Changes in intra-host genetic diversity according to lesion severity in longitudinal HPV16 samples.

Jean-Marc Costanzi¹, Milan Stosic S¹, Alexander Hesselberg Løvestad¹, Ole Herman Ambur², Trine Rounge B³, and Irene Kraus Christiansen¹

¹Akershus Universitetssykehus HF
²OsloMet - storbyuniversitetet
³Universitetet i Oslo

March 12, 2024

Abstract

Human papillomavirus type 16 (HPV16) is the most common cause of cervical cancer, but most infections are transient and with lesions not progressing to cancer. There is a lack of specific biomarkers for diagnosis and risk stratification. This study aimed to explore the intra-host HPV16 genomic variation in longitudinal samples from HPV16-infected women with different cervical lesion severity (normal, low-grade, and high-grade). The TaME-seq deep sequencing protocol was used to generate whole genome HPV16 sequences of 102 samples collected over time from 40 individuals. Single nucleotide variants (SNVs) and intra-host single nucleotide variants (iSNVs) were identified in the viral genomes. A majority of individuals had a unique set of SNVs and these SNVs were stable over time. Overall, the number of iSNVs and APOBEC3-induced iSNVs were significantly lower in high-grade relative to normal and low-grade samples, respectively. A significant increase in the number of APOBEC3-induced iSNVs over time was observed for normal samples when compared to high-grade. Our results provide new insight into the dynamics of HPV16 within-hosts evolution. Low number of iSNVs and APOBEC3-induced iSNVs, characteristics of high-grade lesions, could potentially serve as biomarkers to guide triage of HPV-induced cervical precancerous lesions.
unique set of SNVs and these SNVs were stable over time. Overall, the number of iSNVs and APOBEC3-induced iSNVs were significantly lower in high-grade relative to normal and low-grade samples, respectively. A significant increase in the number of APOBEC3-induced iSNVs over time was observed for normal samples when compared to high-grade. Our results provide new insight into the dynamics of HPV16 within-hosts evolution. Low number of iSNVs and APOBEC3-induced iSNVs, characteristics of high-grade lesions, could potentially serve as biomarkers to guide triage of HPV-induced cervical precancerous lesions.

Keywords

Human papillomavirus, HPV16, cervical cancer, APOBEC3, longitudinal study, SNV, iSNV, whole genome sequencing, intra-host variation

Introduction

Human papillomavirus type 16 (HPV16) is the most carcinogenic HPV genotype, responsible for almost 50% of all cervical cancer cases. Cervical cancer prevention relies on vaccination, screening programs and treatment of precancerous lesions. While HPV vaccines protect against the most carcinogenic HPV-types, it is expected to take several decades before cervical cancer is eliminated as a worldwide public health problem. Meanwhile, screening programs are essential for early detection of HPV infections and guide triage of HPV-induced cervical precancerous lesions. Currently, cervical cancer screening programs rely either on cytology or HPV-based screening, with HPV-based screening being clinically more sensitive but with a lower clinical specificity, especially for younger women. To increase the clinical specificity, thereby reducing overtreatment, identification and implementation of biomarkers are needed.

HPV16 belongs to the family *Papillomaviridae*, genus *Alphapapillomavirus* and species *Alphapapillomavirus* 9(https://ictv.global/). Estimates suggest the presence of more than 390 HPV genotypes, with each type having a minimum of 10% nucleotide variation in the L1 gene region. Each genotype encompasses several lineages exhibiting a whole genome variation between 1% and 10%. Each lineage can be further divided into sublineages with a variation between 0.5% and 1% of the whole genome sequence. Previous studies have shown HPV16 sublineages to correlate with differential cervical cancer risk. While HPV is considered a slowly evolving virus with a relatively stable genome, multiple studies have uncovered considerable genomic variability also within the sublineages level. This variability could explain the differences in the carcinogenic potential of different HPV infections. Genomic variability can be investigated as single nucleotide variants (SNVs), relative to a consensus reference for the sublineage, and intra-host single nucleotide variants (iSNVs) within an individual sample. The within-host consensus is defined as the genome with the most abundant sequenced nucleotide in each position of the HPV genome. This provides a unique within-host consensus per individual. In this study, SNVs will be determined by comparing the within-host consensus HPV genome to their respective sublineages reference genomes (Figure 1). Furthermore, our investigation extends to the deep-sequenced viral genomic diversity of a virus population within an individual sample, i.e iSNVs.

It has been hypothesized that iSNVs could serve as markers for the potential progression of an HPV infection to cervical cancer. A lower amount of iSNVs has previously been associated with cervical precancerous lesions/cancers when compared to normal/low-grade lesion cases. While not fully understood, iSNVs have been linked to the activity of the anti-viral innate host-defence enzymes, apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3 (APOBEC3). APOBEC3 is a family of cytidine deaminases catalysing the deamination of cytidine to uracil with altered base-pairing properties to cause deoxycytidine (C) to deoxythymidine (T) mutations in the HPV genome during its synthesis. APOBEC3s preferentially targets TCN trinucleotide context (where N can be any nucleotide). APOBEC3-induced mutations have been associated with viral clearance and are less likely to be found in individuals with pre-cancerous or cancerous lesions. Conversely, the role of APOBEC3s has also been linked to genomic instability and cancer progression.
A persistent HPV infection might result in an increase in APOBEC3 expression which in turn can lead to an accumulation of somatic mutations.

SNVs, iSNVs and APOBEC3-induced mutations have previously been investigated in cross-sectional studies. To our knowledge, this is the first study to examine how iSNVs and APOBEC3-induced iSNVs change over time according to lesion severity in longitudinal samples. Previous longitudinal studies have either solely focused on SNVs or not tested for association between lesion severity and iSNVs/APOBEC3-induced iSNVs.

Unlike cross-sectional studies, which provide a snapshot in time, analyses of longitudinal samples can provide a better understanding of the viral variability over time within an individual. By following the dynamics of mutations over time it is possible to reveal patterns that can help predicting medical outcomes and ultimately guide triage. The objectives of this study are to investigate if iSNVs and/or APOBEC3-induced iSNVs evolve over time according to diagnostic categories defined as normal, low-grade and high-grade. Also, we aim to see if SNVs are stable and traceable within individuals over time.

Methods

Sample selection

Collection of cervical cell samples in ThinPrep PreservCyt solution (Hologic, Marlborough, MA) was carried out between 2005 and 2008 from women participating in the cervical cancer screening program in Norway, including opportunistic screening samples, and stored in a research biobank. The cervical samples were classified into three different diagnostic categories: normal, low-grade and high-grade, based on both cytology of the specific samples and/or on related histological findings. If histology of the pretreatment biopsy or of the cone material was performed within 90 days of the cytology, histological diagnosis was prioritised for sample categorising. Cell samples from women at the time of treatment were also collected for the biobank. For these, only the histological cone diagnosis was available. Samples were categorized primarily by histology or by cytology when no histology was available. Samples were categorised as normal if they were found normal by histology or by cytology. Samples were categorised as low-grade if diagnosed with cervical intraepithelial neoplasia grade 1 (CIN1) by histology, or as low-grade squamous intraepithelial lesion/ atypical squamous cells of undetermined significance (LSIL/ASC-US) by cytology. Samples were categorised as high-grade if diagnosed with CIN2 or higher by histology, or as atypical squamous cells cannot exclude high-grade squamous intraepithelial lesion / high-grade squamous intraepithelial neoplasia (ASC-H/HSIL) by cytology. If an individual was submitted to treatment, all following samples were excluded. The samples were pelleted and stored at -80°C in a research biobank at Akershus University Hospital.

DNA extraction

From the biobank, samples were available as both cell material and as extracted DNA. Some of the extracted DNA samples were exhausted and were replaced with new extracts from cell material. The extraction was done using the eMAG system (Biomérieux, USA). For all samples, DNA concentration was measured using a Quant-iT dsDNA Assay Kit, high-sensitivity (HS) (Life Technologies, USA).

Library preparation

The sequencing library was prepared using the tagmentation-assisted multiplex PCR enrichment sequencing (TaME-seq 2) protocol. A detailed protocol can be found at dx.doi.org/10.17504/protocols.io.dm6gpjxy5gzp/v1. Sequencing was performed on a Novaseq SP platform using 151 bp paired-end reads.

iSNV calling

HISAT2 (v 2.2.1) was used to align the reads to the reference genome. From this alignment, all reads mapping to the HPV reference genome and their quality were extracted using mpileup from BCFtools (v1.12). The
A detailed procedure is described in [1]. The mpileup output was used to call intra-host single nucleotide variants (iSNVs). The calling was done in a reference-free manner where the most frequent base in each position was defined as the consensus variant and the second most frequent as the iSNV. Nucleotide positions with read depth <100x or with Phred scores <30 were filtered out. Additionally, positions with read depths >100x but >500x required the iSNVs to have an observed frequency >5%. The required frequency was decreased to >1% for positions with read depth >500x. Only samples with a mean sequencing depth of >300x were included in the analysis. All iSNVs identified in repetitive homopolymeric regions within the non-coding region and the upstream regulatory region were removed from the analysis.

Potential sublineage coinfections were removed from the analysis in a sublineage reference dependant manner. If a sample had more than 5% of iSNVs matching another sublineage defining site, it was considered a likely sublineage coinfection and hence removed.

For iSNV statistical analysis, the sum of iSNVs per sample was used. Nucleotide diversity (\(\pi\)) was calculated for each site using the number of differences between the major and the minor variant sequencing depth, divided by the total number of comparisons. The results were then averaged for each sample over the entire genome. Finally, APOBEC3-induced iSNVs were filtered from iSNVs and defined as the total number of C to T mutations in a TCW (TCA and TCT) trinucleotide context per sample. This trinucleotide context was chosen as it is the preferred motif of APOBEC3A/B which is associated with HPV infection.

Statistical analysis

Mixed models were created to investigate the relationship between iSNVs, \(\pi\) or APOBEC3 mutation signature (dependent variables) and diagnostic categories, time and individual ID (independent variables). Time was defined as the difference in days between the first sample (set as zero) and all consecutive samples for each individual. Each model was based on the combination of one dependent variable with all independent variables with an interaction between diagnostic category and time as diagnostic category can vary over time. Because the dataset had multiple samples per individual (longitudinal dataset), the individual ID was included as a random intercept. A generalized linear mixed model with a negative binomial distribution (to account for overdispersion) was used for iSNVs, and a linear mixed model was used for \(\pi\) (detailed formulas in Table S1, S2). To analyse APOBEC3-induced iSNVs, a zero-inflated generalized linear mixed model with a Poisson distribution was used (detailed formulas in Table S3). Residuals of all models were checked visually and no overdispersion was detected in the generalized linear mixed models (tested with the R package DHARMa v0.4.6).

SNV calling

To investigate the SNVs, consensus sequences for all samples were generated by calling the nucleotide with the highest depth at each position (min. 20\(\times\)). Only samples with a genomic coverage >80% of the reference genome were used in further analysis. In total 101 samples were selected for further analysis. Sublineage identification involved independent alignment of each sample’s consensus sequence with 16 different HPV16 sublineage references, utilising clustalO (v1.2.3). The sample-reference sequence relationships were determined with maximum likelihood (ML) trees, employing RAxML-NG v1.2.0 with the following specifications: a general time-reversible model, 50 randomised parsimony initiation trees, and 1000 bootstrap replications. A consensus sequence would be categorized as a sublineage if they shared a clade and if the clade bootstrap value exceeded 70. The majority of samples sequences were identifiable as either A1 or A2. Then a constrained phylogenetic tree was made with all four lineage reference sequences and all sample sequences. All sample sequences grouped with reference sequences corresponding to lineage A (Figure S1). Finally, single nucleotide variants (SNVs) were identified whenever there was a discrepancy between a nucleotide in the sample and its corresponding nucleotide in the sublineage reference sequence. Any SNV corresponding to specific differences between A1 and A2 sublineages were omitted from further analysis.
Results

Dataset description

In total, 271 samples from 90 individuals were sequenced in two runs using the TaME-seq 2 protocol for this and other studies. From this dataset, 125 samples (72 individuals) were categorized as normal, 53 (41 individuals) low-grade, 93 (55 individuals) high-grade and one lacked diagnostic categorisation. After study specific filtering, the dataset consisted of 102 samples from 40 different individuals. These filtering steps involved removing 33 technical replicate samples (used in another study), one sample with untraceable diagnostic category, 49 samples retrospectively found collected after conization, and 50 samples with mean sequencing depth <300x. After these filtering steps, 32 individuals with only one sample available were excluded in this longitudinal study. Finally, four samples were excluded as they were found to have potential coinfections with different HPV16 sublineages. The number of samples per diagnostic category included here are presented in Table 1. The number of samples per individual ranged between 2 and 5 (mean: 2.55). The time difference in days between the first and the last sample for an individual ranged between 37 and 814 days (mean: 340 days).

iSNVs:

Overall, the number of iSNV within a sample ranged from 1 to 61 (Table 2 and Figure S2 for result per individual). The mean number of iSNVs per diagnostic category was 20.3, 18.2 and 12.3 for normal, low-grade and high-grade, respectively (Table 2). The number of iSNVs did not significantly change over time within individuals for any of diagnostic group. However, when not accounting for time, we observed a significant difference (p-value: 0.011) in the number of iSNVs between high-grade and normal diagnostic categories, with high-grade samples having a significantly lower number of iSNVs than normal samples on average (Figure 2A). No significant difference in the number of iSNVs was observed between low-grade and high-grade or low-grade and normal (Figure 2A and Table S1).

Nucleotide diversity (π):

The mean nucleotide diversity (π) per diagnostic category was 1.64e-3, 1.66e-3 and 1.82e-3 for high-grade, low-grade and normal respectively (Table 2 and Figure S3 for result per individual). The models revealed that π did not significantly change overtime (Figure 2B, Table S2). Moreover, no significant differences in π were observed between diagnostic categories when not accounting for time (Figure 2B).

APOBEC3-induced iSNVs:

The number of C to T mutations in TCA and TCT context ranged from 0 to 19 within samples (Table 2 and Figure S4 for result per individual). The mean number of those mutations per diagnostic category was 2.59, 2.74 and 1.35 for normal, low-grade and high-grade, respectively (Table 2). When not accounting for time, the number of APOBEC3-induced iSNVs was observed to be significantly higher for low-grade when compared to high-grade (p-value: 0.0015) (Table S3 and Figure 2C). The model showed that time alone had no significant effect on the number of APOBEC3-induced iSNVs (Table S3). However, the interaction between diagnostic categories and time difference was significant (p-value: 0.025) (Table S3). This implies that the number of APOBEC3-induced mutations changed differently over time depending on whether the samples were normal, low-grade or high-grade. Normal samples had a significant increase in APOBEC3-induced iSNVs over time, compared with high-grade samples (Figure 2C). Conversely, the number of APOBEC3-induced iSNVs did not significantly change over time for low-grade and high-grade samples. These diagnostic categories had a stable level of APOBEC3-induced iSNVs throughout the sample collection period.

SNVs variation over time

When investigating how SNVs may vary over time within individuals, we found that most individuals kept the same SNV pattern in all their samples collected at different time points. Only 5 out of 40 individuals (12.5%) were observed to acquire new SNVs or lose SNVs over time (Figure 3). Among those, four individuals acquired new SNVs between two time points and three individuals lost SNVs between time points (Figure
3). The number of newly acquired SNVs per individual between two time points ranged from 1 to 4. Finally, we also observed, over the entire dataset, that every individual except for two had a unique set of SNVs.

Discussion

Taking a longitudinal approach, we looked at genomic variation of HPV16 within individual hosts over time, uncovering lesion-specific characteristics and potential factors that influence viral genome dynamics. Our results reveal distinct genetic variation patterns associated with different stages of cervical lesion severity. This highlights the complex interaction between HPV and its host and the potential implications for cervical cancer progression diagnostics and triage.

The results showing a significantly higher total number of iSNVs in the normal diagnostic category relative to high-grade are in line with our cross-sectional study. observed a trend of HPV16 positive non-progressive samples having a higher number of iSNVs (minor nucleotide variants, MNVs) than CIN2+ samples, which corroborates our findings. The observed higher total number of iSNVs in normal and low-grade diagnostic category samples could be linked to the HPV life cycle with high rates of viral replication in early infection stages.

A wide range of factors including polymerase fidelity and cytidine deaminases (APOBEC3) are known to make viral replication a rich source of mutations. The lowest number of APOBEC3-induced iSNVs was observed in high-grade samples, and this number was stable over time in our model. Those results are in line with previous studies that detected a lower number of APOBEC3-induced iSNVs in higher grade/cancer samples. The lower number of total iSNVs in high-grade lesions might be a result of clonal expansion, a cell population bottleneck where specific infected cells with a growth advantage proliferate faster to gain tissue/sample dominance alongside lesion progression. This selection on growth could be further driven by the adaptive tumour progression model which suggests that certain cells could gain their competitive advantage due to specific host and/or viral mutations that could favour their clonal expansion.

Our model showed that the number of APOBEC3-induced iSNVs increased significantly in the normal category over time (Figure 2C and Table S3). To our knowledge this is the first time this has been observed in HPV infections. This result strengthens the proposed link between viral replication rates, infection persistence and APOBEC3 induced mutations. Finding that APOBEC3-induced iSNVs are accumulating over time in the normal category (Figure 2C) corroborates with previous studies showing that HPV infections trigger APOBEC3 expression as part of the host innate anti-viral immune defence associated with viral clearance. Also, the number of APOBEC3-induced iSNVs was found significantly higher in low-grade samples compared to high-grade in a time-independent manner to further support the link to the replicative stages of viral life cycle (Figure 2C). We therefore hypothesise that the number of APOBEC3-induced iSNVs accumulates during the early and productive stages of infection and plateau in the low-grade stage of infection. This would explain the high starting number of APOBEC3-induced iSNVs in the low-grade category and its stability over time. We also speculate that the low number of APOBEC3-induced iSNVs could reflect a suppressed APOBEC3 activity which would result in a persistent infection and increase the chances of high-grade lesions.

Although a substantial proportion of the iSNVs in the early diagnostic categories could be attributed to APOBEC3, this factor cannot alone fully explain the total number of iSNVs observed in normal samples. This indicates that other factors than APOBEC3s, as reviewed by , might be responsible for the observed total iSNVs.

Interestingly, we observed no significant difference in nucleotide diversity ($\pi$) between diagnostic categories. This is in line with the results from although their nucleotide diversity calculation was based on SNV variation. While nucleotide diversity in this study was calculated from iSNVs, the similar results indicate that this diversity might not generally play an important role in the development of precancerous lesions. Nucleotide diversity was also found to be stable over time, which is in line with our iSNV and SNV results.
It is challenging to track a specific APOBEC3-induced mutation or a specific iSNV within individuals over time due to their low frequency in the intra-host population. Indeed, methods such as tagmentation, PCR amplification and sequencing, which are all part of the TaME-seq2 protocol, are stochastic processes making the repeated detection of specific low frequency variants difficult. The TaME-seq2 protocol can however reliably detect specific SNVs with an extremely low error rate. In this study, SNV mutations were found to be conserved over time for most individuals (87.5% of individuals had all SNVs conserved overtime) as also reported by other studies. Genome conservation through time is a good indication of a persistent infection. While we observed five individuals (12.5%) losing or acquiring mutations between samples, most other SNVs in those samples remained stable over time indicating the same ongoing HPV infection. Further investigations revealed that these SNVs were polymorphic in at least one of the samples for each individual (Table S4), with the reference genome base always present as either an iSNV or SNV. These observed SNV changes may not indicate actual shifts in viral allele frequencies in intra-host viral populations over time but could rather be artifacts of stochasticity in PCR or sequencing reactions. Furthermore, variations in sample material might influence the detection of specific SNVs as they might be highly heterogeneous. To investigate the potential viral genomic variation within a heterogeneous sample, future study might employ targeted cell sampling such as the Cervical cell lift, laser dissection of tissue samples or single cell sequencing.

Our results provide new insights into the dynamics of within hosts HPV16 evolution and its association with cervical lesion severity. They also highlight the significant role of APOBEC3s in promoting intra-host variation in normal and low-grade samples. However, our study also has some limitations. First, we used a relatively small sample size and a short time frame (min:37 days and max: 814 days), which might limit the generalisability and power of our findings. Despite these limitations, we could still identify significant iSNV characteristics of different diagnostic categories and importantly that APOBEC3-induced iSNVs accumulated over time in the normal category. Second, we focused only on HPV16, and other types might have different patterns of genetic diversity and APOBEC3-induced iSNVs. Third, we did not investigate the functional consequences of the iSNVs and APOBEC3-induced iSNVs on viral gene expression and protein function, which could affect viral fitness and pathogenicity. To overcome these limitations, future studies should include larger and more diverse cohorts of HPV-infected women, different HPV genotypes and sublineages, and women with longer follow-up periods.

In conclusion, this longitudinal study on cervical samples from HPV16 positive women revealed that iSNVs and APOBEC3-induced iSNVs vary according to diagnostic categories and over time. These findings suggest that HPV16 genomic variations are influenced by both viral and host factors that may affect the outcome of HPV infection and cervical carcinogenesis. Additionally, we hypothesize that a suppressed APOBEC3 activity, indicated by the low number of APOBEC3-induced iSNVs, could contribute to the persistence of infection and an increased risk of high-grade lesions. Finally, our results could have implications for HPV diagnostics in cervical cancer prevention, as they could help identify candidate biomarkers of infection persistence and progression.

Acknowledgements:

We would like to thank A. Kathrine Lie, dr. med., Department of Pathology, Oslo University Hospital, for valuable contribution defining clinical categories.

The sequencing service was provided by the Norwegian Sequencing Centre (https://www.sequencing.uio.no), a national technology platform hosted by Oslo University Hospital and the University of Oslo supported by the Research Council of Norway and the South-Eastern Regional Health Authority.
Data availability statement:

The data presented in this article are available upon reasonable requests directed to the corresponding authors, with compliance with the General Data Protection Regulation (GDPR), and ethical approval from the Norwegian Regional Committee for Medical and Health Research Ethics (REC).

Funding statement:

This work was supported by a post-doctoral research grant to JMC from the South- Eastern Norway Regional Health Authority, project number 2020010 and a by a PhD grant to AHL and MS from Faculty of Health Sciences, Oslo Metropolitan University. The funders had no role in study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

Conflict of interest disclosure:

The authors have no conflict of interest to declare.

Ethics approval statement:

The study was approved by the Regional Committee for Medical and Health Research Ethics, Oslo, Norway (REK 2017/447) and by Akershus University Hospital’s Data Protection Official (2017-109).

References

Table legends:

**Table 1**: Number of individuals, samples analysed and mean mapping statistics for each diagnostic category.

**Table 2**: Mean, min and max of the number of iSNV, nucleotide diversity $\pi$ and Number of C$>$T mutation in T*A and T*T context (APOBEC3 mutation signature) for each diagnostic category.

Figure legends:

**Figure 1**: Graphical abstract of the study outlining the study design and main results. Summary of the dataset, the different methods used to estimate intra-host variation on the left side and simplified representation and interpretation of the main results on the right side. Methodological outputs are matched by colour to their respective main results.

**Figure 2**: The predicted relationship with 95% confidence interval between A) the number of iSNV per sample, B) nucleotide diversity ($\pi$) per sample or C) the number of C$>$T mutations in T*A and T*T context (representing APOBEC3-induced iSNVs) per sample, and time difference from the first sample within an individual (in days) for each diagnostic category. Points represent the raw observations for each sample. In B) one outlier was removed from the figure for ease of reading. The outlier ($\pi = 0.00654$) was kept in all analyses.
Figure 3: Circos plot representing the variation in SNVs between samples within individuals. Each circos plot represents an individual, light and dark grey circles are samples from the same individual (also labelled S1, S2, S3). The outside circle represents the HPV gene regions. The dark blue dots are the SNVs compared to the reference genome. The position within the grey circle represents the frequency of this particular SNV, a SNV closer to center has a lower frequency. SNV with a frequency close to 1 are positioned on the outer part of the grey circle. The squares highlight the changes, a red square represents an acquired SNV between two samples, a green square represent a lost SNV between two samples and a yellow square represent an SNV that is first acquired then lost.
Hosted file

Table1.xlsx available at https://authorea.com/users/716164/articles/724222-changes-in-intra-host-genetic-diversity-according-to-lesion-severity-in-longitudinal-hpv16-samples

Hosted file

Table2.xlsx available at https://authorea.com/users/716164/articles/724222-changes-in-intra-host-genetic-diversity-according-to-lesion-severity-in-longitudinal-hpv16-samples