. TaME-seq genome coverage varied between samples and strongly correlated with the initial HPV load. Since overlapping high-resolution data is required to compare MNVs at different time points of an infection, sample inclusion was limited to ≥100× coverage across > 45% of the genome. Consequentially, the mutational patterns observed in this study are often observed on stretches of DNA rather than whole-genome results. It is worth noting that the mutational profiles described in the present analysis, reflect the complete population of HPV16 variants in each sample. No distinction is made between potential co-infections of the same type to prevent potential bias. To compensate for varying viral load of the input material on the resulting sequencing coverage, a downsampling analysis was performed of high-coverage samples, which showed similar results to the original analysis. Therefore, we expect sequence coverage differences to be of limited influence on the observed mutation patterns. However, one 200 nt genomic region was poorly covered in all samples (position 4800–5000), possibly due to scarcity of primers. Potential MNVs in this region may therefore be underreported. One sample with the highest coverage (> 10 fold higher than most other samples) and viral load, reported the least number of MNVs (n = 15). This illustrates how MNVs may not reach 1% frequency against a massive backdrop of major variants in a competitive amplification step.

MNVs were generally found to differ between consecutive samples. The identification of a number of MNVs which were conserved in consecutive samples (Table S3) suggests that this is at least partially caused by sequence coverage and depth. Uncommon MNVs around the detection cutoff will vary in detection and frequency due to PCR and sequencing stochasticity. In addition, the sequencing resolution dictates the number of variants detected from an expected larger mutational pool. It is likely that highly prevalent MNVs are more frequently detected than MNVs around the detection cutoff, although a correlation between MNV prevalence and consecutive detection could not be confirmed for our dataset. It is likely that each sample preparation step leads to a selection of MNVs from the total pool, making redetection of MNVs over time difficult. Furthermore, biological differences between baseline and follow-up samples account for a large portion of MNVs that could not be repeatedly detected. A high viral load at baseline suggests that many MNVs can be detected, while a low viral load at follow-up suggests that only a limited number could be detected. This could explain how often even prevalent MNVs could not be detected in follow-up samples. To our knowledge, this dataset is among the first to describe MNVs in follow-up samples, implying that there could be methodological inefficiencies in the redetection of MNVs from follow-up samples. Further research is required to determine the optimal approach for this.

In this study, QIAGEN Multiplex PCR kit with HotStar Taq DNA Polymerase was used, which, like other polymerases lacking proofreading, could introduce a T > C prone error bias [41]. Additionally, some of the observed transitions could be caused by the Illumina platform [47]. However, as described in the methods section, the use of paired-end reads and a cutoff for calling minor variants (> 1%) should minimize bias from these sources. Furthermore, MNVs in 35 individual genome positions were detected repeatedly in consecutive samples and 138 MNVs in both the separate F and R amplification reactions, suggesting robustness for our observations.

The samples used here were obtained from a retrospective cohort study, which was initially aimed at identifying *Chlamydia trachomatis* infections, and later adapted for HPV purposes [17]. Due to the age of the women recruited for this study (16–29 years old), and the fact that they were recruited for *C. trachomatis* purposes, it is unlikely that the study participants have high-grade cytological malignancies, although this could not be confirmed. The longitudinal nature of this study combined with our inclusion criteria, also means that sample size is relatively small. Since this study was originally conducted to assess *C. trachomatis* status, an effect of such infections might be apparent in the mutation rates of the samples tested in the present analysis. However, since only one of the fifteen infections analyzed here was *C. trachomatis* positive, we could not compare mutation rates between *C. trachomatis* positive and negative individuals.

In summary, this study reports a multitude of MNVs observed through whole genome, deep sequencing of HPV16 infection with longitudinal follow-up. The mutation profiles identified in this study suggested non-APOBEC-related pathways causing mutations in HPV16 infections in young women. Most MNVs were detected incidentally, however, some MNVs could be detected separately or repeatedly over time, suggesting robustness in mutational profiles and at least partial conservation of MNVs. Some of the MNVs identified repeatedly were associated with malignant infection outcomes in other studies, potentially suggesting clinical relevance in longitudinal tracking of MNVs.

Funding

Ministry of Health, Welfare and Sports, the Netherlands and the South-Eastern Norway Regional Health Authority (Grant ID 2016020) are greatly acknowledged for funding.

Ethical approval

This study was approved by the Medical Ethical Committee of the Vrije Universiteit University Medical Centre (VUmc) Amsterdam (2007/239).

CRediT authorship contribution statement

Sonja Lagström: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Pascal van der Weele:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Trine Ballestad Rounge:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Project administration, Supervision, Resources, Software, Validation, Writing - review & editing. **Irene Kraus Christiansen:** Conceptualization, Funding acquisition, Supervision, Methodology, Project administration, Resources, Writing - review & editing. **Audrey J. King:** Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Resources, Writing - review & editing. **Ole Herman Ambur:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Project administration, Supervision, Validation, Writing - review & editing.

Declaration of Competing Interest

None declared

Acknowledgement

The authors would like to thank the CSI group, study investigators and laboratory personnel for their contributions. CSI group: I.V.F. van den Broek (National Institute for Public Health and the Environment, Bilthoven, The Netherlands), E.E.H.G. Brouwers (South Limburg Public Health Service), J.S.A. Fennema (Amsterdam Public Health service), H.M. Götz (Municipal Public Health Service Rotterdam-Rijmond), C.J.P.A. Hoebe (South Limburg Public Health Service), R.H. Koekenbier (Amsterdam Public Health Service), E.L.M. Op de Coul (National Institute for Public Health and the Environment, Bilthoven, The Netherlands), L.L. Pars (STI AIDS Netherlands), S.M. van Ravesteijn (Municipal Public Health Service Rotterdam-Rijnmond). Medical Microbiological Laboratories: A.A.T.P. Brink (Maastricht University Medical Center, Maastricht), A. Luijendijk (Erasmus Medical Center, Rotterdam), A.G.C.L. Speksnijder (Public Health Laboratory, Amsterdam), P.F.G. Wolffs (Maastricht University Medical Center, Maastricht). Finally, we would like to thank three anonymous reviewers and Ignacio G. Bravo for insightful peer reviews.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jcv.2019.08.003.

References

[1] J.G. Baseman, L.A. Koutsky, The epidemiology of human papillomavirus infections, J. Clin. Virol. 32 (Suppl 1) (2005) S16–24.
[2] J.M. Walboomers, et al., Human papillomavirus is a necessary cause of invasive cervical cancer worldwide, J. Pathol. 189 (1) (1999) 12–19.
[3] A.F. Rositch, et al., Patterns of persistent genital human papillomavirus infection among women worldwide: a literature review and meta-analysis, Int. J. Cancer 133 (6) (2013) 1271–1285.
[4] K. Van Doorslaer, Evolution of the papillomaviridae, Virology 445 (1-2) (2013) 11–20.
[5] I.G. Bravo, M. Felez-Sanchez, Papillomaviruses: viral evolution, cancer and evolutionary medicine, Evol. Med. Public Health 2015 (1) (2015) 32–51.
[6] R.D. Burk, A. Harari, Z. Chen, Human papillomavirus genome variants, Virology 445 (1-2) (2013) 232–243.
[7] A.A. Chen, et al., Human papillomavirus 18 genetic variation and cervical Cancer risk worldwide, J. Virol. 89 (20) (2015) 10680–10687.
[8] L. Mirabello, et al., HPV16 sublineage associations with histology-specific Cancer risk using HPV whole-genome sequences in 3200 women, J. Natl. Cancer Inst. 108 (9) (2016).
[9] M. Cullen, et al., Deep Sequencing of HPV16 Genomes: A New High-Throughput Tool for Exploring the Carcinogenicity and Natural History of HPV16 Infection, Papillomavirus Research, 2015.
[10] P. van der Weele, C. Meijer, A.J. King, Whole-genome sequencing and variant analysis of human papillomavirus 16 infections, J. Virol. 91 (19) (2017).
[11] P. van der Weele, C. Meijer, A.J. King, High whole-genome sequence diversity of human papillomavirus type 18 isolates, Viruses 10 (2) (2018) 68.
[12] L. Mirabello, et al., HPV16 E7 genetic conservation is critical to carcinogenesis, Cell 170 (6) (2017) 1164–1174 e6.
[13] R.S. Dube Mandishora, et al., Genotypic diversity of anogenital human papillomavirus in women attending cervical cancer screening in Harare, Zimbabwe, J. Med. Virol. 89 (9) (2017) 1671–1677.
[14] R.S. Dube Mandishora, et al., Intra-host sequence variability in human papillomavirus, Papillomavirus Res. 5 (2018) 180–191.
[15] C.M. de Oliveira, et al., High-level of viral genomic diversity in cervical cancers: a Brazilian study on human papillomavirus type 16, Infect. Genet. Evol. 34 (2015) 44–51.
[16] J. Doorbar, et al., The biology and life-cycle of human papillomaviruses, Vaccine 30 (Suppl 5) (2012) F55–70.
[17] S.D. Kang, et al., Effect of productive human papillomavirus 16 infection on global gene expression in cervical epithelium, J. Virol. 92 (20) (2018).
[18] J.R. Chapman, M.R. Taylor, S.J. Boulton, Playing the end game: DNA double-strand break repair pathway choice, Mol. Cell 47 (4) (2012) 497–510.
[19] C. Moody, Mechanisms by which HPV induces a replication competent environment in differentiating keratinocytes, Viruses 9 (9) (2017).
[20] R. Sanjuan, P. Domingo-Calap, Mechanisms of viral mutation, Cell. Mol. Life Sci. 73 (23) (2016) 4433–4448.
[21] A. Koito, T. Ikeda, Intrinsic immunity against retrotransposons by APOBEC cytidine deaminases, Front. Microbiol. 4 (2013) 28.
[22] L.B. Alexandrov, et al., Signatures of mutational processes in human cancer, Nature 500 (7463) (2013) 415–421.
[23] N.A. Wallace, K. Munger, The curious case of APOBEC3 activation by cancer-associated human papillomaviruses, PLoS Pathog. 14 (1) (2018) p. e1006717.
[24] C.J. Warren, et al., APOBEC3A functions as a restriction factor of human papillomavirus, J. Virol. 89 (1) (2015) 688–702.
[25] S. Mori, et al., Human papillomavirus 16 E6 upregulates APOBEC3B via the TEAD transcription factor, J. Virol. 91 (6) (2017).
[26] M.B. Burns, et al., APOBEC3B is an enzymatic source of mutation in breast cancer, Nature 494 (7437) (2013) 366–370.
[27] M.B. Burns, N.A. Temiz, R.S. Harris, Evidence for APOBEC3B mutagenesis in multiple human cancers, Nat. Genet. 45 (9) (2013) 977–983.
[28] S. Lagström, et al., TaME-seq: an efficient sequencing approach for characterisation of HPV genomic variability and chromosomal integration, Sci. Rep. 9 (1) (2019).
[29] M. Mollers, et al., Prevalence, incidence and persistence of genital HPV infections in a large cohort of sexually active young women in the Netherlands, Vaccine 31 (2) (2013) 394–401.
[30] I.V. van den Broek, et al., Systematic selection of screening participants by risk score in a Chlamydia screening programme is feasible and effective, Sex. Transm. Infect. 88 (3) (2012) 205–211.
[31] I.V. van den Broek, et al., Evaluation design of a systematic, selective, internet-based, Chlamydia screening implementation in the Netherlands, 2008-2010: implications of first results for the analysis, BMC Infect. Dis. 10 (2010) 89.
[32] B. Kleter, et al., Development and clinical evaluation of a highly sensitive PCR-reverse hybridization line probe assay for detection and identification of anogenital human papillomavirus, J. Clin. Microbiol. 37 (8) (1999) 2508–2517.
[33] B. Kleter, et al., Novel short-fragment PCR assay for highly sensitive broad-spectrum detection of anogenital human papillomaviruses, Am. J. Pathol. 153 (6) (1998) 1731–1739.
[34] P. van der Weele, et al., Correlation between viral load, multiplicity of infection, and persistence of HPV16 and HPV18 infection in a Dutch cohort of young women, J. Clin. Virol. 83 (2016) 6–11.
[35] J.J. Kozich, et al., Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform, Appl. Environ. Microbiol. 79 (17) (2013) 5112–5120.
[36] K. Van Doorslaer, et al., The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis, Nucleic Acids Res. 41 (Database issue) (2013) D571–8.
[37] D. Kim, B. Langmead, S.L. Salzberg, HISAT: a fast spliced aligner with low memory requirements, Nat. Methods 12 (4) (2015) 357–360.
[38] Y. Hirose, et al., Within-host variations of human papillomavirus reveal APOBEC signature mutagenesis in the viral genome, J. Virol. 92 (12) (2018).
[39] I. Kukimoto, et al., Genetic variation of human papillomavirus type 16 in individual clinical specimens revealed by deep sequencing, PLoS One 8 (11) (2013) p. e80583.
[40] C.J. Warren, et al., Roles of APOBEC3A and APOBEC3B in human papillomavirus infection and disease progression, Viruses 9 (8) (2017).
[41] C. Brandariz-Fontes, et al., Effect of the enzyme and PCR conditions on the quality of high-throughput DNA sequencing results, Sci. Rep. 5 (2015) 8056.