Characterisation of human papillomavirus genomic variation and chromosomal integration in cervical samples

by

Sonja Lagström

Thesis submitted to the Faculty of Medicine, University of Oslo
For the Degree of Doctor of Philosophy (PhD)

Department of Microbiology and Infection Control
Akershus University Hospital

Department of Research
Cancer Registry of Norway

Lørenskog, Norway
2020

Challenges are what make life interesting and overcoming them is what makes life meaningful.

Joshua J. Marine

ACKNOWLEDGEMENTS

The work for this doctoral thesis was carried out at the Department of Microbiology and Infection Control, Akershus University Hospital (Ahus) and at the Department of Research, Cancer Registry of Norway (KRG). The work was funded by a grant from the South-Eastern Norway Regional Health Authority. I was part of the PhD programme at the Faculty of Medicine, University of Oslo (UiO) that provided the courses and supported the PhD project.

First, I would like to thank my supervisors Trine B. Rounge (KRG, UiO), Irene Kraus Christiansen (Ahus) and Ole Herman Ambur (Oslo Metropolitan University, OsloMet) for sharing your expertise in HPV and guiding me through this thesis. You were such a great combination of supervisors with different backgrounds and skills but with same curiosity and enthusiasm for research. I am privileged to have had the three of you as my supervisors, thank you! I also want to thank my co-supervisor Truls M. Leegaard (Ahus, UiO), enabling me to be part of the PhD programme at the University of Oslo.

In addition to my supervisors, I would like to thank Mari Nygård (KRG) and Ameli Tropé (KRG) for sharing your knowledge on HPV and cervical cancer and giving me valuable advice along the way. I am also grateful to Pekka Ellonen (Institute for Molecular Medicine Finland, FIMM), my former boss at SeqLab. This is where this whole journey started when I was first introduced to amplicon sequencing, then to Trine Rounge and HPV and finally I was involved in the development of a new HPV sequencing method. I am so happy that I was always welcome to visit SeqLab after I left my job there, to learn new things and to share the latest developments in my project. I am also grateful to have had the opportunity to work together with Alexander Hesselberg Løvestad (OsloMet), our latest addition to the team. It has been a pleasure to share scientific and not-so-scientific discussions with you.

In addition, I would like to thank our collaborators from Finland and the Netherlands. Thank you Maija Lepistö (FIMM) for offering your methodological expertise, and Audrey J. King (National Institute for Public Health and the Environment, RIVM) and Pascal van der Weele (RIVM) for the interesting collaboration opportunity. I am also grateful to have met my first African friend, Racheal Mandishora (International Agency for Research on Cancer, IARC), during the project. You have introduced me to a whole new world of HPV in Africa. It is true what you said when we met for the first time in Cape Town, South Africa: "it's like we have always known each other".

Before starting in the PhD project, I had to pack my things and move to the other side of the Nordic countries. I could not have wished for a better start in Norway at these two great workplaces. I have been privileged to work with amazing and kind people. My special thanks go to Hanne Haugland and Mona Hansen for all the support and help in the HPV lab and to Roger Meisal for helping me the first months of the project. I want to thank also the rest of the FoU group at Ahus for the great atmosphere both at work and outside of it. I had wonderful colleagues at KRG, thanks especially to Sinan Uğur Umu for teaching me bioinformatics and R, and to Elina Vinberg and Marcin Wojewodzic for all your help with different projects, nice floor ball training and other activities outside the work. In addition, I would like to thank Anna Frengen and other people in the EpiGen lab at Ahus for your help and guidance.

I would also like to thank NORBIS national research school in bioinformatics, biostatistics and systems biology and Health Innovation School (UiO, Norwegian University of Science and Technology and Karolinska Institutet) for offering me the possibility to expand my knowledge and think outside the box. During the seminars and courses I have met funny and enthusiastic young talents from so many different fields. We had great opportunities to visit new wonderful places in the Nordics. This also made it possible for me to experience the Northern lights for the first time in my life.

Finally, I would like to thank my family and friends for always being there for me. Special thanks go to my parents and my husband Gustaf: without your encouragement, I would never have moved to Norway. Thank you Gustaf for your endless support, love and cooking skills. It was always great to come home to a ready-made dinner. Lastly, the biggest reason to be effective and to get home as early as possible even during the final phase of the thesis, is Hilda, our wonderful little daughter and the joy of our life.

Lørenskog, April 2020

Sonja Lagström

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ADC	Adenocarcinoma
AIS	Adenocarcinoma *in situ*
APOBEC	Apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like
APOT	Assay of papillomavirus oncogene transcripts
ASC-H	Atypical squamous cells, cannot exclude high-grade lesion
ASC-US	Atypical cells of undetermined significance
BAM	Binary Alignment/Map
CIN	Cervical intraepithelial neoplasia
CNV	Copy number variants
Ct	Cycle threshold
DIPS	Detection of integrated papillomavirus sequences
GDPR	General Data Protection Regulation
HPV	Human papillomavirus
HSIL	High-grade squamous intraepithelial neoplasia
IARC	International Agency for Research on Cancer
Indel	Insertion or deletion
LBC	Liquid-based cytology
LCR	Long control region
LEEP	Loop electrosurgical excision procedure
LSIL	Low-grade squamous intraepithelial neoplasia
MNV	Minor nucleotide variant
NCR	Non-coding region
NGS	Next generation sequencing
ORF	Open reading frame
PaVE	Papillomavirus Episteme
PCR	Polymerase chain reaction
PE	Paired-end
QC	Quality control
qPCR	Quantitative polymerase chain reaction
pRB	Retinoblastoma protein
RNA-seq	RNA sequencing
SAM	Sequence Alignment/Map

SBS	Sequencing-by-synthesis
SCC	Squamous cell carcinoma
SE	Single-end
SNP	Single-nucleotide polymorphism
TaME-seq	Tagmentation-assisted multiplex PCR enrichment sequencing
URR	Upstream regulatory region
VAF	Variant allele frequency
VLP	Virus-like particle
WES	Whole exome sequencing
WGBS	Whole genome bisulfite sequencing
WGS	Whole genome sequencing

LIST OF ORIGINAL PAPERS

I. **Lagström S**, Umu SU, Lepistö M, Ellonen P, Meisal R, Christiansen IK, Ambur OH, Rounge TB. TaME-seq: An efficient sequencing approach for characterisation of HPV genomic variability and chromosomal integration. *Scientific Reports* 2019;9:524. doi: 10.1038/s41598-018-36669-6.

II. **Lagström S***, van der Weele P*, Rounge TB, Christiansen IK, King AJ, Ambur OH. HPV16 whole genome minority variants in persistent infections from young Dutch women. *Journal of Clinical Virology* 2019;119:24-30. doi: 10.1016/j.jcv.2019.08.003.
* Denotes equal contribution.

III. **Lagström S**, Hesselberg Løvestad A, Umu SU, Ambur OH, Nygård M, Rounge TB, Christiansen IK. HPV16 and HPV18 type-specific APOBEC3 and integration profiles in different diagnostic categories of cervical samples. *Manuscript submitted.*

1 INTRODUCTION

1.1 Human papillomavirus and cancer

Human papillomavirus (HPV) causes nearly 5% of all cancers worldwide [1]. HPV infection is a necessary cause of virtually all cervical cancers [2, 3], the fourth most common cancer in women worldwide [4]. HPV is also associated with a significant proportion of cancers in the oropharynx (31% of cancer cases are attributable to HPV) and anogenital regions, including vulvar (25%), vaginal (78%), penile (50%), and anal (88%) cancers [1, 5]. HPV infection is the most common sexually transmitted infection worldwide and more than 70 % of sexually active individuals of both sexes will be infected during their lifetime [6, 7]. However, only a small fraction of HPV infections will persist and cause progression to cancer [8].

The global HPV-related disease burden is larger in women than men. Worldwide, 8.6% of all cancers in women are attributable to HPV, while only 0.8% of cancers in men are caused by HPV. Cervical cancer is the most common HPV-induced cancer, causing 90% of all the cancer cases caused by HPV [1]. Cervical cancer affects more than 500,000 women worldwide, causing 266,000 deaths each year [4]. There is a considerable variation in cervical cancer incidence across geographical regions; 70 % of all the cervical cancer cases occur in less developed countries, while these countries account for >85 % of deaths caused by cervical cancer [1, 4].

To date, more than 200 HPV types have been identified but only a few types are known to cause cancer [9, 10]. Based predominantly on the association with cervical cancer, HPVs are divided into high-risk and low-risk types. Twelve high-risk HPV types (16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, and 59) have been classified as carcinogenic to humans according to the International Agency for Research on Cancer (IARC) [11]. The IARC working group considered eight HPV types (26, 53, 66, 67, 68, 70, 73, and 82) as probably or possibly carcinogenic [11, 12]. Of the high-risk types, HPV16 and 18 are associated with about 70% of all cervical cancers [13]. Low-risk types, including HPV6 and HPV11, generally cause benign diseases such as genital warts [14].

1.2 Natural history and pathology of cervical cancer

The major steps of cervical carcinogenesis include HPV infection with one or more high-risk HPV types, persistence rather than clearance, progression to cervical precancer and invasive cancer (Figure 1) [8, 15]. The peak prevalence of HPV infection occurs around 20–25 years of age and is associated with sexual transmission in youth [15]. More than 90% of HPV infections clear within two years after acquisition, and the remaining infections have a high potential for persistence [16].

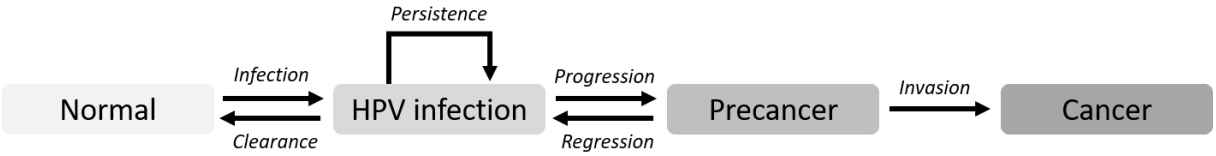

Figure 1. Natural history model of cervical carcinogenesis showing shifts between pathological diagnostic categories and the HPV infection. Adapted with permission from [17].

A persistent infection is a prerequisite for progression to cervical intraepithelial neoplasia (CIN). Low-grade neoplasia CIN1 is a histopathological sign of HPV infection, while CIN2 and CIN3 are considered as precancers and have the potential to develop to cancer [6, 15]. HPV infections persisting for more than two years are highly linked to precancer that is usually developed within 5–10 years [18]. In Norway, 1.5% of the screened population between 25 and 69 years of age were diagnosed with precancer in 2016 [19]. For all ages, the regression rate for CIN2 is estimated to be 50–70% [8]. There is little data available on the spontaneous regression of CIN3 because of the ethical reasons of not treating CIN3 lesions that hold the risk of developing into cancer [20]. Nevertheless, it has been estimated that the regression of CIN3 is likely close to 20–30% [8, 21].

Screening-detected precancers are treated to prevent them from developing into cervical cancer [15]. In unscreened populations, the peak of invasive cervical cancer occurs from about 35 to 55 years of age. The progression from CIN3 to cervical cancer is impossible to predict accurately due to ethical reasons, but based on an earlier study, estimates suggest a 30%–50% risk of invasion from precancer within 30 years [20]. The overall estimated risks for persistence of HPV infection and progression to invasive cervical cancer over time are presented in Figure 2.

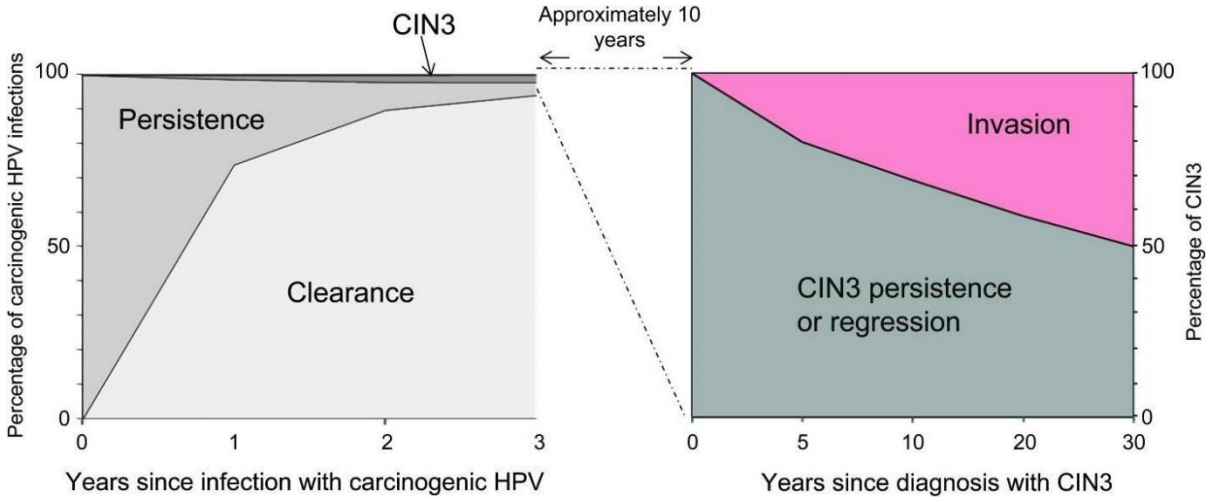

Figure 2. Risk of persistence of HPV infection and estimated risk of progression to invasive cancer over time. Reprinted with permission from [22].

HPV infects basal epithelial cells at the squamo-columnar junction between the squamous epithelium and the columnar epithelium in the cervix, which is highly susceptible to HPV infection. The virus makes its entry into the basal epithelial cells mainly through microlesions [23]. For squamous cell carcinomas (SCC), the lesions mainly occur in squamous epithelial cells at the squamo-columnar junction, while adenocarcinomas (ADC) arise in glandular cells in the endocervical canal. The columnar epithelium in the endocervical canal is replaced by squamous epithelium over time as part of a normal process depending on the woman's age, parity and hormonal status. A new squamo-columnar junction is formed between the newly formed squamous epithelium and the columnar epithelium, and the metaplastic epithelial area referred to as the transformation zone (Figure 3) [24, 25].

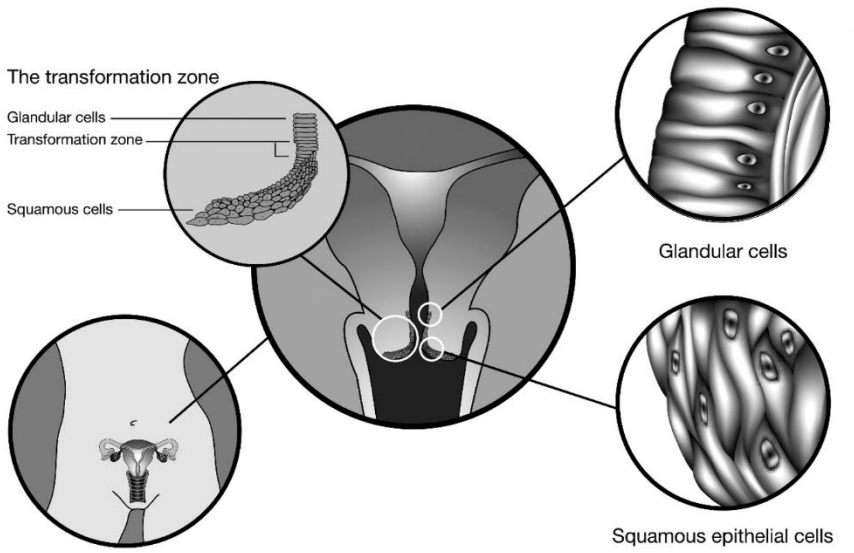

Figure 3. Location of the transformation zone, squamous and glandular cells in the cervix. Reprinted with permission from [26].

The majority of cervical cancers are SCCs, whereas ADCs of the uterine cervix accounts for 10–20% of all cervical cancers worldwide [27]. Incidence rates of ADC are increasing in the developed countries, including Norway [28, 29]. HPV16 is the most predominant type in SCC, while HPV18 and HPV45 are to a higher degree associated with ADC [13].

1.3 Cervical cancer prevention

1.3.1 HPV vaccination

HPV vaccination is used as primary prevention to control high-risk HPV infections and HPV-related cancers. Currently three prophylactic vaccines are commercially available: the bivalent Cervarix® (GlaxoSmithKline, London, UK), the quadrivalent Gardasil® (Merck, Kenilworth, New Jersey) and the nonavalent Gardasil® 9 (Merck, Kenilworth, New Jersey). All three vaccines protect against the most prevalent types HPV16 and HPV18 [15]. In addition to HPV16 and HPV18, Gardasil® protects against HPV6, HPV11, and Gardasil® 9 provides protection against HPV6, HPV11, HPV31, HPV33, HPV45, HPV52, and HPV58 [30, 31]. HPV vaccines consist of virus-like particles (VLP) derived from the HPV capsid protein L1, and they are designed to elicit virus-neutralising antibody responses [32]. VLPs resemble native virus particles but they contain no viral genetic material and are therefore non-infectious [15].

All vaccines are highly efficacious when administered according to the protocol prior to HPV exposure [33]. All three vaccines have been tested in large phase III randomised controlled trials in women 15–26 years of age. In these trials, efficacy of >90% against persistent HPV infection and precancer was shown in individuals without HPV infection at trial entry and at the completion of the three-dose immunisation trials [34-36]. Vaccination programmes typically target girls 9–13 years of age but also boys are targeted in recent years [37-39]. Today more than 80 countries, the majority being high- or upper-middle income countries, have introduced national HPV vaccination programmes. Low- and lower-middle-income countries have often the highest burden of cervical cancer and the most need for vaccination, but there are financial barriers to introduce vaccination programmes [40].

1.3.2 Cervical cancer screening

Cervical cancer screening is used as secondary prevention of cervical cancer [15]. Nevertheless, screening remains the most important prevention tool for cervical cancer in low- and lower-middle-income countries where no vaccination programmes are available [40]. The aim of screening is to detect precancers and early cancers that can be treated to reduce cancer mortality. A screening programme comprises screening of individuals, follow-up of individuals with a positive screening test and treatment if needed. In a cost-effective screening programme, the screened condition should be an important health problem to benefit screening. A screening test needs to be accurate and acceptable to the target population, most of whom are healthy, requiring an approach with minimal harm. Finally, an effective and well-tolerated treatment for individuals with a positive screening test must be available [41, 42].

Main tests for cervical cancer screening in developed countries are cervical cytology and HPV testing to detect high-risk HPV types. These two screening methods can also be combined, either as co-testing or one method being the primary screening method and the other being used as a secondary test to substantiate results from the primary test. Cytology, microscopic evaluation of the cells from cervical samples, is still the most frequently used screening test [15]. Today, the most common way to prepare samples for examination is liquid-based cytology (LBC); the most widely used LBC technologies are ThinPrep Pap Test (Hologic, Inc., Marlborough, MA) and BD SurePath™ liquid-based Pap test (Becton, Dickinson and Company, Franklin Lakes, NJ) [43]. LBC involves collection of cervical epithelial cells using a sampling brush, and preservation of the sample in a suspension before preparing the microscope slide or DNA extraction [44, 45]. Cytology screening has lower sensitivity than HPV testing and it requires shorter screening intervals to secure good sensitivity [46, 47]. HPV testing has proven to be more sensitive, but less specific than cytology as most HPV infections are transient [48, 49]. Several large, randomised clinical trials have demonstrated that HPV DNA testing in primary screening provides earlier detection of precancers and a reduced number of cancers during follow-up, allowing extended screening intervals [48].

The Norwegian cervical cancer screening programme was implemented in 1995 and invites women between 25 to 69 years of age to attend screening every three years when cytology is used as the primary screening method. In 2015, a controlled implementation pilot was started in four Norwegian counties to replace cytology by HPV testing as primary screening method for women 34–69 years of age [50]. By 2022, HPV testing will replace cytology as the primary screening method

in Norway for this age group, with a screening interval of five years. Also, as from 2018, the screening algorithm in Norway discriminates between HPV genotypes 16 and 18 versus other high-risk types, with a closer follow-up for HPV16 and HPV18 positive women [51].

1.3.3 Classification of cervical neoplasia

Different cytological and histological classification systems have been developed for cancer screening purposes (Figure 4). Precancerous lesions in cytological samples are commonly classified according to the Bethesda system [15]. The squamous lesions are classified as low-grade squamous intraepithelial lesions (LSIL) and high-grade squamous intraepithelial lesions (HSIL). The Bethesda system also allows uncertain results, using the terms atypical squamous cells- uncertain significance (ASC-US) and atypical squamous cells, cannot exclude HSIL (ASC-H) [52].

Histological classification of squamous cervical lesions includes the CIN scale, which is based on the severity of dysplasia [53]. It distinguishes CIN1 (mild dysplasia), CIN2 (moderate dysplasia) and CIN3 (severe dysplasia and carcinoma *in situ*) by the proportion of epithelium replaced by undifferentiated cells [54]. Precancerous lesions in glandular cells are not graded and are classified as adenocarcinoma *in situ* (AIS) [55]. A biopsy is taken for the histological analysis to diagnose CIN or cervical carcinoma [56] which is usually combined with colposcopy, a diagnostic procedure to visually inspect the illuminated cervix under magnification [57].

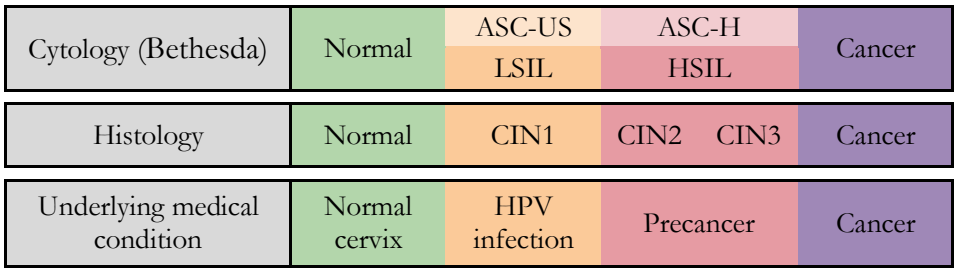

Figure 4. Cytological (Bethesda) and histological classification systems. Adapted with permission from [58].

1.3.4 Treatment of cervical lesions

According to the Norwegian guidelines, women diagnosed with high-grade cervical lesions by cytology (i.e., ASC-H or HSIL) are referred directly to colposcopy and biopsy [59]. Removal of abnormal cells is recommended in women when diagnosed with CIN2 or more severe lesions [60]. Loop electrosurgical excision procedure (LEEP) is the most common treatment of precancers in

developed countries [61]. A small electrical wire loop is used to remove abnormal cells from the cervix, and the procedure is performed under local anaesthesia. The majority of precancers would never progress to cancer in the absence of treatment [62] but the available screening methods cannot distinguish between precancers that will progress to cancer from those that regress. Therefore, management of CIN2 and CIN3 involves some overtreatment, but this is considered acceptable as long as excessive overtreatment is avoided [63]. All women with diagnosed cervical cancer are treated, and the type of treatment is determined by cancer stage [15].

1.4 Molecular biology of HPV

1.4.1 Genome structure

All papillomaviruses, including HPVs, have a circular double-stranded DNA genome of approximately 8000 bp (Figure 5) [64]. The HPV genome contains eight open reading frames (ORF) organised in three main regions [15, 65]. The early region (E) encodes genes E1, E2, E4, E5, E6 and E7, which are expressed early in the HPV life cycle [15]. E1 and E2 proteins are involved in viral replication and regulation of viral gene transcription [66]. E4 contributes to genome amplification and virus synthesis [67], while E5 contributes to the productive stage of the viral life cycle [68]. An essential function of E6 and E7, also referred to as the HPV oncogenes, is to drive cell cycle re-entry and genome amplification in the differentiating epithelial layers [66]. The late region (L) encodes genes L1 and L2, which are expressed late in the HPV life cycle and involved in virus capsid assembly [15]. The upstream regulatory region (URR), that is also referred to as the long control region (LCR), is a non-coding region, harbouring transcription factor-binding sites and controlling gene expression [66, 69]. The short non-coding region (NCR) between the genes E5 and L2 harbours a weak promoter activity for the L2 gene [70, 71].

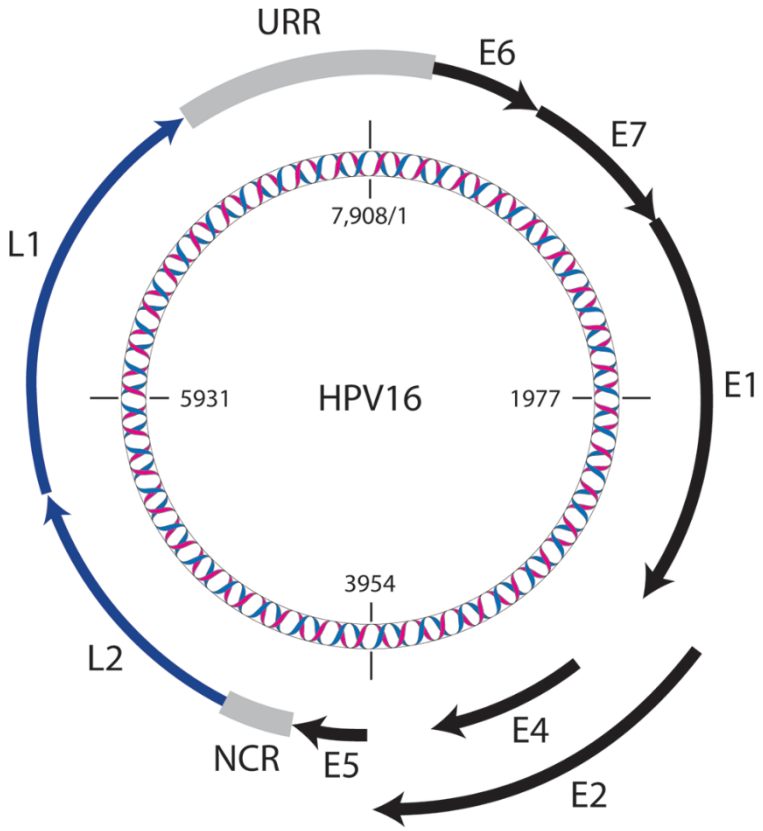

Figure 5. Schematic representation of the HPV genome, exemplified by HPV16, showing the location of early (E) and late (L) genes, URR and NCR. Adapted with permission from [72].

1.4.2 HPV classification

Papillomaviruses are a highly diverse family of viruses [65]. HPVs are suggested to co-evolve slowly with their hosts because they replicate their genome using the host replication machinery with a high degree of proof-reading, leading to low mutation rates [73, 74]. To date, more than 200 HPV types have been identified, infecting skin and mucosa in humans [9]. The L1 gene is the most conserved gene across HPV types. Therefore, identity at the nucleotide level in the L1 gene has been used for HPV classification [64]. Each individual HPV type shares at least 90% sequence identity in the L1 gene nucleotide sequence. Based on the nucleotide sequences of the L1 gene and often some additional HPV genes in different HPV types, a phylogenetic tree can be constructed to visualise the relationship between the different types. High-risk HPV types associated with cervical cancer [11, 75] are phylogenetically clustered within alpha-HPVs that contain alpha-5, alpha-6, alpha-7 and alpha-9 species groups (Figure 6) [76].

Figure 6. Phylogenetic tree based on the nucleotide sequence of HPV genes E7, E1, E2, L2 and L1. High-risk alpha-HPVs are highlighted by yellow, and grouped by alpha-5, alpha-6, alpha-7 and alpha-9 species groups. Reprinted with permission from [11].

Isolates of the same HPV types that differ by 1–10% across the genome are referred to as variant lineages, while sublineages of each HPV type have 0.5–1% sequence differences in the genome [76, 77]. The isolates of the same HPV type are closely related and nucleotide changes in the genome are not always evenly distributed throughout the genome. Therefore, the full genome sequence is the most accurate to classify HPV variant lineages and sublineages [76, 78].

The prevalence of the different HPV types and variant lineages in the population differs between geographical regions [79-81]. Worldwide, HPV16 is the most frequently detected HPV type of all HPV infections in cervix [79, 82]. It is also the most carcinogenic HPV type, alone responsible for more than 60% of all cervical cancers worldwide [13]. It remains unclear why HPV16 has higher prevalence and carcinogenicity compared to other closely related types, for example the alpha-9 sister viruses HPV31 and HPV35 [65]. The prevalence of HPV18 and HPV45, both from the alpha-7 species group, is higher in cancer than CIN3, indicating that lesions associated with these types are more likely to progress to cancer [83]. Additionally, it has been well established that HPV variant lineages carry different risks for disease outcomes despite the close phylogenetic relatedness [80, 81, 84-86].

1.4.3 HPV life cycle

The HPV life cycle, that is strictly controlled by the host cell differentiation, begins when the virus accesses the epithelial basal layer, usually through microlesions of mucosa or skin (Figure 7) [23]. Viral replication is performed by cellular polymerases in synchrony with replication of the cellular genome [23]. The HPV genome is maintained at low copy number episomes in the infected basal layer cells [66]. After migrating from the basal membrane, the infected cells initiate the differentiation program. E6/E7-mediated proliferation of basal cells induces the productive phase of the viral life cycle [65]. Following differentiation of epithelial cells, the expression of E6 and E7 is replaced by E1 and E2, that are thought to be essential for the initial genome amplification [66]. In addition, E4 and E5 contribute to replicating the virus to high copy number in the viral genome amplification process [23]. L1 and L2 are expressed in the upper layers of the epithelium, resulting in production of new virions that are released from the epithelial surface [69]. The non-enveloped icosahedral capsid of HPV is made up of 360 copies of the L1 gene product. The minor capsid protein L2 is thought to participate in virus capsid assembly and plays essential roles in the infectious entry pathway of HPV [87]. The HPV life cycle takes 2–3 weeks, which is the time for a cervical cell to migrate from the basal layers to the surface of epithelium and differentiate [88].

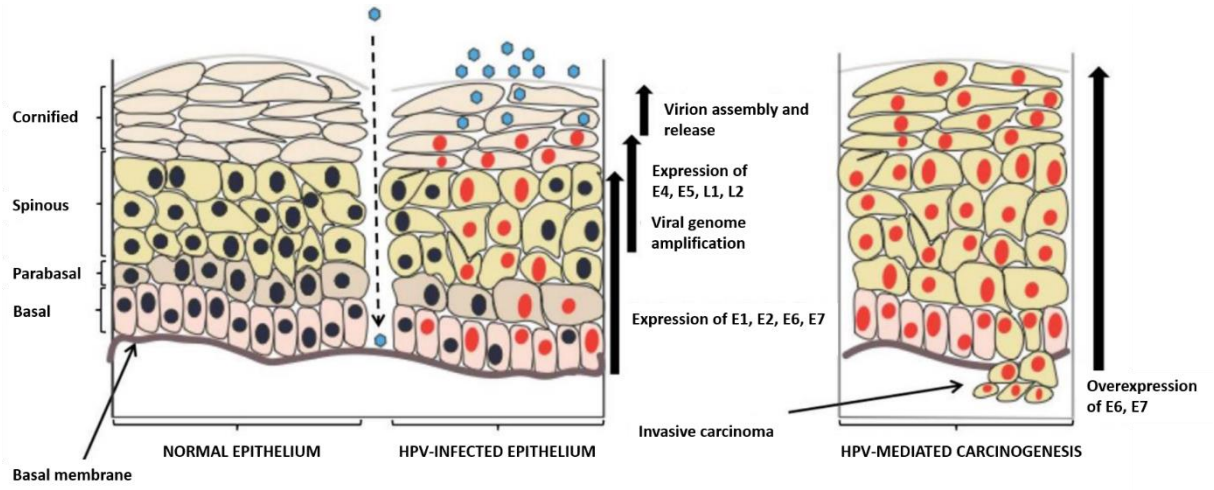

Figure 7. The HPV life cycle and HPV-mediated carcinogenesis. The left panel shows the HPV life cycle and expression of HPV genes during the cycle from infection to virus release. In the right panel, overexpression of E6 and E7 results in increased cell division and inhibition of the normal cellular differentiation, leading to malignant transformation of the infected cells. Adapted with permission from [89].

1.5 HPV-mediated cervical carcinogenesis

1.5.1 Molecular mechanisms of carcinogenesis

The life cycle of both high-risk and low-risk HPV types is similar but high-risk types have a unique ability to activate cell proliferation in the basal layers [15]. E6 and E7 proteins are essential for the viral life cycle since they increase viral fitness and viral production by driving cell cycle entry and genome amplification in the differentiating epithelial layers [90]. Although the E6 and E7 activity is present in both high-risk and low-risk HPV types, their role in low-risk types is largely insufficient to trigger the development of precancerous lesions or cancer [15].

Cell differentiation is required to complete the infectious life cycle of the virus, followed by virus assembly and release [6]. HPV gene expression may become deregulated in early neoplasia, although the mechanism is not fully understood. The increased activity of E6 and E7 underlies the development of cervical lesions, resulting in increased cell division, inhibiting both normal cellular differentiation and apoptosis [66]. Finally, the cells remain involved in cell-cycle progression, resulting in genomic instability that enables genetic alterations to accumulate [23]. Ultimately, this causes the malignant transformation of the infected cell, leading to precancerous lesions and eventually invasive cancer (Figure 7) [6].

Oncoproteins E6 and E7 are less conserved proteins and more specialised than the other viral proteins [91]. The increased expression and activity of E6 and E7 promotes cellular proliferation and inhibits the normal cell differentiation [6], most notably through the inhibition of tumour suppressor proteins p53 and retinoblastoma protein (pRB) (Figure 8) [92, 93]. p53 has a crucial role in aberrant cell cycle progression by inducing cell cycle arrest or apoptosis [66]. Inhibition of pRB activates the pRB/E2F pathway, promoting cell cycle entry and DNA synthesis [92, 94].

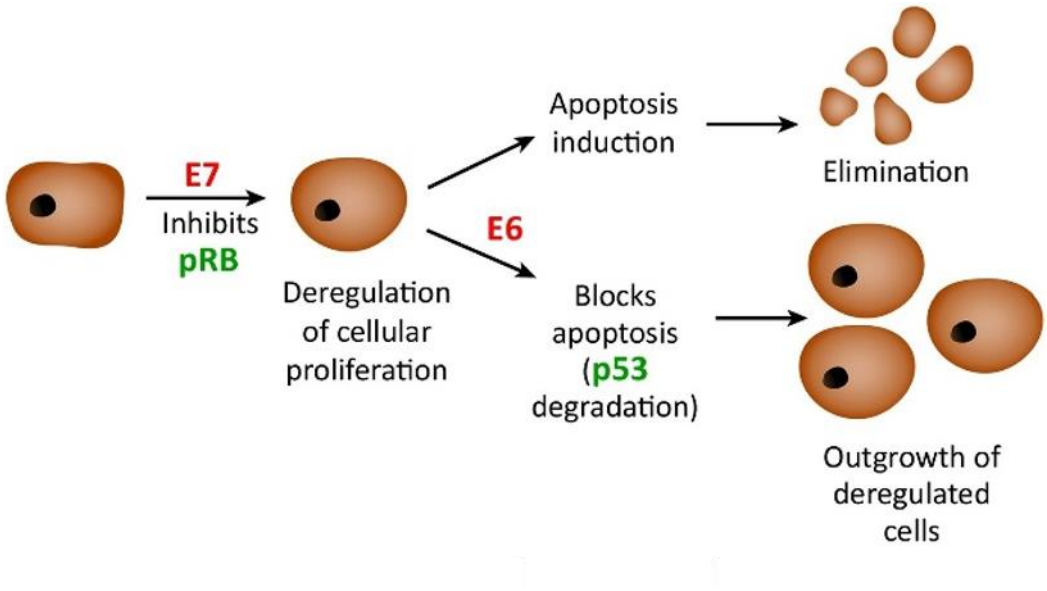

Figure 8. E6 and E7 oncoproteins during malignant cell transformation. E7 targets pRB, and the inhibition of pRB leads to stimulation of cellular proliferation. The expression of E6 induces the degradation of p53, allowing the outgrowth of deregulated cells. Adapted with permission from [95].

Deregulation of E6/E7 gene expression alone is not sufficient for the development of cervical cancer. Invasion requires the accumulation of additional alterations in the host genome, which is facilitated by the persistent overexpression of E6 and E7 [15]. In addition to deregulation of cell cycle control and accumulation of genetic damage, the development of invasive cancer also depends on immune evasion mechanisms that enable the virus to be undetected [6]. HPV has evolved several mechanisms to inhibit host antivirus natural and adaptive responses, of which avoidance of antigen presentation is suggested to be the primary mechanism of immune evasion [96].

1.5.2 Chromosomal integration

HPV integration into the host genome has been widely studied and is regarded as a driving event in cervical carcinogenesis [97-99]. However, the exact mechanisms for integration and the role of integration in cancer progression is not fully understood [100]. HPV integration events can be detected in precancerous lesions but the frequency of HPV integration events increases as cells progress to invasive cancer [101-103]. While integration events are very frequently detected in HPV-associated cancers, they are not required for cervical cancer development [99].

Disruption or complete deletion of the E1 or E2 genes, which regulate the expression of viral oncoproteins E6 and E7, is often observed upon integration, resulting in constitutive expression

of the E6 and E7 oncogenes [66, 104, 105], which in turn inactivate of cell cycle checkpoints and lead to genomic instability [99]. HPV-infected cells can contain either intact episomal HPV DNA, integrated DNA or both. HPV can also be integrated as viral-host head-to-tail concatemers (multiplied copies of same DNA sequence), often leading to amplification of expression of oncogenes E6 and E7 [106]. If the cancer cells harbour exclusively episomal HPV DNA, the viral genome may have acquired other genetic or epigenetic changes, such as methylation of the E2 binding sites in URR, which may result in dysregulated E6/E7 gene expression [107, 108].

Viral integration may lead to rearrangements, deletions, and amplification in the host genome as well as disruption or modified expression of cellular genes, including oncogenes and tumour-suppressor genes, which may promote carcinogenesis [99, 109]. Chromosomal integration sites are distributed across the human genome [110]. However, integration sites in certain regions, such as chromosomal loci 3q28, 8q24.21 and 13q22.1, have been reported more often than others, suggesting a non-random distribution of integration sites [111, 112]. Indeed, the hot-spot regions are often associated with common fragile sites [112, 113] and transcriptionally active regions of the genome [100, 110], which may expose these regions for viral integration [110]. These hot-spot regions are gene-rich regions with several important oncogenes and tumour-suppressor genes [100]. In addition, regions of identical short sequences, defined as microhomology, between viral and human genomic sequences have also been found at integration breakpoints [102, 114].

A recent study of The Cancer Genome Atlas Research Network [115] showed that HPV integration occurred in >80% of HPV positive cervical cancers. Of these, HPV integration was observed in all HPV18 positive samples and in 76% of the HPV16 positive samples [115]. This result is consistent with other observations of integration frequencies in HPV16 and HPV18 positive cervical lesions [98, 116]. Integration in high-risk HPV types other than HPV16 and HPV18 is less studied. Higher amount of integrated HPV with increasing lesion severity is reported for HPV31, 33, 45, 52, 58, but the integration frequencies vary considerable between the different HPV types [103, 117-119].

1.5.3 HPV genomic variation

Studies have shown that closely related HPV variant lineages can differ in their carcinogenic potential [80, 81, 85, 86, 120]. HPV16, the most carcinogenic HPV type, can be divided into four main variant lineages (A, B, C, D), and into ten sublineages (A1, A2, A3, A4, B1, C1, D1, D2, D3,

D4) (Figure 9) [76]. Although the HPV16 variant lineages differ only as little as 1% (~80 bp) at the whole genome level, the non-A variant lineages (B/C/D) are associated with an increased risk of CIN3 or cervical cancer compared to A variant lineage [80, 120]. However, the distribution of HPV16 variants worldwide is specific to geography and ethnicity, limiting the possibility to compare the carcinogenic potential of the HPV16 variants in a standardised way across all regions [80, 85, 121]. In contrast, different HPV18 variant lineages represent equal risk of cancer [81, 122].

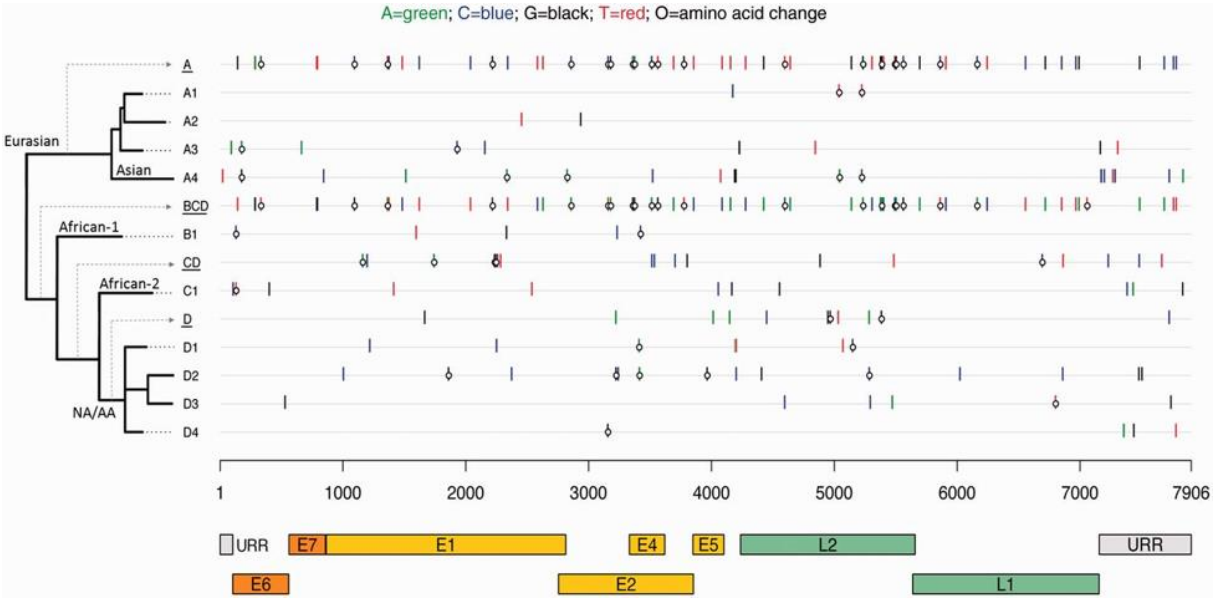

Figure 9. Nucleotide and amino acid changes across the complete genome of HPV16 variant lineages and sublineages. The x axis shows HPV16 genomic positions, aligned according to the sublineage in the phylogenetic tree on the y axis. The nucleotide changes are colour-coded as shown at the top figure, and the open circles represent amino acid changes. Reprinted with permission from [120].

Below the levels of variant lineages and sublineages, there is another level of variability with <0.5% differences in the viral genomes [123]. This level of variability is still scarcely studied. A few studies have reported a high level of variability in HPV16 and HPV18 genomes, indicating high genetic diversity of HPV16 and HPV18 between infected individuals [124-127]. Also high intra-host HPV variability with low-frequency minor nucleotide variants (MNVs) has been reported [123, 128, 129]. Since the HPV genome replication is dependent on host cell high-fidelity DNA polymerases [73], other mechanisms may underlie the high mutation rate in HPV genomes [123].

It has been suggested that the apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like 3 (APOBEC3) may be an important source of HPV genome mutagenesis [130]. APOBEC3 proteins are expressed in epithelial cells, which are the target cells of HPV infection. The APOBEC3 protein binds to single-stranded DNA and causes cytidine (C) to thymidine (T)

mutations during viral genome synthesis [131]. A preferred trinucleotide context for ABOPEC3 has been shown to be TCN, where N is any nucleotide [132]. APOBEC3 is part of APOBEC protein family consisting of 11 enzymes that have important roles in the innate immune defence. APOBEC proteins edit the viral genome, and the mutations may lead to defects in viral genome replication, thereby restricting the viral life cycle [133]. APOBEC3-mediated mutagenesis has been reported in HPV positive cervical samples, mainly in women with low-grade or benign HPV infections [123, 134]. APOBEC-induced mutations have also been described in many other viruses [135-137], as well as in several human cancers [138, 139].

Nucleotide variation can be categorised into synonymous, non-synonymous and nonsense nucleotide substitutions. Synonymous nucleotide substitutions do not alter the amino-acid sequence in a protein, whereas non-synonymous or nonsense substitutions lead to an altered amino-acid sequence or an early stop codon in the encoded protein, respectively [140]. Non-synonymous substitutions are mostly deleterious and are eliminated by purifying or negative selection. Substantial genetic diversity can occur during the course of viral infection, and the evolutionary process may be important for the persistence of the infection. Non-synonymous variants leading to amino acid changes may be favoured through positive or diversifying selection if they increase the survival of the virus, e.g. by evading the host immune system [141]. Selection pressure on protein-coding sequences is commonly estimated by the ratio of the nonsynonymous substitution rate (dN) to the synonymous substitution rate (dS). If the ratio dN/dS is higher than 1, positive selection is assumed to have occurred, while the ratio dN/dS less than 1 indicates negative selection [142, 143].

1.6 Molecular approaches in HPV screening and research

1.6.1 HPV detection and genotyping

HPV testing is widely used in cervical cancer screening programmes to detect early disease. As a result, more than 200 molecular HPV tests are commercially available [144, 145]. The first commercial HPV tests reported results only as HPV positive or negative with no HPV genotype information [144]. Newer generation of commercially available methods is based on HPV or RNA detection and report 1) HPV16 and HPV18 genotypes separately, and other high-risk types as pooled results, 2) ≥5 genotypes and pooled detection of the remaining high-risk HPV genotypes, and 3) full genotyping with individual identification of carcinogenic HPV genotypes [144, 145].

The sensitivity of the commercial assays for HPV detection is based on clinically validated cut-off values specific for each test [45].

In Norway, HPV tests used in primary and secondary screening should fulfil the international criteria for HPV screening [146]. Totally, six HPV tests fulfil these criteria and are approved for primary cervical cancer screening in Norway, and the different HPV tests are approved to be used with SurePath and/or ThinPrep LBC systems [147]. HPV genotyping, identifying at least the most carcinogenic HPV types (mainly HPV16 and HPV18), has been shown to be a potential tool in screening and can be used for risk stratification of cervical lesions [145, 148]. HPV tests reporting full genotyping information are widely used research tools for epidemiological studies, vaccine development, implementation and monitoring of vaccination programmes [145].

Today, techniques based on polymerase chain reaction (PCR) are widely used in HPV assays because they are highly sensitive and non-laborious [149]. During cycling in a conventional PCR, the DNA polymerase extends a pair of oligonucleotide primers that flank the region of interest, resulting in exponential amplification of a single double-stranded DNA molecule [150]. Real-time PCR, also called quantitative PCR (qPCR), is used to amplify and simultaneously detect and quantify the absolute or relative amount of DNA molecules using fluorescent dyes or probes, allowing real-time quantification of the amplified target DNA sequence [149].

HPV PCR assays commonly use consensus primers that target conserved regions of the L1 gene [151-154]. The consensus primers are used to amplify multiple HPV genotypes in a single reaction, but use of consensus primer systems has limitations with regard to the detection sensitivity of specific HPV types [155, 156]. For better sensitivity, HPV type-specific primers are used in various assays, usually targeting the E6 and E7 oncogenes [155, 157, 158], which harbour most nucleotide sequence variation between different HPV types and are retained by the infected cells due to their importance in the carcinogenesis [159]. HPV detection can also be based on detection of mRNA from the E6/E7 oncogenes [150].

1.6.2 Characterisation of HPV integration

Over the last years, several methods have been established to facilitate the identification of integrated HPV genomes [149]. The assays provide information about the presence of integrated viral genome, which may indicate accumulation of chromosomal damages in the infected cells,

eventually leading to development of cervical cancer [99]. Several PCR-based methods are developed for the detection of HPV integration into the host genome, including the detection of integrated papillomavirus sequences (DIPS), the assay of papillomavirus oncogene transcripts (APOT) and real-time PCR [149].

The DIPS method is an adapter ligation-based PCR assay. Adapter ligation is followed by nested PCR: the first round with HPV-specific primers and the second round with HPV and adapter-specific primers. DIPS reveals the locus of integration but the method is unable to discriminate between integrated and episomal DNA [160].

APOT is a method that can identify the episomal, mixed and integrated viral forms with high sensitivity. The method is based on the detection and analysis of HPV E6 and E7 transcripts. After the reverse transcription and adapter ligation of the total RNA, the cDNA is subjected to nested PCR using a forward primer specific to HPV and a reverse primer specific to the adapter. APOT identifies integration through active transcripts of E1 or E2 genes, but the method ignores integration in other HPV genomic regions [161].

HPV integration detection using real-time PCR is based on quantification of the E2 and E6 genes. Viral physical status is determined through E2/E6 ratio analysis: equivalent amounts of E2 and E6 genes indicate the presence of episomal DNA, while disruption events in the E2 gene indicate the mixed or integrated viral forms. The method provides significant sensitivity in identifying viral integration events through E2/E6 ratio analysis, but it is unable to detect integration events elsewhere in the HPV genome [162-164].

1.6.3 Next-generation sequencing technologies

For more than two decades the gold standard of sequencing was Sanger sequencing [165], but it lacks the sensitivity needed to detect low-frequency mutations due to its variant allele detection limit of 10–20% [166]. The advances in next-generation sequencing (NGS) technologies, such as whole genome sequencing (WGS), whole exome sequencing (WES), whole genome bisulfite sequencing (WGBS) and RNA sequencing (RNA-seq), have provided new comprehensive tools for genome research [167]. The NGS technologies enable massively parallel sequencing analysis, high throughput, and detection of low-frequency mutations for reduced cost [168, 169]. Several NGS platforms have been developed, including MiSeq, NextSeq, HiSeq and NovaSeq sequencing

platforms (Illumina), Ion Torrent (Life Technologies), PacBio (Pacific Biosciences) and Nanopore (Oxford Technologies), which all are currently in use. With a rapidly advancing field, many NGS platforms, including 454 pyrosequencing (Roche) and SOLiD (Applied BioSystems), have already become obsolete [149, 170].

Illumina has a dominant position in terms of market share and the amount of sequence data their platforms can produce [170]. Illumina sequencing platforms are based on sequencing-by-synthesis (SBS) technology [168]. Illumina platforms require preparation of sequencing libraries where fixed adapters are linked to the target molecule. There is an increasing number of template preparation methods for different applications on the Illumina platforms [171, 172]. First, the sequence library with adapters is denatured to single strands and it is inserted into a flow cell. Bridge amplification is performed subsequently to generate template clusters. Finally, the sequencing starts when the four different fluorescently tagged nucleotides (ddATP, ddGTP, ddCTP, ddTTP) synthesise the new strand according to the sequence at the template. The signal is detected through a charge-coupled device [168]. The Illumina platforms support read lengths up to 2×300 bp [149, 173].

Illumina platforms are known to cause sequence-specific errors in inverted repeats and nucleotide patterns GGC and GGT [174, 175]. Because sequencing on the Illumina platforms is performed in cycles and reagents may be affected over time, signal and thus base calling quality can be affected towards the end of the run. The decrease of signal quality with increasing cycle number is a well-established characteristic on Illumina platforms [176, 177]. Paired-end (PE) sequencing that allows sequencing of both ends of a fragment, produce data of better quality compared to single-end (SE) sequencing since errors can be corrected by overlapping the sequenced paired-end reads [176].

1.6.4 NGS applications in HPV research

The NGS technology provides a wide range of application for HPV genotyping, detection of multiple HPV infections and analysis of HPV genomic variability and integration [149, 178]. The first studies using NGS for HPV genotyping were based on the analysis of L1 gene amplicons using Roche 454 sequencing technology [179, 180]. Several PCR-based HPV genotyping assays using Illumina platforms or IonTorrent have been developed, including conventional multiplex PCR approaches [181-184] and rolling circle amplification [185]. Furthermore, hybridisation based-target capture technologies to enrich target genome have recently been used for HPV genotyping

[186, 187]. Since these NGS genotyping strategies rely on PCR enrichment or target capture, design of the primers and probes is an important part of the assay [178].

To study HPV genomic variability, a few studies report the use of long-range PCR to amplify the whole HPV genome followed by a NGS library preparation [129, 188]. A multiplex PCR-based amplification panel for targeted sequencing of the HPV16 genome has been used in several studies of HPV genomic variation and HPV gene conservation [116, 124, 127, 189]. Methods targeting specific HPV genomic regions are used to study nucleotide changes and variants in a larger number of samples for epidemiological purposes and risk stratification [128, 190]. Initially, WGS and WES were used to analyse HPV integration sites in cervical cancer samples [106, 191]. Recently methods with reduced sequencing costs have been developed for detection of integration sites, including PCR-based approaches [192], target capture technologies [101, 102, 193] and RNA-seq [194].

1.6.5 NGS data analysis

The NGS data analysis is arranged in a stepwise process called pipeline [195]. A typical pipeline consists of: 1) quality control of sequencing reads, including filtering and trimming of the reads, 2) aligning reads to a reference genome, 4) identification of variants, and 5) annotation of variants [169]. If no reference exists for the sequenced genome, step 2 can be replaced by a *de-novo* genome assembly that is a method assembling the short sequencing reads to create a full-length sequence or genome without a reference sequence [195].

Raw sequencing outputs are nucleotide base calls and their quality values, which are usually stored in the form of FASTQ files [196], the standard input for most NGS pipelines. Quality values assigned to each base are based on Phred quality scores that are defined in terms of the estimated probability of incorrect base call [197, 198]. For instance, probability of incorrect base call with a Phred score 20 is 1 in 100, and with a Phred score 30 it is 1 in 1000. A quality control (QC) of raw sequence reads is an initial check of the input data. The QC step detects part of the sequencing artefacts created during library preparation and sequencing, while many of the artefacts may become apparent at later steps of the analysis [195]. Trimming of low quality reads or bases at the end of reads and filtering of any non-biologically relevant sequences (e.g. adapters), that could otherwise lead to confusing or biased results, are usually performed in the QC step [199, 200].

The next step, mapping, is matching of the reads to the reference genome. This is performed by aligning reads to the reference genome(s) to which they are most similar in terms of nucleotide sequence. Sequence mapping is usually the most time and memory-consuming step of a pipeline [201]. The well-known BLAST [202] algorithm is too slow to use for NGS because of the massive amount of short reads. Therefore, specific time and memory-optimised aligning algorithms, short read aligners, are developed. Aligners vary in their methods, computer resource usage and sensitivity, and they may result in different mapping results [203]. Some commonly used aligners are BWA [204] and Bowtie [205]. The Sequence Alignment/Map (SAM) and Binary Alignment/Map (BAM) are generic alignment formats for storing NGS read alignments [206].

Variant calling from NGS data refers to identifying variant nucleotides in the sequence relative to the reference genome [207]. Variant calling includes small-scale variants [208], such as SNPs (single-nucleotide polymorphisms), insertions and deletions (indels) [209], and large-scale structural variants, copy number variants (CNV), inversions, and translocations [195]. Several variant callers are available; some of them are designed for germline variant calling, while others are more suitable for calling somatic variants [195]. Variant callers usually apply minimum depth of coverage and quality of called bases for filtering and trimming [207]. Finally, variant calling is based on a predefined threshold for variant allele frequency (VAF), which the proportion of the variant base of all bases at a given position [210]. Many pipelines also apply PCR duplicate removal step, removing sequences from amplification of same original PCR products, before variant calling to minimise the risk of false positive calls [211]. An annotation step is often performed after the variant calling. The most common way to annotate the variants is to provide database links to public variant databases, such as dbSNP or 1000 Genome Projects [212]. Finally, annotation is often followed by visualisation that can be useful for interpreting results [213].

2 AIMS OF THE STUDY

High-risk HPV is identified as a necessary cause of cervical cancer. Nevertheless, only a small fraction of HPV infections may progress to cancer, indicating that additional molecular factors, such as nucleotide variation in the HPV genome and chromosomal integration, contribute to the carcinogenic process. Current HPV tests used in screening programmes have a high sensitivity for the detection of HPV infection but they are unable to predict the risk of persistence of infection and progression to cervical cancer. The extent and nature of both HPV genomic variability and integration events can reveal new insight into HPV-induced carcinogenesis and can be used for assessing risk of developing cervical cancer.

The aim of this study was to develop a novel NGS method to characterise HPV genomic variation and integration, and to apply the method on clinical samples to explore the HPV genomic events contributing to HPV-induced carcinogenesis. The genomic events in HPV positive samples were assessed and analysed in both longitudinal follow-up and cross-sectional study settings.

Three specific research objectives were assessed in separate papers as follows:

Paper I: To develop a cost-effective NGS method for simultaneous characterisation of HPV genomic events, such as genomic variability and chromosomal integration, with reduced cost and hands-on time in the laboratory.

Paper II: To characterise genomic variation at the minor variant level in persistent HPV16 infections in follow-up samples from same women.

Paper III: To compare HPV minor nucleotide variation and integration profiles in HPV16 and HPV18 positive cervical samples with different morphology.

3 MATERIALS AND METHODS

3.1 Sample material and study design

Samples included in *Papers I-III* are listed in Table 1. Additional details, such as recruitment and inclusion criteria and detailed methods, are reported in the indicated publications.

Table 1. Study samples, HPV types and study designs in *Papers I-III*.

Study	Samples	HPV types	Study design	Additional details
Paper I	31 HPV positive LBC samples with ASC-US/ LSIL cytology 4 HPV positive cervical cancer cell lines 3 HPV plasmids	16, 18, 31, 33, 45	Method development and validation	
Paper II	59 HPV16 positive vaginal self-swabs with unknown cytology	16	Longitudinal follow-up study	[125, 214-217]
Paper III	157 HPV16 positive samples, including samples with normal/ASC-US/LSIL cytology and CIN2/CIN3/AIS/cancer histology* 75 HPV18 positive samples, including samples with normal/ASC-US/LSIL cytology and CIN2/CIN3/AIS histology*	16, 18	Cross-sectional study	[218, 219]

* Cytological samples taken at the time of histological diagnosis.

Paper I

Anonymised LBC samples from routine cervical cancer screening, diagnosed with ASC-US or LSIL, were included. DNA from commercial cervical cancer cell lines CaSki (HPV16), SiHa (HPV16), HeLa (HPV18) and MS751 (HPV45) (ATCC, Manassas, VA) were used as positive controls for the specified HPV types. In addition, WHO international standards for HPV16 and HPV18 (NIBSC, Potters Bar, Hertfordshire, UK) and a plasmid containing the strain HPV33 [220] were used as positive controls for method development purposes.

Paper II

Vaginal self-swabs were obtained from the *Chlamydia trachomatis* Screening and Implementation study performed in the Netherlands [215-217]. All the samples were HPV16 positive but the cytology was not performed on the samples. Samples with a persistent HPV16 infection in at least

three consecutive time points were included, with the median sampling interval being 48 weeks (95% CI: 46–51 weeks; min: 17, max: 63 weeks).

Paper III

Cervical samples were collected from women attending the cervical cancer screening programme in Norway between January 2005 and April 2008 [218, 219], and stored in a research biobank at Akershus University Hospital. For this study, a non-progressive category of infection was defined, consisting of: 1) samples with normal cytology (at enrolment and during the preceding two years, and with no previous history of treatment for cervical neoplasia) and 2) ASC-US/LSIL samples from women with no history of cervical abnormality and with no follow-up diagnosis within four-year follow-up. Cytological samples in each category of progressive disease were included, representing women with histologically confirmed CIN2, CIN3, AIS and cervical cancer, including cases of SCC and ADC. Finally, all samples at the biobank tested positive for HPV16 and/or HPV18, alone or together with other HPV types, were included in the study, with the exception of HPV16 CIN3 category of which a random selection of 50 cytological samples was included.

3.2 DNA extraction and HPV genotyping

Paper I

HPV positive samples with the cobas 4800 HPV test (Roche Molecular Diagnostics, Pleasanton, CA) were subjected to DNA extraction using NucliSENS easyMag (BioMerieux Inc., France). The samples were genotyped using the MGP PCR protocol [154], followed by HPV type-specific hybridisation using Luminex suspension array technology [221] or the Anyplex™ II HPV28 assay (Seegene, Inc., Seoul, Korea).

Paper II

Total DNA was isolated using Total Nucleic Acid Isolation Kit and the MagnaPure96 platform (Roche Molecular Systems, Pleasanton, CA) according to the manufacturer's instructions. HPV genotyping was previously performed using the SPF_{10}-DEIA-$LiPA_{25}$ platform (DDL Diagnostic Laboratory, Voorburg, Netherlands) [153, 222].

Paper III

Nucleic acid extraction and HPV genotyping were performed in the original studies [218, 219]. Total nucleic acid was extracted using the semi-automatic miniMag or automatic easyMag (BioMerieux Inc., France) extraction protocol, suitable for both HPV DNA and mRNA testing. HPV DNA testing was previously performed by Amplicor HPV test (Roche Molecular Systems, Pleasanton, CA) followed by genotyping by Linear Array HPV assay (Roche Molecular Systems, Pleasanton, CA). HPV mRNA testing was performed by PreTect HPV Proofer (PreTect AS, Klokkarstua, Norway).

3.3 DNA concentration and viral load

In *Papers I* and *II*, DNA concentration was measured by Qubit™ dsDNA BR Assay Kit (Thermo Fisher Scientific, Waltham, MA). In *Paper III*, DNA concentration was measured by Quant-iT™ Broad-Range dsDNA Assay Kit (Thermo Fisher Scientific, Waltham, MA).

In *Paper II*, viral load was determined using an adapted L1-targeting qPCR protocol [223]. Amplification was performed on the Roche LightCycler 480 platform (Roche Molecular Systems, Pleasanton, CA) according to the detailed protocol [214]. Viral load was determined based on cycle threshold (Ct) values in relation to the HPV16 containing plasmid that was used as a standard.

3.4 TaME-seq

The in-house developed sequencing assay TaME-seq (tagmentation-assisted multiplex PCR enrichment sequencing) was used for sample preparation in *Papers I-III*. Detailed TaME-seq workflow is described below.

3.4.1 Primer design

HPV16, 18, 31, 33, and 45 whole genome reference and variant sequences were obtained from the Papillomavirus Episteme (PaVE) database [224, 225]. All the reference and variant sequences within an HPV type were aligned using the multiple sequence alignment tool ClustalO [226]. Sequence alignment was converted to a consensus sequence for each HPV type in CLC Sequence viewer (v7.7.1, QIAGEN Aarhus A/S). Primer design was performed using Primer3 [227].

Sequencing primers on both strands were designed using HPV consensus sequences as the source sequence. HPV18 primers are illustrated as an example in Figure 10. Primer specificity was controlled using a BLAST search [202] against the human genome (GRCh38/hg38). Cross-binding between HPV types was evaluated against HPV16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66 and 68 consensus sequences (based on PaVE database reference and variant sequences as described earlier). Primers were modified by adding an Illumina-compatible adapter tail (5'-AGACGTGTGCTCTTCCGATCT-3') to the 5'-end. Primers were synthesised by Thermo Fisher Scientific, Inc. (Waltham, MA).

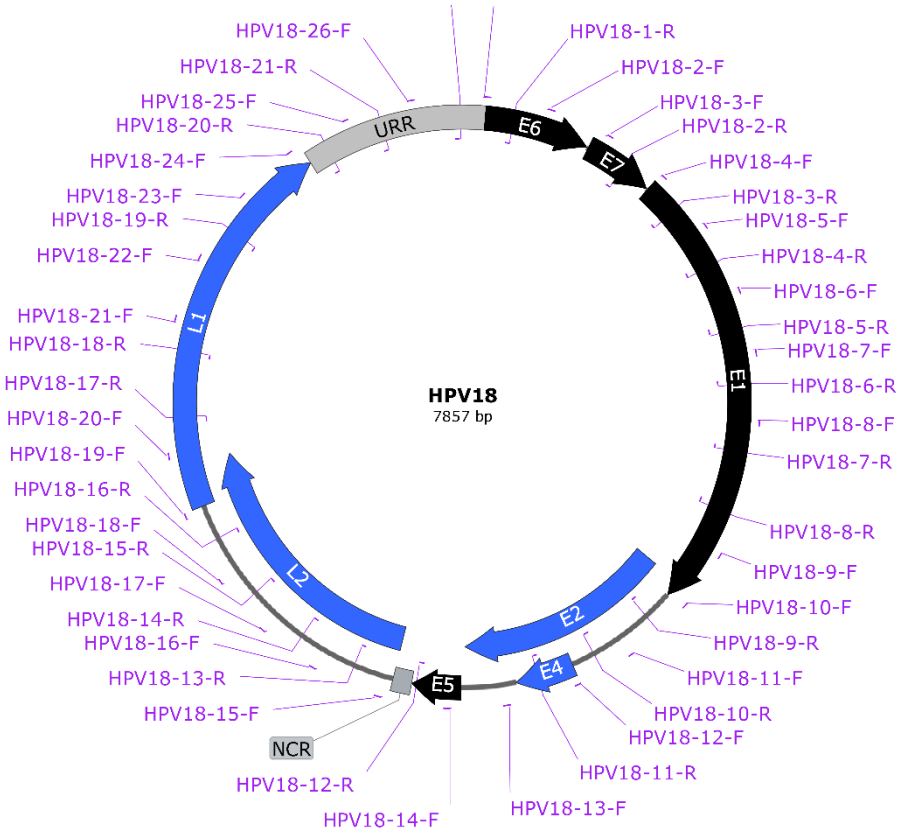

Figure 10. HPV18 genome and location of HPV18 primers. HPV genes are marked in black (early genes), blue (late genes) and grey (URR, NCR). Primers were designed on both strands (F/R), and marked in lilac. Figure is created with SnapGene (v4.2.6).

3.4.2 Library preparation and sequencing

Primer pools for each HPV type were prepared by separately pooling equimolar amount of F and R primers. Sample DNA was diluted to 2.5 ng/µl, or if the original DNA concentration was <2.5 ng/µl samples remained undiluted. Samples were subjected to tagmentation using Nextera DNA library prep kit (Illumina, Inc., San Diego, CA) with following modifications: i) reaction volume

was downscaled to 20 µl, ii) DNA input amount varied from 0.96 ng to 20 ng, iii) incubation was performed at 55°C for 4 minutes. Tagmented DNA was purified using DNA Clean & Concentrator™-5 columns (Zymo Research, Irvine, CA) according to the manufacturer's instructions or ZR-96 DNA Clean & Concentrator™-5 plates (Zymo Research, Irvine, CA) according to the Nextera® DNA Library Prep Reference Guide (15027987 v01). Tagmented DNA was subjected to PCR amplification for target enrichment. Amplification was performed in 20 µl containing 5 µl of tagmented DNA, 10 µl of Qiagen Multiplex PCR Master Mix (Qiagen, Hilden, Germany), 2 µl of Q-solution (Qiagen, Hilden, Germany), 0.75 µM of HPV primer pool, 0.5 µM of i7 index primer [228], and 1 µl of i5 Nextera index primer (Illumina, Inc., San Diego, CA). The cycling conditions were as follows: initial denaturation at 95 °C for 5 minutes; 30 cycles at 95 °C for 30 seconds, at 58 °C for 90 seconds and at 72 °C for 20 seconds; final extension at 68 °C for 10 minutes.

Following amplification, libraries were pooled in equal volumes and purified with Agencourt® AMPure® XP beads (Beckman Coulter, Brea, CA) according to the manufacturer's instructions. The quality and quantity of the pooled libraries were assessed on Agilent 2100 Bioanalyzer using Agilent High Sensitivity DNA Kit (Agilent Technologies Inc., Santa Clara, CA) and by qPCR using KAPA DNA library quantification kit (Kapa Biosystems, Wilmington, MA). Sequencing was performed on the Illumina MiSeq or HiSeq2500 platforms (Illumina, Inc., San Diego, CA) as 125 or 150 bp paired-end reads.

3.5 Sequencing data analysis

Sequencing data analysis was performed using in-house Python scripts collected together by the Snakemake workflow management system [229].

3.5.1 Sequence alignment

Raw paired-end reads were trimmed for Illumina adapters, HPV primers, quality (-q 20) and finally for minimum length (-m 50) using cutadapt (v1.10) [230]. Trimmed reads were mapped to human (GRCh38/hg38) and HPV reference genomes obtained from the PaVE database [224, 225] using HISAT2 (v2.1.0) [231]. Mapping statistics and sequencing coverage were counted using Pysam package [206].

3.5.2 Sequence variation analysis

For *Papers I-III*, mapped nucleotide counts over HPV reference genomes and average mapping quality values of each nucleotide were retrieved from BAM files and variant calling was performed using an in-house R (v3.5.1) script. Nucleotides observed ≤2 times in each position were filtered out. Final variant calling parameters used in each paper are listed below.

Paper I

Nucleotides with mean Phred quality score of <20 were filtered out. MNVs were called if VAF was >0.2% and sequencing depth was ≥100×. Samples with <20000 reads were excluded from the analysis.

Paper II

Nucleotides with mean Phred quality score of <30 were filtered out. Variants were called if VAF was >1%. Samples with <45% of the genome covered by minimum 100× were excluded from the analysis.

Paper III

Nucleotides with mean Phred quality score of <20 were filtered out. Variants were called if VAF was >1%. Samples with a mean sequencing depth of <300× were excluded from the analysis. The ratio of non-synonymous to synonymous substitutions (dN/dS) was calculated to indicate potential positive or negative selection affecting protein-coding genes.

3.5.3 Mutational signature analysis

Paper II and *III*

All nucleotide substitutions based on sequence variation analysis were classified into six base substitutions, C>A, C>G, C>T, T>A, T>C, and T>G, and then into 96 trinucleotide substitution types including information on the bases immediately 5' and 3' of the mutated base. Analysis was performed using an in-house R (v3.5.1) script.

3.5.4 Construction of phylogenetic tree

Paper II

Consensus sequences were created from BAM alignment using samtools (v1.8) mpileup (-E -d 200000 -L 200000) followed by bcftools (v1.6) (call -c --ploidy 1) and vcfutils.pl tools. Phylogenetic tree was constructed using MUSCLE (v.3.8.1551) to align sequences, IQtree (v1.5.5) to infer maximum likelihood phylogeny and FigTree (v1.4.3) to visualise the alignment.

3.5.5 Detection of integration sites and HPV genomic deletions

Paper I and *III*

The paired-end reads that mapped (HISAT2) with one read to a human chromosome and the other read to the target HPV reference genome were identified as discordant read pairs. Junction reads were identified to determine the exact position of HPV-human integration breakpoints; previously unmapped reads were re-mapped using the LAST (v876) aligner (options -M -C2) [232]. Positions covered by unique ≥2 read pairs (HISAT2) or by ≥3 junction reads (LAST) were considered as potential integration breakpoints. Integration detection was not based on reads sharing the same start and end coordinates as these reads were considered as potential PCR duplicates.

In *Paper III*, any sample with a mean depth of >1000× and <85% of the genome covered by minimum 100× were manually inspected using IGV (v2.3.09) to detect HPV genomic deletions of >1 kb.

3.6 Validation of integration sites

Paper I and *III*

Sanger sequencing was used to validate all the potential HPV integration sites detected by the integration analysis. Primers were designed to flank the integration breakpoint, with one primer binding site locating in the human genome and one in the HPV genomes. DNA strand used for primer design was based on human and HPV genome orientations at the breakpoint. SAM flags [206] of discordant read pairs were used to verify the genome orientation and finally one primer on each human and HPV +/- DNA strand was selected for validation (Table 2).

Table 2. Genome and read orientations at the integration breakpoint and +/- DNA strand used for primer design.

Genome orientation at the integration breakpoint		SAM flag of the discordant read pair		Primers designed on +/- DNA strand	
Human	**HPV**	**Human**	**HPV**	**Human**	**HPV**
→	→	97	145	+	+
→	←	65	129	+	-
←	←	81	161	-	-
←	→	113	177	-	+

All primers designed for validation of integration sites are listed in Table 3. PCR was performed on selected samples and PCR products were sequenced on the ABI® 3130xl/3100 Genetic Analyzer 16-Capillary Array (Thermo Fisher Scientific Inc., Waltham, MA) using BigDye™ Terminator v1.1 cycle sequencing kit (Thermo Fisher Scientific Inc., Waltham, MA). Sequences were analysed using Sequencher (v5.4) and a BLAST [202] or BLAT [233] search.

Table 3. Integration validation primers.

HPV breakpoint	Human breakpoint	F primer (5'-3')	R primer (5'-3')	Study
HPV16:494	chr20:26341342	GGAACCAACCCAAATGTCCA	AGCCACTGTGTCCTGAAGAA	*Paper I*
HPV16:2987	chrX:145708231	GAGCTCCTGTTCACCAAACC*	CGAGGACAAGGAAAACGATG*	
HPV16:3631	chr19:55310043	AGCCGTGGTTCTCAACTAGG	TACATCCCGTACCCTCTTCC	
HPV16:7123	chr20:26357640	TCCCTTTCAGAGAGCACGTT	ACAAGCACATACAAGCACATACA	
HPV18:1561	chr7:74525628	CCTGTCATCCCAGCACTTTG	ACGGAGGCTATAGACAACGG	
HPV18:6528	chr7:74515883	CGAAGGCTGTGGAGAGAAGA	ACCCTGTGCCTTATGTAACCA	
HPV16:1073	chr10:28607045	ACGAAGCCAGTTAAAGGTAGAC	GGGATGCTATATCAGATGACGAG	*Paper III*
HPV16:2082	chr8:127912379	AAGTTGGTGGATGGGGAGAG	GAGCCTCCAAAATTGCGTAGT	
HPV16:3724	chr1:209432931	ACCCTGACAGCTGAGAGGTA	GACCCATACCAAAGCCGTC	
HPV16:4182	chr8:127881839	TGTTAGTTAGGCCGGTCTCG	CACAACATTACTGGCGTGCT	
HPV16:4220	chr8:127848840	AAGAGACTAGCTGGCATCCC	TACCCGACCCTGTTCCAATT	
HPV16:5123	chr10:28606989	ACTGGGTCATGTAGTGTTCGT	ATTGTGGAGACCCTGGAACT	
HPV16:5170	chr4:113525879	GGTCATTGTTGTGGGATTTGGA	TGCACCACCAAAAGGAATTGT	
HPV16:6815	chr3:129503742	CTTCACTGAGAGGAGCCGTC	TTGGCCTTCAATCCTGCTTG	
HPV18:168	chr6:136155580	CATGGGCAAGATTCAGGCTT	TGGGCACTGCTCCTACATAT	
HPV18:688	chr10:79344152	GCACTGGTAATGATCTCAGCC	ACCCTGTGTCTGTTGCATTT	
HPV18:936	chr4:128855029	GAGAATCTCCCAAAGCTGCTG	CCAGCCGTTACAACCCGTG	
HPV18:990	chr4:74546646	ACTCAACTCAGGTGACATCAAT	GCAACACTTGTGCATCATTGT	
HPV18:4676	chr9:129328942	CCGGCCTACTCCCATCTTAC	CTGACACTTGTTGGTAGGCC	
HPV18:4894	chr8:98848067	AGGGAGCACTGAGAAGTCAC	GCAGGCCTATGTAGACGGAT	
HPV18:4918	chr4:38110991	CAGGTCACGTGGGTAGAGAG	GCAGGCCTATGTAGACGGAT	
HPV18:5749	chr7:101884655	CCGCCAAAGGAGACAGACC	CACGGCCAATTTCCACTCC	

* Primers from [192].

3.7 Statistical analysis

In *Paper III*, statistical analyses were performed using the non-parametric Kruskal-Wallis test in R (v3.5.1). To confirm that the data did not follow normal distribution the Shapiro-Wilk test of normality was performed. A p-value of <0.05 was considered statistically significant.

3.8 Ethical aspects

The use of clinical patient samples in *Papers I* and *III* was approved by the Regional Committees for Medical and Health Research Ethics, Norway (2017/447). The use of sample material in *Paper II* was approved by the Medical Ethical Committee of the Vrije Universiteit University Medical Centre (VUmc) Amsterdam (2007/239). All clinical data was pseudonymised.

In addition, sequencing results for HPV16, HPV18 and HPV33 plasmids used in *Paper I* were not analysed due to third-party material rights.

3.9 Patent application

Together with Technology Transfer Office Inven2 (Oslo, Norway), we have filed a patent application on the TaME-seq method (Appendix 1: Patent application). The patent was filed in November 2018, and a preliminary report was received from the European Patent Office in May 2019. Patent application was optimised with new data that was sent in November 2019 to strengthen the protection of the method.

4 SUMMARY OF RESULTS

Table 4 summarises the main findings in *Papers I-III*.

Table 4. Main findings in *Papers I-III*.

Study	Sequencing statistics	Variation	APOBEC3	Integration
Paper I	• 154.8 million raw reads obtained • 47% of the reads mapped to HPV	• HPV genes had 0–28% sites with nucleotide variation	–	• Known integration sites validated • Novel integration sites found
Paper II	• 36/59 samples had >45% of the genome covered by minimum 100×	• Total number MNVs 1717, on average 48 MNVs per genome • 35 MNVs were recaptured in the follow-up samples • 1.67 times more non-synonymous substitutions	• 23% of the mutations C>T substitutions • Associated with APOBEC3 activity	–
Paper III	• 80/157 HPV16 and 51/75 HPV18 samples included in the analysis • 1.05 billion read pairs analysed	• 3747 MNVs in 131 samples • Increased number of variation in HPV16 NCR • For most genes in HPV16, dN/dS >1 and in HPV18 dN/dS ≈ 1	• APOBEC3-related C>T substitutions in HPV16 non-progressive and CIN2 categories • Similar APOBEC3 pattern not observed in HPV18 samples	• Integrations in 13% of HPV16 and 59% of HPV18 samples • Novel integration sites in HPV16 NCR • Integrations in HPV E1/E2 genes and in or close to human cancer-related genes

4.1 Paper I

We developed a next-generation sequencing approach named TaME-seq (tagmentation-assisted multiplex PCR enrichment sequencing) to characterise HPV genomic variability and chromosomal integration. To validate the method, HPV positive cervical cancer cell lines (n=4), HPV positive plasmids (n=3), and HPV16, 18, 31, 33 and 45 positive LBC samples (n=21) were analysed.

Totally 154.8 million raw reads were generated, and the mean sequencing depth per sample ranged from 303 to 273898. In total, 47% of the reads were mapped to HPV genomes, demonstrating the excellent target enrichment capacity of the TaME-seq method, resulting in less off-target reads and therefore reduced sequencing cost. The results showed considerable HPV genomic variability; up to 28% sites in each HPV gene had nucleotide variation. Most nucleotide variation were observed in one of the clinical samples, showing 21% variable sites (1641/7858 bases) throughout the HPV45 genome. The method confirmed previously reported chromosomal integration sites and HPV deletions in the HPV positive cell lines. In addition, novel integration sites were found in the CaSki cell line and in one clinical samples.

TaME-seq laboratory workflow is straightforward using standard laboratory procedures and sample preparation and sequencing are cost-effective. The method can easily be applied to large sample cohorts, representing an excellent choice for the characterisation of HPV genomic variation and chromosomal integration.

4.2 Paper II

The TaME-seq method was applied to samples from women with persistent HPV16 infection to assess HPV16 genomic variation at the minor variant level in multiple follow-up samples. Subset chosen for the study included women with either three (n=13) or four rounds (n=5) of persistent HPV16 infection. Participants supplied up to four samples, and median interval time between sampling was 48 (range 17–63) weeks. Cytology was not performed on the samples.

In total, 59 samples were processed and 61% (36/59) fulfilled the criteria for further analysis, which was >45% genome coverage by minimum 100×. Three infections were excluded from the analysis due to poor sequencing coverage, resulting in 15 infections that were followed over time. 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.

By using a >100× sequencing depth and a 1% variant frequency cut-off, a total of 1717 MNVs were detected in 36 samples. On average 48 (range 15–82) MNVs per genome were observed. Majority (67%) of MNVs were T>C substitutions, and the second-most abundant were C>T mutations (23%), latter being associated with the APOBEC3 activity. There were 1.67 times more

non-synonymous than synonymous mutations in the samples. 35 MNVs were recaptured in the follow-up samples from eleven women.

4.3 Paper III

Cervical samples positive for HPV16 and/or HPV18 were sequenced using the TaME-seq protocol. A total of 80 HPV16 positive samples and 51 HPV18 positive samples passed the mean sequencing depth criteria of 300× reads. Samples were categorised based on the HPV type and diagnostic category of non-progressive disease (HPV16 n=21, HPV18 n=12), CIN2 (HPV16 n=27, HPV18 n=9), CIN3/AIS (HPV16 n=27, HPV18 n=30) and cervical cancer (HPV16 n=5). In total, 1.05 billion read pairs were analysed and the samples had on average 77.7% of the genome covered by a minimum depth of 100×.

Overall, 3747 MNVs were found in the analysed 131 samples, showing similar numbers and frequencies of MNVs between the diagnostic categories and HPV types. Only the short NCR between E5 and L2 in HPV16 harboured considerably more variation in the diagnostic categories CIN2, CIN3/AIS and cancer. HPV16 showed predominantly more nonsynonymous variants ($dN/dS > 1$), while HPV18 genes had equal amounts of nonsynonymous and synonymous variants ($dN/dS \approx 1$) in most of the genes. APOBEC3-related C>T nucleotide substitutions were observed in HPV16 non-progressive samples and to a slightly less extent in HPV16 CIN2 samples. The same mutational patterns were not detected in HPV18 samples.

The integration frequency was higher in all HPV18 positive diagnostic categories compared to the HPV16 categories, with the proportion of samples with integration being 13% (10/80) for HPV16 and 59% (30/51) for HPV18 positive samples. Interestingly, one HPV16 positive cancer sample harboured two integration breakpoints in NCR, while in the HPV18 positive samples integration breakpoints were located in all HPV genomic regions except NCR. For HPV16 and HPV18 combined, a significant part of the integration breakpoints were observed in the HPV genes E1 or E2. In the human genome, integration breakpoints were detected in or close to cancer-related genes in all cancer samples, and in 65%, 38% and 34% of CIN3/AIS, CIN2 and non-progressive samples, respectively.

5 DISCUSSION

The overall aim of this thesis was to develop a novel NGS method to characterise HPV genomic variation and chromosomal integration, and to apply the method on clinical HPV samples to explore the genomic events, which contribute to HPV-induced carcinogenesis. Three specific objectives were outlined, and they will be addressed in this section. The main findings are discussed in respective papers, but in this sections the topics will be linked together and discussed in-depth. The chapter starts with a broader discussion of the methodology, which was an essential part of the thesis, followed by a discussion of the main findings.

5.1 Methodological considerations

5.1.1 Sample material

Different study designs were applied first to validate the method and further to characterise HPV genomic events both over time and across different HPV types and diagnostic groups. HPV genotype was determined in all samples before applying the TaME-seq method. As a result, incorrectly genotyped samples would lead to failure with TaME-seq since HPV-type specific primers are used for enrichment of the HPV genome. Most of the samples included in *Papers I-II* were previously characterised by Sanger sequencing or NGS, which gave us useful information on nucleotide sequence and possible variation when applying the TaME-seq method on clinical samples for the first time. Original DNA samples had been properly stored, mainly in -80°C, which is desired for long-term storage of DNA to maintain DNA quality and integrity [234].

For single samples, DNA concentration was measured to use a recommended input DNA amount for the Nextera tagmentation reaction. For some samples, the DNA concentration was too low for an optimal input in the Nextera reaction. Nevertheless, low amount of input material in Nextera tagmentation has been reported to result in good sequencing results [235]. Since the TaME-seq method targets HPV sequences, low viral load may be a possible cause of low or no sequencing yield. HPV viral load is shown to be higher in high-grade lesions compared to samples with normal cytology [236], therefore it could be assumed that samples with normal cytology or low-grade lesions would generally have lower sequencing yield.

5.1.2 Library preparation and NGS

For analysing rare genomic events, such as low-frequency variation and integration into the human genome, we could only use NGS technologies [170]. Several NGS approaches has been developed for investigating HPV genomes, but the methods are often limited to assessing either the genomic variation [84, 188] or integration [192, 193]. Some methods, including WGS [101, 106] and target capture technologies [102, 186, 237], can be used simultaneously for both purposes, but these methods have some limitations. Disadvantage of the WGS is low on-target (HPV) mapping leading to lower sequencing coverage unless sequencing throughput is increased, which in turn increases the cost [238]. Cost for sample preparation applying target capture technologies remain high and the laboratory workflows are time-consuming [239]. Table 5 summarises a selection of different NGS-based methods used in HPV genome research, their target applications and limitations.

Table 5. NGS-based methods used for different HPV research applications.

Method	Laboratory workflow	Research application			Limitations	Additional details
		Genomic variation	Integration	Genotyping		
TaME-seq	Nextera and HPV type-specific amplification	Yes	Yes	No	HPV-type specific approach	
Ion AmpliSeq HPV16 panel	47 overlapping amplicons	Yes	No	No	HPV-type specific approach, suitable only for variation analysis	[84]
Full-circle PCR	Long-range PCR followed by NGS library preparation	Yes	No	No	HPV-type specific approach, suitable only for variation analysis	[188]
TEN16 Assay	Nextera, blocking of the DNA 3'-ends, multiplex enrichment	No	Yes	No	Detects integrations in one DNA strand only	[192]
HIVID	Bead-based capture technology	Yes	Yes	Yes	Laborious, high reagent costs	[102, 237]
WGS	Traditional WGS workflow	Yes	Yes	Yes	High sequencing costs	[101, 106]

For TaME-seq, only up to 20 ng of DNA is needed, which is an advantage since a limited amount of clinical material is often available. Different commonly used polymerases were tested but QIAGEN Multiplex PCR Master Mix with HotStar *Taq* DNA Polymerase was chosen as the polymerase used in the TaME-seq workflow because it showed the best capacity to enrich HPV sequences and produced least off-target sequences. The TaME-seq method was proven to have an excellent target enrichment capacity, yielding on average in 47% of raw reads mapping to the target HPV reference genomes. Other NGS-based approaches report much lower HPV mapping ratios, varying between 0.001–35% [102, 187]. When HPV sequences are successfully enriched in the sample, sequencing yields in less off-target (human) reads, reducing the overall sequencing costs. Lastly, the method is based on PCR, and the HPV type-specific primers are designed to cover the HPV genomes evenly. Still, the sequencing coverage is uneven for certain genomic regions and some regions might even lack coverage, which may be caused by suboptimal primer design, performance or poor sequencing alignment.

NGS technologies are prone to errors that may lead to incorrect conclusions [195]. Most of the library preparation protocols, including TaME-seq, involves a PCR amplification step. PCR can introduce erroneous sequences into the pool of amplified DNA molecules. The efficiency with which PCR amplifies a sequence may vary between the sequences depending on factors including sequence composition and secondary structure [240]. A high GC content can reduce amplification efficiency, causing uneven amplification of different sequences [241]. Another source of uneven amplification in PCR is stochasticity. Stochastic amplification may have a significant impact on sequence representation when specific sequences are present at a very low copy number [240]. When synthesising a new DNA strand, DNA polymerase makes errors, including single nucleotide substitutions and short indels [242]. Polymerases have different error rates that also depend on experimental conditions [240]. The estimates of *Taq* polymerase fidelity vary, but recent results based on single-molecule sequencing report an error rate (per base per cycle) of 1.5×10^{-4} [243], being in line with earlier studies [244, 245]. The *Taq* polymerase lacks proofreading activity and may introduce bias, mainly A>G and T>C substitutions, during the amplification [243].

5.1.3 Sequencing analysis

To analyse the data produced by TaME-seq, an in-house analysis pipeline was developed. Different commonly used sequencing aligners [246] were tested to find out which one performed best for our purposes. HISAT2, a fast spliced aligner, originally developed for RNA-seq data [231], gave

the best sequence mapping results compared to other tested aligners. To further optimise the alignment process, use of an accurate reference genome is essential [195]. HPV genomes are known to vary and several variant lineage genomes are published [76], but for a less complex alignment process we chose to use HPV reference genomes [224] as the target genomes. Therefore, HPV genomes harbouring nucleotide variation [80, 81] could affect the sequence alignment by causing more mismatches in the alignment process. Finally, PCR duplicate removal was not applied because most of the reads would be lost before the subsequent analysis [247].

There are several freely available variant calling pipelines [248] that compare the sequences to the reference genome to call the non-reference bases as variants [207]. Since the HPV reference genomes alone do not reflect the diversity of HPV variant lineages, we decided to call MNVs based on the detected minor alleles, despite the reference base. Based on known error rates, a variant calling threshold must be set according to the study aims. When reducing number of false positive by applying trimming and filtering, number of false negative may increase [207]. If using a low threshold for variant calling, there is a chance to call false positive variants [166]. In *Paper I*, a threshold of 0.2% for variant calling was set to show the variant detection capacity of TaME-seq. In *Papers II* and *III*, clinical samples were analysed and it was crucial to minimise the risk of calling false positive variants. Therefore we increased the variant calling threshold to 1%. In addition, our variant calling pipeline was based on a stepwise evaluation of the MNVs in both F and R reactions to minimise the risk of false discoveries.

Several integration analysis tools are available for detection and characterisation of integration breakpoints [249-251], but none of them were optimal for our protocol. The tools were unable to report the exact integration breakpoints [249], to detect non-integrated viruses [250] or were not flexible to be included as part of other pipelines [251]. Integration analysis included in our pipeline was based on a two-step analysis to strengthen the findings from each individual analysis, which was essential especially for detecting rare integration events. Here, we applied filtering steps to exclude potential PCR duplicates, which could otherwise be reported as false positive findings. A real integration site could be missed if it was covered with too few supporting reads, but skipping the filtering step would have led to detection of multiple false positive integration sites. Validation of integration sites was performed by Sanger sequencing, which is still the most commonly used method to validate NGS results [252].

5.1.4 Statistical analysis

The Kruskal-Wallis test is a non-parametric test used to compare two or more independent groups. The choice of using a non-parametric test was based on non-normal distribution of the tested variables [253]. The significant result in a Kruskal-Wallis test indicates that there are group differences, but it does not indicate which groups differ [254]. The standard deviation in our data was large and the samples size was relative small, resulting mostly in non-significant statistical results. Bigger sample size would be needed to confirm the findings statistically [255].

5.2 Discussion of results

Traditionally, HPV genomic variation has been assessed to classify HPV into variant lineages and sublineages [76]. Today, several studies have used NGS for analysing HPV genomic variation, but it is still widely in use to categorise the variants into HPV variant lineages [120, 127, 256]. The HPV variant lineage classification is based on the major nucleotide in each genomic position, and ignores the minor nucleotide variation [120]. However, recent studies have reported a high intra-host HPV variability with low-frequency variants [123], which may be evidence of intra-host evolution and adaptation, being important for HPV survival and carcinogenicity. The high HPV genomic variability found in the samples is surprising because HPV is suggested to have a low evolutionary rate [65].

In line with the other studies showing high HPV variability, we reported on average 48 MNVs in early HPV16 infections (*Paper II*), 25–37 and 21–27 MNVs in HPV16 and HPV18 positive samples of different morphology, respectively (*Paper III*). The total amount of variation seem to be relatively stable during the different stages of infection. Interestingly, we detected the same MNVs in several follow-up samples with early HPV infection (*Paper II*), which might reveal new insight into the importance of specific MNVs for persistence of infection and carcinogenesis. Overall, we detected more non-synonymous than synonymous variation both in early and later stages of infection, indicating positive selection of the HPV genomes. While non-synonymous variants are usually eliminated, certain non-synonymous variants may be favoured if they increase the survival of the virus [141]. At the HPV gene level, HPV16 NCR harboured most overall variation and HPV16 E7 showed least non-synonymous variation. The NCR sequence is known to vary considerably [123, 127, 257], but it is unknown if the variation could contribute to HPV-induced carcinogenic process. Interestingly, HPV E7 is previously shown to be conserved, suggested to be critical for the

carcinogenesis [124]. Conservation of the E7 gene in precancers and cancers may be a result of clonal evolution, leading to loss of diversity and expansion of dominating variants that are important for the carcinogenic process [258].

It has been suggested that APOBEC3 may be an important source of mutagenesis in the HPV genomes [124], and APOBEC3-related mutation profiles have been reported in HPV16, HPV52 and HPV58 positive cervical samples [123, 134, 259]. To our knowledge, mutational signatures in HPV18 has not yet been studied. We showed distinct APOBEC3-related MNV profiles in HPV16 positive low-grade lesions, but not in early infections or high-grade lesions. This supports the suggestion that APOBEC3 becomes active in the course of infection trying to eliminate the virus [131], but at a more severe stage of disease, the virus may evade host restriction and the APOBEC3 mutagenesis is replaced by other mechanisms important for carcinogenesis. A recent study, showing similar results, suggests that infections with APOBEC3-related mutations are likely to be benign and associated with viral clearance [134]. Similarly in the early infections, with no distinct APOBEC3-related MNVs, it may be suggested that APOBEC3 has not become active yet as the infection is still at an early stage. APOBEC3-related mutational patterns were not found in HPV18 positive samples, suggesting separate mechanisms causing genomic variation in HPV18 genomes.

Other sequencing methods, e.g. Sanger sequencing, are not able to detect low-frequency variation but NGS has enabled deep genome analysis of HPV genomes [166]. Still, most recent studies interpret the HPV genomic variation as co-infections of different variant lineages [84] or they exclude the low-frequency variants from the analysis [124, 189]. In recent years, studies have reported within-host minor nucleotide variation [129, 188]. Only very recently, it has been more widely accepted to consider the HPV genome more dynamic, undergoing mutagenesis also during an infection. However, the studies reporting HPV genomic variation are careful when reporting low-frequency variation [134] and might use VAF thresholds far above the NGS detection limits [166]. Indeed, this project adds to a growing evidence base, suggesting a paradigm shift from a stable HPV genome to a HPV that evolves during the infection and adapts to its environment to ensure its survival.

We observed more integration events in HPV18 positive samples compared to HPV16 positive samples, which is in line with previously reported results [103, 115, 186]. The genomic location of the integration sites was determined in both HPV and human genomes. Interestingly, a large proportion of the integration breakpoints was detected in the E1 or E2 gene. E1 and E2 regulate

expression of the HPV oncogenes E6 and E7 and disruption in E1 or E2 may lead to overexpression of the oncogenes [104, 105]. Similar findings have been reported in several studies [102, 192, 260], confirming frequent disruption of the E1 and E2 genes. We also observed an increased number of integrations in or close to human cancer-related genes with increasing lesion severity, and we reported several integration breakpoints in known chromosomal hot-spot regions [111], supporting the hypothesis of a non-random distribution of integration sites in high-grade lesions of cancer cases [112]. Integration breakpoints detected in human oncogenes or genomic deletion in the E1 or E2 genes could be a sign of a more aggressive infection or could indicate poorer disease prognosis in a clinical context, even with normal cytology or low-grade lesions, requiring follow-up with shorter intervals.

HPV16 is the most predominant genotype found in SCC, while HPV18 is more often associated with ADC [13]. These two types are suggested to differ in their target cell specificity; HPV16 affects predominantly squamous cells, while HPV18 to a higher degree induces lesions in glandular cells [83]. HPV18 is suggested to promote a higher degree of genomic instability and progress faster from CIN3 to invasive cervical cancer than other HPV types [83, 261]. This may partly be attributable to the anatomic location of HPV18-related cervical cancers, which may be more difficult to detect, often locating higher up in the cervical canal [262]. However, considering the high number of integration events and different mutational patterns in HPV18 compared to HPV16, it may be likely that these HPV types utilise different mechanisms to infect the cervical cells and to induce cervical carcinogenesis. Certainly, previously reported results show different DNA methylation patterns [263], numbers of E6* splice variants [264], and mechanistic signatures of integrations [193] for HPV16 and HPV18, supporting our hypothesis of different molecular mechanisms to induce cancer for the two genotypes.

Finally, TaMe-seq revealed novel integration sites in well-characterised CaSki cell line, proving the capacity of the method for deeper analysis of cell lines used widely for HPV research [265, 266]. We could also verify that the MS751 cell line only contains HPV45 sequences [267], and that no HPV18 sequences are present [268]. Immortalised human cancer cell lines are widely used in biomedical research, being at the same time easy to manipulate and molecularly characterise. A detailed characterisation is fundamental before using cell lines [269]. TaME-seq is proven to be an excellent and cost-efficient tool for genomic characterisation of cervical cancer cell lines, verifying previously reported results and revising incorrect information.

5.3 Future research and implications of results

For future studies, sample size needs to be increased in the different diagnostic categories to explore more closely the genomic differences between the groups. TaME-seq covers several high-risk HPV types, currently including HPV16, HPV18, HPV31, HPV33 and HPV45. By comparing the genomic events between several HPV types, we can obtain valuable knowledge on molecular mechanisms of the specific types involved in carcinogenesis. Also, current results should be supplemented with follow-up samples stored at the biobank [219]. The samples were collected from women originally diagnosed with ASC-US/LSIL followed by a new sampling after a certain time. This setting would reveal HPV genomic changes within an infected individual, and verify if the findings have prognostic value for assessing the risk of developing cervical cancer. In addition to basic HPV research, the method can be used in epidemiological studies to study HPV diversity and geographic distribution, as well as in vaccine surveillance studies to investigate the effect of the vaccine on the prevailing HPV types. For developing new HPV vaccines, the method can reveal important details at the molecular level. Finally, TaME-seq can easily be applied to other viruses by changing the virus-specific primers.

The long-term aim of this project is to use TaME-seq for cervical cancer risk assessment. Persistent infection with a high-risk HPV type is necessary but not sufficient for development of cervical cancer [150], and other factors and molecular events influence whether the infection persist and progresses to cervical cancer [69]. Cervical cancer screening programmes aim for early detection of HPV and treatment of precancerous lesions, which is important to prevent the progression to cervical cancer [41]. At the same time, the current management of precancers involves some overtreatment and unnecessary follow-up of lesions that would otherwise have regressed spontaneously [63]. Therefore, new molecular approaches are needed to uncover HPV genomic events arising during the carcinogenic process, and which could be used for personalised cancer risk assessment. We also suggested that HPV16 and HPV18 may use different molecular mechanisms to induce cancer, suggesting exploration of different follow-up strategies for HPV16 and HPV18 positive patients.

HPV genomic variation at the minor nucleotide level is scarcely studied, but specific HPV variant lineages [80, 81] and nucleotide variants [124] are associated with higher risk of developing cervical lesions [124]. Increased or decreased HPV genomic variation at the whole genome level, in specific HPV genes or specific MNVs stratified by lesion severity could serve as prognostic markers for

cervical cancer development. Here, more research needs to be done to confirm if the level of HPV genomic variation differs between the diagnostics groups and if specific MNVs are associated with higher risk of developing precancer or cancer. While the association between HPV low-frequency genomic variation and cervical carcinogenesis still is unclear, use of HPV integration as prognostic marker for cervical cancer has been suggested in several studies [101, 163, 186], but the methods used for integration analysis remain either time-consuming or costly, hindering the implementation of such methods in routine diagnostics. With TaME-seq, integration breakpoints were detected within or close to cancer-related human genes and HPV genes where the integration is known to occur more frequently. Such integration breakpoints could serve as prognostic markers for more aggressive cases, requiring follow-up with shorter intervals to detect high-grade lesions before they develop to cancer.

NGS can be considered as an established technology for research applications across the life science field. It still remains challenging to store, analyse and translate the huge amount of genomic data into medical and biological context, requiring substantial bioinformatics expertise [270, 271]. Recently, use of NGS has expanded into clinical environments facilitating diagnostics and enabling more personalised treatments. However, using NGS both in research and in clinical diagnostics is accompanied by ethical and legal challenges. The General Data Protection Regulation (GDPR) was implemented in Europe in 2018 to protect the subjects' rights while allowing the processing of personal data, such as health data. The GDPR is also applied in research context, including clinical and translational research areas [272]. The GDPR provides individuals a right to access their personal data, including sequencing data [273]. However, it is still debated if and how incidental or additional findings, referring to findings that may have relevance for health but are unrelated to the research aim or the diagnostic test, from sequencing data should be reported back to the subject [274]. Many countries still have different policies concerning the return of incidental findings from sequencing data due to the ethical and practical complexity of the topic [273].

Finally, developing new products, e.g. diagnostic or screening tests, requires time and money. A patent is an instrument to secure investments and to cover development costs. It is also evidence of innovation, increasing competitive advantage and helping to achieve a greater share of the market [275]. While there is no guarantee that a patent application will be granted, the possibility to file a patent still indicates the novelty of the innovation, increasing the interest for the innovation in the scientific community and strengthening the chances to receive funding for future work.

6 CONCLUSIONS

This thesis aimed to characterise HPV genomic variation at the minor nucleotide level and chromosomal integration, and to explore HPV genomic events contributing to HPV-induced carcinogenesis.

We developed a unique deep-sequencing protocol TaME-seq for deep analysis of HPV genomic variation and integration; we have also filed a patent application for the method. By using the TaME-seq method, we could show that overall HPV genome variability is higher than assumed based on earlier estimates. Especially the low-frequency MNVs are predominantly detected in the HPV genomes. A high number of MNVs were found in all samples from early infections to cancer, with a noticeable part of HPV16 positive samples showing APOBEC3-related nucleotide substitutions. The findings also revealed previously known integrations sites in well-characterised cervical cancer cell lines and integration sites in several clinical samples, locating both in previously defined hot-spot and novel loci. HPV integration into the host genome is defined as a driving event in carcinogenesis and identification of integration sites can reveal important insight to the carcinogenic process.

The TaME-seq method could potentially be a valuable method for assessing the risk of developing cervical cancer. Current HPV tests used in cervical cancer screening are unable to predict the risk of persistence of HPV infection and progression to cervical cancer. An additional HPV test in cervical cancer screening would enable more personalised follow-up, improving detection of lesions with higher risk of progression and reducing unnecessary follow-up and treatment of women with minimal risk of developing high-grade lesions or cancer.

7 REFERENCES

1. de Martel C, Plummer M, Vignat J, Franceschi S. Worldwide burden of cancer attributable to HPV by site, country and HPV type. *Int J Cancer*. 2017;141:664-70.
2. Walboomers JM, Jacobs MV, Manos MM, Bosch FX, Kummer JA, Shah KV, Snijders PJ, Peto J, Meijer CJ, Munoz N. Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *J Pathol*. 1999;189:12-9.
3. zur Hausen H. Papillomaviruses in the causation of human cancers - a brief historical account. *Virology*. 2009;384:260-5.
4. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer*. 2015;136:E359-86.
5. Bosch FX, Broker TR, Forman D, Moscicki AB, Gillison ML, Doorbar J, Stern PL, Stanley M, Arbyn M, Poljak M, Cuzick J, Castle PE, Schiller JT, Markowitz LE, Fisher WA, Canfell K, Denny LA, Franco EL, Steben M, Kane MA, et al. Comprehensive control of human papillomavirus infections and related diseases. *Vaccine*. 2013;31 Suppl 6:G1-31.
6. Crosbie EJ, Einstein MH, Franceschi S, Kitchener HC. Human papillomavirus and cervical cancer. *The Lancet*. 2013;382:889-99.
7. Bosch FX, de Sanjose S. Chapter 1: Human papillomavirus and cervical cancer--burden and assessment of causality. *J Natl Cancer Inst Monogr*. 2003:3-13.
8. Moscicki AB, Schiffman M, Burchell A, Albero G, Giuliano AR, Goodman MT, Kjaer SK, Palefsky J. Updating the natural history of human papillomavirus and anogenital cancers. *Vaccine*. 2012;30 Suppl 5:F24-33.
9. Bzhalava D, Eklund C, Dillner J. International standardization and classification of human papillomavirus types. *Virology*. 2015;476:341-4.
10. The International Human Papillomavirus (HPV) Reference Center. Available on https://www.hpvcenter.se/ on January 13, 2020.
11. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Biological agents. Volume 100 B. A review of human carcinogens. *IARC Monogr Eval Carcinog Risks Hum*. 2012;100:1-441.
12. Arbyn M, Tommasino M, Depuydt C, Dillner J. Are 20 human papillomavirus types causing cervical cancer? *J Pathol*. 2014;234:431-5.
13. de Sanjose S, Quint WGV, Alemany L, Geraets DT, Klaustermeier JE, Lloveras B, Tous S, Felix A, Bravo LE, Shin H-R, Vallejos CS, de Ruiz PA, Lima MA, Guimera N, Clavero O, Alejo M, Llombart-Bosch A, Cheng-Yang C, Tatti SA, Kasamatsu E, et al. Human papillomavirus genotype attribution in invasive cervical cancer: a retrospective cross-sectional worldwide study. *The Lancet Oncology*. 2010;11:1048-56.
14. Egawa N, Doorbar J. The low-risk papillomaviruses. *Virus Res*. 2017;231:119-27.
15. Schiffman M, Doorbar J, Wentzensen N, de Sanjose S, Fakhry C, Monk BJ, Stanley MA, Franceschi S. Carcinogenic human papillomavirus infection. *Nat Rev Dis Primers*. 2016;2:16086.
16. Plummer M, Schiffman M, Castle PE, Maucort-Boulch D, Wheeler CM, Group A. A 2-year prospective study of human papillomavirus persistence among women with a cytological diagnosis of atypical squamous cells of undetermined significance or low-grade squamous intraepithelial lesion. *J Infect Dis*. 2007;195:1582-9.
17. Schiffman M, Wentzensen N. Human papillomavirus infection and the multistage carcinogenesis of cervical cancer. *Cancer Epidemiol Biomarkers Prev*. 2013;22:553-60.
18. Schiffman M, Castle PE, Jeronimo J, Rodriguez AC, Wacholder S. Human papillomavirus and cervical cancer. *The Lancet*. 2007;370:890-907.

19. Cancer Registry of Norway. Livmorhalsprogrammet, årsrapport 2016. Oslo, Norway: Cancer Registry of Norway, 2018.
20. McCredie MRE, Sharples KJ, Paul C, Baranyai J, Medley G, Jones RW, Skegg DCG. Natural history of cervical neoplasia and risk of invasive cancer in women with cervical intraepithelial neoplasia 3: a retrospective cohort study. *The Lancet Oncology*. 2008;9:425-34.
21. Motamedi M, Bohmer G, Neumann HH, von Wasielewski R. CIN III lesions and regression: retrospective analysis of 635 cases. *BMC Infect Dis*. 2015;15:541.
22. Schiffman M, Wentzensen N, Wacholder S, Kinney W, Gage JC, Castle PE. Human papillomavirus testing in the prevention of cervical cancer. *J Natl Cancer Inst*. 2011;103:368-83.
23. Doorbar J, Quint W, Banks L, Bravo IG, Stoler M, Broker TR, Stanley MA. The biology and life-cycle of human papillomaviruses. *Vaccine*. 2012;30 Suppl 5:F55-70.
24. Sellors JW, R S. Colposcopy and treatment of cervical intraepithelial neoplasia: a beginners' manual. Lyon, France: IARC Press, 2003.
25. Silverberg SG, Ioffe OB. Pathology of cervical cancer. *Cancer J*. 2003;9:335-47.
26. Cancer Institute NSW. Available on https://www.cancer.nsw.gov.au/cervical-screening-nsw/your-cervical-screen-results/an-unsatisfactory-result on February 18, 2020.
27. Adegoke O, Kulasingam S, Virnig B. Cervical cancer trends in the United States: a 35-year population-based analysis. *Journal of women's health (2002)*. 2012;21:1031-7.
28. van der Horst J, Siebers AG, Bulten J, Massuger LF, de Kok IM. Increasing incidence of invasive and in situ cervical adenocarcinoma in the Netherlands during 2004-2013. *Cancer Med*. 2017;6:416-23.
29. Alfsen GC, Thoresen SO, Kristensen GB, Skovlund E, Abeler VM. Histopathologic subtyping of cervical adenocarcinoma reveals increasing incidence rates of endometrioid tumors in all age groups: a population based study with review of all nonsquamous cervical carcinomas in Norway from 1966 to 1970, 1976 to 1980, and 1986 to 1990. *Cancer*. 2000;89:1291-9.
30. Joura EA, Giuliano AR, Iversen OE, Bouchard C, Mao C, Mehlsen J, Moreira ED, Jr., Ngan Y, Petersen LK, Lazcano-Ponce E, Pitisuttithum P, Restrepo JA, Stuart G, Woelber L, Yang YC, Cuzick J, Garland SM, Huh W, Kjaer SK, Bautista OM, et al. A 9-valent HPV vaccine against infection and intraepithelial neoplasia in women. *N Engl J Med*. 2015;372:711-23.
31. Lehtinen M, Dillner J. Clinical trials of human papillomavirus vaccines and beyond. *Nat Rev Clin Oncol*. 2013;10:400-10.
32. Stanley M, Pinto LA, Trimble C. Human papillomavirus vaccines--immune responses. *Vaccine*. 2012;30 Suppl 5:F83-7.
33. World Health Organization. Human papillomavirus vaccines: WHO position paper, May 2017-Recommendations. *Vaccine*. 2017;35:5753-5.
34. Schiller JT, Castellsague X, Garland SM. A review of clinical trials of human papillomavirus prophylactic vaccines. *Vaccine*. 2012;30 Suppl 5:F123-38.
35. Munoz N, Kjaer SK, Sigurdsson K, Iversen OE, Hernandez-Avila M, Wheeler CM, Perez G, Brown DR, Koutsky LA, Tay EH, Garcia PJ, Ault KA, Garland SM, Leodolter S, Olsson SE, Tang GW, Ferris DG, Paavonen J, Steben M, Bosch FX, et al. Impact of human papillomavirus (HPV)-6/11/16/18 vaccine on all HPV-associated genital diseases in young women. *J Natl Cancer Inst*. 2010;102:325-39.
36. Castellsague X, Munoz N, Pitisuttithum P, Ferris D, Monsonego J, Ault K, Luna J, Myers E, Mallary S, Bautista OM, Bryan J, Vuocolo S, Haupt RM, Saah A. End-of-study safety, immunogenicity, and efficacy of quadrivalent HPV (types 6, 11, 16, 18) recombinant vaccine in adult women 24-45 years of age. *Br J Cancer*. 2011;105:28-37.
37. Palefsky JM, Giuliano AR, Goldstone S, Moreira ED, Jr., Aranda C, Jessen H, Hillman R, Ferris D, Coutlee F, Stoler MH, Marshall JB, Radley D, Vuocolo S, Haupt RM, Guris D,

Garner EI. HPV vaccine against anal HPV infection and anal intraepithelial neoplasia. *N Engl J Med*. 2011;365:1576-85.
38. Brisson M, Bénard É, Drolet M, Bogaards JA, Baussano I, Vänskä S, Jit M, Boily M-C, Smith MA, Berkhof J, Canfell K, Chesson HW, Burger EA, Choi YH, De Blasio BF, De Vlas SJ, Guzzetta G, Hontelez JAC, Horn J, Jepsen MR, et al. Population-level impact, herd immunity, and elimination after human papillomavirus vaccination: a systematic review and meta-analysis of predictions from transmission-dynamic models. *The Lancet Public Health*. 2016;1:e8-e17.
39. Folkehelseinstituttet. Vaksine mot humant papillomavirus (HPV): Rapport fra arbeidsgruppe nedsatt av Nasjonalt folkehelseinstitutt for å vurdere om HPV-vaksine til gutter skal tilbys i program i Norge. Oslo, Norway: Folkehelseinstituttet, 2016.
40. Gallagher KE, LaMontagne DS, Watson-Jones D. Status of HPV vaccine introduction and barriers to country uptake. *Vaccine*. 2018;36:4761-7.
41. World Health Organization. IARC handbooks of cancer prevention. Volume 10: Cervix cancer screening: IARC Press, International Agency for Research on Cancer, 2005.
42. Wilson JMG, Jungner G, World Health O. Principles and practice of screening for disease / J. M. G. Wilson, G. Jungner. Geneva: World Health Organization; 1968.
43. Kenyon S, Sweeney BJ, Happel J, Marchilli GE, Weinstein B, Schneider D. Comparison of BD Surepath and ThinPrep Pap systems in the processing of mucus-rich specimens. *Cancer Cytopathol*. 2010;118:244-9.
44. Cuzick J, Ahmad AS, Austin J, Cadman L, Ho L, Terry G, Kleeman M, Ashdown-Barr L, Lyons D, Stoler M, Szarewski A. A comparison of different human papillomavirus tests in PreservCyt versus SurePath in a referral population—PREDICTORS 4. *J Clin Virol*. 2016;82:145-51.
45. Burd EM. Human Papillomavirus Laboratory Testing: the Changing Paradigm. *Clin Microbiol Rev*. 2016;29:291-319.
46. Arbyn M, Ronco G, Anttila A, Meijer CJ, Poljak M, Ogilvie G, Koliopoulos G, Naucler P, Sankaranarayanan R, Peto J. Evidence regarding human papillomavirus testing in secondary prevention of cervical cancer. *Vaccine*. 2012;30 Suppl 5:F88-99.
47. Gage JC, Schiffman M, Katki HA, Castle PE, Fetterman B, Wentzensen N, Poitras NE, Lorey T, Cheung LC, Kinney WK. Reassurance against future risk of precancer and cancer conferred by a negative human papillomavirus test. *J Natl Cancer Inst*. 2014;106.
48. Ronco G, Dillner J, Elfström KM, Tunesi S, Snijders PJF, Arbyn M, Kitchener H, Segnan N, Gilham C, Giorgi-Rossi P, Berkhof J, Peto J, Meijer CJLM. Efficacy of HPV-based screening for prevention of invasive cervical cancer: follow-up of four European randomised controlled trials. *The Lancet*. 2014;383:524-32.
49. Mayrand MH, Duarte-Franco E, Rodrigues I, Walter SD, Hanley J, Ferenczy A, Ratnam S, Coutlee F, Franco EL. Human papillomavirus DNA versus Papanicolaou screening tests for cervical cancer. *N Engl J Med*. 2007;357:1579-88.
50. Nygård M, Andreassen T, Berland J, Hagen B, Hagmar B, Iversen O-E, Juvkam K-H, Kristiansen IS, Lønnberg SV, Sørby SW, Vintermyr OK, Aarseth H-P. HPV-test i primærscreening mot livmorhalskreft. Kontrollert implementering og evaluering av forbedret helsetjeneste. Oslo, Norway: Helsedirektoratet, 2013.
51. The Norwegian Cervical Cancer Screening Programme. HPV i primærscreening. 2019. Available on https://www.kreftregisteret.no/screening/livmorhalsprogrammet/Helsepersonell/screeningstrategi-og-nasjonale-retningslinjer/HPV-i-primarscreening/ on March 30, 2020.
52. Solomon D, Davey D, Kurman R, et al. The 2001 bethesda system: Terminology for reporting results of cervical cytology. *JAMA*. 2002;287:2114-9.
53. Richart RM. Cervical intraepithelial neoplasia. *Pathol Annu*. 1973;8:301-28.

54. Buckley CH, Butler EB, Fox H. Cervical intraepithelial neoplasia. *J Clin Pathol*. 1982;35:1-13.
55. Zaino RJ. Symposium part I: adenocarcinoma in situ, glandular dysplasia, and early invasive adenocarcinoma of the uterine cervix. *Int J Gynecol Pathol*. 2002;21:314-26.
56. Gage JC, Hanson VW, Abbey K, Dippery S, Gardner S, Kubota J, Schiffman M, Solomon D, Jeronimo J, Group fTALTS. Number of Cervical Biopsies and Sensitivity of Colposcopy. *Obstet Gynecol*. 2006;108:264-72.
57. Baasland I, Hagen B, Vogt C, Valla M, Romundstad PR. Colposcopy and additive diagnostic value of biopsies from colposcopy-negative areas to detect cervical dysplasia. *Acta Obstet Gynecol Scand*. 2016;95:1258-63.
58. Herbert A, Bergeron C, Wiener H, Schenck U, Klinkhamer P, Bulten J, Arbyn M. European guidelines for quality assurance in cervical cancer screening: recommendations for cervical cytology terminology. *Cytopathology*. 2007;18:213-9.
59. The Norwegian Cervical Cancer Screening Programme. Kvalitetsmanual for Livmorhalsprogrammet. 2019. Available on https://www.kreftregisteret.no/screening/livmorhalsprogrammet/Helsepersonell/Faglig-Radgivningsgruppe/kvalitetsmanual2/ on March 16, 2020.
60. Rådgivningsgruppen for Masseundersøkelsen mot livmorhalskreft. Quality assurance manual: Cervical Cancer Screening Programme. Oslo, Norway: Cancer Registry of Norway, 2014.
61. Khan MJ, Smith-McCune KK. Treatment of cervical precancers: back to basics. *Obstet Gynecol*. 2014;123:1339-43.
62. Rodriguez AC, Schiffman M, Herrero R, Hildesheim A, Bratti C, Sherman ME, Solomon D, Guillen D, Alfaro M, Morales J, Hutchinson M, Katki H, Cheung L, Wacholder S, Burk RD. Longitudinal study of human papillomavirus persistence and cervical intraepithelial neoplasia grade 2/3: critical role of duration of infection. *J Natl Cancer Inst*. 2010;102:315-24.
63. Petry KU. Management options for cervical intraepithelial neoplasia. *Best Pract Res Clin Obstet Gynaecol*. 2011;25:641-51.
64. de Villiers EM, Fauquet C, Broker TR, Bernard HU, zur Hausen H. Classification of papillomaviruses. *Virology*. 2004;324:17-27.
65. Bravo IG, Felez-Sanchez M. Papillomaviruses: Viral evolution, cancer and evolutionary medicine. *Evol Med Public Health*. 2015;2015:32-51.
66. Doorbar J, Egawa N, Griffin H, Kranjec C, Murakami I. Human papillomavirus molecular biology and disease association. *Rev Med Virol*. 2015;25 Suppl 1:2-23.
67. Doorbar J. The E4 protein; structure, function and patterns of expression. *Virology*. 2013;445:80-98.
68. DiMaio D, Petti LM. The E5 proteins. *Virology*. 2013;445:99-114.
69. de Sanjose S, Brotons M, Pavon MA. The natural history of human papillomavirus infection. *Best Pract Res Clin Obstet Gynaecol*. 2018;47:2-13.
70. Maki H, Fujikawa-Adachi K, Yoshie O. Evidence for a promoter-like activity in the short non-coding region of human papillomaviruses. *J Gen Virol*. 1996;77:453-8.
71. Mandal P, Bhattacharjee B, Das Ghosh D, Mondal NR, Roy Chowdhury R, Roy S, Sengupta S. Differential expression of HPV16 L2 gene in cervical cancers harboring episomal HPV16 genomes: influence of synonymous and non-coding region variations. *PLoS One*. 2013;8:e65647.
72. Smith B, Chen Z, Reimers L, van Doorslaer K, Schiffman M, Desalle R, Herrero R, Yu K, Wacholder S, Wang T, Burk RD. Sequence imputation of HPV16 genomes for genetic association studies. *PLoS One*. 2011;6:e21375.
73. Van Doorslaer K. Evolution of the papillomaviridae. *Virology*. 2013;445:11-20.

74. Gottschling M, Stamatakis A, Nindl I, Stockfleth E, Alonso A, Bravo IG. Multiple evolutionary mechanisms drive papillomavirus diversification. *Mol Biol Evol*. 2007;24:1242-58.
75. Munoz N, Bosch FX, de Sanjose S, Herrero R, Castellsague X, Shah KV, Snijders PJ, Meijer CJ. Epidemiologic classification of human papillomavirus types associated with cervical cancer. *N Engl J Med*. 2003;348:518-27.
76. Burk RD, Harari A, Chen Z. Human papillomavirus genome variants. *Virology*. 2013;445:232-43.
77. Chen Z, Schiffman M, Herrero R, Desalle R, Anastos K, Segondy M, Sahasrabuddhe VV, Gravitt PE, Hsing AW, Burk RD. Evolution and taxonomic classification of human papillomavirus 16 (HPV16)-related variant genomes: HPV31, HPV33, HPV35, HPV52, HPV58 and HPV67. *PLoS One*. 2011;6:e20183.
78. Burk RD, Chen Z, Van Doorslaer K. Human Papillomaviruses: Genetic Basis of Carcinogenicity. *Public Health Genomics*. 2009;12:281-90.
79. Bruni L, Diaz M, Castellsague X, Ferrer E, Bosch FX, de Sanjose S. Cervical human papillomavirus prevalence in 5 continents: meta-analysis of 1 million women with normal cytological findings. *J Infect Dis*. 2010;202:1789-99.
80. Clifford GM, Tenet V, Georges D, Alemany L, Pavon MA, Chen Z, Yeager M, Cullen M, Boland JF, Bass S, Steinberg M, Raine-Bennett T, Lorey T, Wentzensen N, Walker J, Zuna R, Schiffman M, Mirabello L. Human papillomavirus 16 sub-lineage dispersal and cervical cancer risk worldwide: Whole viral genome sequences from 7116 HPV16-positive women. *Papillomavirus Res*. 2019;7:67-74.
81. Chen AA, Gheit T, Franceschi S, Tommasino M, Clifford GM. Human Papillomavirus 18 Genetic Variation and Cervical Cancer Risk Worldwide. *J Virol*. 2015;89:10680-7.
82. Bzhalava D, Guan P, Franceschi S, Dillner J, Clifford G. A systematic review of the prevalence of mucosal and cutaneous human papillomavirus types. *Virology*. 2013;445:224-31.
83. Tjalma WA, Fiander A, Reich O, Powell N, Nowakowski AM, Kirschner B, Koiss R, O'Leary J, Joura EA, Rosenlund M, Colau B, Schledermann D, Kukk K, Damaskou V, Repanti M, Vladareanu R, Kolomiets L, Savicheva A, Shipitsyna E, Ordi J, et al. Differences in human papillomavirus type distribution in high-grade cervical intraepithelial neoplasia and invasive cervical cancer in Europe. *Int J Cancer*. 2013;132:854-67.
84. 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, et al. Deep sequencing of HPV16 genomes: A new high-throughput tool for exploring the carcinogenicity and natural history of HPV16 infection. *Papillomavirus Res*. 2015;1:3-11.
85. Cornet I, Gheit T, Iannacone MR, Vignat J, Sylla BS, Del Mistro A, Franceschi S, Tommasino M, Clifford GM. HPV16 genetic variation and the development of cervical cancer worldwide. *Br J Cancer*. 2013;108:240-4.
86. Chan PK, Zhang C, Park JS, Smith-McCune KK, Palefsky JM, Giovannelli L, Coutlee F, Hibbitts S, Konno R, Settheetham-Ishida W, Chu TY, Ferrera A, Alejandra Picconi M, De Marco F, Woo YL, Raiol T, Pina-Sanchez P, Bae JH, Wong MC, Chirenje MZ, et al. Geographical distribution and oncogenic risk association of human papillomavirus type 58 E6 and E7 sequence variations. *Int J Cancer*. 2013;132:2528-36.
87. Buck CB, Cheng N, Thompson CD, Lowy DR, Steven AC, Schiller JT, Trus BL. Arrangement of L2 within the papillomavirus capsid. *J Virol*. 2008;82:5190-7.
88. Stanley MA. Epithelial cell responses to infection with human papillomavirus. *Clin Microbiol Rev*. 2012;25:215-22.
89. Tomaic V. Functional Roles of E6 and E7 Oncoproteins in HPV-Induced Malignancies at Diverse Anatomical Sites. *Cancers (Basel)*. 2016;8.

90. Doorbar J. Molecular biology of human papillomavirus infection and cervical cancer. *Clin Sci (Lond)*. 2006;110:525-41.
91. McBride AA. Oncogenic human papillomaviruses. *Philos Trans R Soc Lond B Biol Sci*. 2017;372.
92. Roman A, Munger K. The papillomavirus E7 proteins. *Virology*. 2013;445:138-68.
93. Vande Pol SB, Klingelhutz AJ. Papillomavirus E6 oncoproteins. *Virology*. 2013;445:115-37.
94. Munger K, Baldwin A, Edwards KM, Hayakawa H, Nguyen CL, Owens M, Grace M, Huh K. Mechanisms of human papillomavirus-induced oncogenesis. *J Virol*. 2004;78:11451-60.
95. Hoppe-Seyler K, Bossler F, Braun JA, Herrmann AL, Hoppe-Seyler F. The HPV E6/E7 Oncogenes: Key Factors for Viral Carcinogenesis and Therapeutic Targets. *Trends Microbiol*. 2018;26:158-68.
96. Frazer IH. Interaction of human papillomaviruses with the host immune system: a well evolved relationship. *Virology*. 2009;384:410-4.
97. zur Hausen H. Papillomaviruses and cancer: from basic studies to clinical application. *Nat Rev Cancer*. 2002;2:342-50.
98. Pett M, Coleman N. Integration of high-risk human papillomavirus: a key event in cervical carcinogenesis? *J Pathol*. 2007;212:356-67.
99. McBride AA, Warburton A. The role of integration in oncogenic progression of HPV-associated cancers. *PLoS Pathog*. 2017;13:e1006211.
100. Bodelon C, Untereiner ME, Machiela MJ, Vinokurova S, Wentzensen N. Genomic characterization of viral integration sites in HPV-related cancers. *Int J Cancer*. 2016;139:2001-11.
101. Huang J, Qian Z, Gong Y, Wang Y, Guan Y, Han Y, Yi X, Huang W, Ji L, Xu J, Su M, Yuan Q, Cui S, Zhang J, Bao C, Liu W, Chen X, Zhang M, Gao X, Wu R, et al. Comprehensive genomic variation profiling of cervical intraepithelial neoplasia and cervical cancer identifies potential targets for cervical cancer early warning. *J Med Genet*. 2019;56:186-94.
102. 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, et al. Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism. *Nat Genet*. 2015;47:158-63.
103. Vinokurova S, Wentzensen N, Kraus I, Klaes R, Driesch C, Melsheimer P, Kisseljov F, Durst M, Schneider A, von Knebel Doeberitz M. Type-dependent integration frequency of human papillomavirus genomes in cervical lesions. *Cancer Res*. 2008;68:307-13.
104. Jeon S, Allen-Hoffmann BL, Lambert PF. Integration of human papillomavirus type 16 into the human genome correlates with a selective growth advantage of cells. *J Virol*. 1995;69:2989-97.
105. Ziegert C, Wentzensen N, Vinokurova S, Kisseljov F, Einenkel J, Hoeckel M, von Knebel Doeberitz M. A comprehensive analysis of HPV integration loci in anogenital lesions combining transcript and genome-based amplification techniques. *Oncogene*. 2003;22:3977-84.
106. Akagi K, Li J, Broutian TR, Padilla-Nash H, Xiao W, Jiang B, Rocco JW, Teknos TN, Kumar B, Wangsa D, He D, Ried T, Symer DE, Gillison ML. Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. *Genome Res*. 2014;24:185-99.
107. Kim K, Garner-Hamrick PA, Fisher C, Lee D, Lambert PF. Methylation Patterns of Papillomavirus DNA, Its Influence on E2 Function, and Implications in Viral Infection. *J Virol*. 2003;77:12450-9.
108. Chaiwongkot A, Vinokurova S, Pientong C, Ekalaksananan T, Kongyingyoes B, Kleebkaow P, Chumworathayi B, Patarapadungkit N, Reuschenbach M, von Knebel

Doeberitz M. Differential methylation of E2 binding sites in episomal and integrated HPV 16 genomes in preinvasive and invasive cervical lesions. *Int J Cancer*. 2013;132:2087-94.
109. Peter M, Stransky N, Couturier J, Hupe P, Barillot E, de Cremoux P, Cottu P, Radvanyi F, Sastre-Garau X. Frequent genomic structural alterations at HPV insertion sites in cervical carcinoma. *J Pathol*. 2010;221:320-30.
110. Christiansen IK, Sandve GK, Schmitz M, Durst M, Hovig E. Transcriptionally active regions are the preferred targets for chromosomal HPV integration in cervical carcinogenesis. *PLoS One*. 2015;10:e0119566.
111. Kraus I, Driesch C, Vinokurova S, Hovig E, Schneider A, von Knebel Doeberitz M, Durst M. The Majority of Viral-Cellular Fusion Transcripts in Cervical Carcinomas Cotranscribe Cellular Sequences of Known or Predicted Genes. *Cancer Res*. 2008;68:2514-22.
112. Schmitz M, Driesch C, Jansen L, Runnebaum IB, Durst M. Non-random integration of the HPV genome in cervical cancer. *PLoS One*. 2012;7:e39632.
113. Thorland EC, Myers SL, Gostout BS, Smith DI. Common fragile sites are preferential targets for HPV16 integrations in cervical tumors. *Oncogene*. 2003;22:1225-37.
114. Parfenov M, Pedamallu CS, Gehlenborg N, Freeman SS, Danilova L, Bristow CA, Lee S, Hadjipanayis AG, Ivanova EV, Wilkerson MD, Protopopov A, Yang L, Seth S, Song X, Tang J, Ren X, Zhang J, Pantazi A, Santoso N, Xu AW, et al. Characterization of HPV and host genome interactions in primary head and neck cancers. *Proc Natl Acad Sci U S A*. 2014;111:15544-9.
115. Cancer Genome Atlas Research N, Albert Einstein College of M, Analytical Biological S, Barretos Cancer H, Baylor College of M, Beckman Research Institute of City of H, Buck Institute for Research on A, Canada's Michael Smith Genome Sciences C, Harvard Medical S, Helen FGCC, Research Institute at Christiana Care Health S, HudsonAlpha Institute for B, Ilsbio LLC, Indiana University School of M, Institute of Human V, Institute for Systems B, International Genomics C, Leidos B, Massachusetts General H, McDonnell Genome Institute at Washington U, et al. Integrated genomic and molecular characterization of cervical cancer. *Nature*. 2017;543:378-84.
116. Cullen AP, Reid R, Campion M, Lorincz AT. Analysis of the physical state of different human papillomavirus DNAs in intraepithelial and invasive cervical neoplasm. *J Virol*. 1991;65:606-12.
117. Cheung JL, Cheung TH, Tang JW, Chan PK. Increase of integration events and infection loads of human papillomavirus type 52 with lesion severity from low-grade cervical lesion to invasive cancer. *J Clin Microbiol*. 2008;46:1356-62.
118. Ho CM, Chien TY, Huang SH, Lee BH, Chang SF. Integrated human papillomavirus types 52 and 58 are infrequently found in cervical cancer, and high viral loads predict risk of cervical cancer. *Gynecol Oncol*. 2006;102:54-60.
119. Marongiu L, Godi A, Parry JV, Beddows S. Human Papillomavirus 16, 18, 31 and 45 viral load, integration and methylation status stratified by cervical disease stage. *BMC Cancer*. 2014;14:384.
120. Mirabello L, Yeager M, Cullen M, Boland JF, Chen Z, Wentzensen N, Zhang X, Yu K, Yang Q, Mitchell J, Roberson D, Bass S, Xiao Y, Burdett L, Raine-Bennett T, Lorey T, Castle PE, Burk RD, Schiffman M. HPV16 Sublineage Associations With Histology-Specific Cancer Risk Using HPV Whole-Genome Sequences in 3200 Women. *J Natl Cancer Inst*. 2016;108.
121. Nicolas-Parraga S, Alemany L, de Sanjose S, Bosch FX, Bravo IG, Ris Hpv TT, groups HVs. Differential HPV16 variant distribution in squamous cell carcinoma, adenocarcinoma and adenosquamous cell carcinoma. *Int J Cancer*. 2017;140:2092-100.
122. Xu HH, Zheng LZ, Lin AF, Dong SS, Chai ZY, Yan WH. Human papillomavirus (HPV) 18 genetic variants and cervical cancer risk in Taizhou area, China. *Gene*. 2018;647:192-7.

123. Hirose Y, Onuki M, Tenjimbayashi Y, Mori S, Ishii Y, Takeuchi T, Tasaka N, Satoh T, Morisada T, Iwata T, Miyamoto S, Matsumoto K, Sekizawa A, Kukimoto I. Within-Host Variations of Human Papillomavirus Reveal APOBEC-Signature Mutagenesis in the Viral Genome. *J Virol*. 2018.
124. Mirabello L, Yeager M, Yu K, Clifford GM, Xiao Y, Zhu B, Cullen M, Boland JF, Wentzensen N, Nelson CW, Raine-Bennett T, Chen Z, Bass S, Song L, Yang Q, Steinberg M, Burdett L, Dean M, Roberson D, Mitchell J, et al. HPV16 E7 Genetic Conservation Is Critical to Carcinogenesis. *Cell*. 2017;170:1164-74 e6.
125. van der Weele P, Meijer C, King AJ. Whole-Genome Sequencing and Variant Analysis of Human Papillomavirus 16 Infections. *J Virol*. 2017;91.
126. van der Weele P, Meijer C, King AJ. High Whole-Genome Sequence Diversity of Human Papillomavirus Type 18 Isolates. *Viruses*. 2018;10.
127. Arroyo-Muhr LS, Lagheden C, Hultin E, Eklund C, Adami HO, Dillner J, Sundstrom K. Human papillomavirus type 16 genomic variation in women with subsequent in situ or invasive cervical cancer: prospective population-based study. *Br J Cancer*. 2018;119:1163-8.
128. Dube Mandishora RS, Gjotterud KS, Lagstrom S, Stray-Pedersen B, Duri K, Chin'ombe N, Nygard M, Christiansen IK, Ambur OH, Chirenje MZ, Rounge TB. Intra-host sequence variability in human papillomavirus. *Papillomavirus Res*. 2018.
129. de Oliveira CM, Bravo IG, Santiago e Souza NC, Genta ML, Fregnani JH, Tacla M, Carvalho JP, Longatto-Filho A, Levi JE. High-level of viral genomic diversity in cervical cancers: A Brazilian study on human papillomavirus type 16. *Infect Genet Evol*. 2015;34:44-51.
130. Vartanian JP, Guetard D, Henry M, Wain-Hobson S. Evidence for editing of human papillomavirus DNA by APOBEC3 in benign and precancerous lesions. *Science*. 2008;320:230-3.
131. Warren CJ, Xu T, Guo K, Griffin LM, Westrich JA, Lee D, Lambert PF, Santiago ML, Pyeon D. APOBEC3A functions as a restriction factor of human papillomavirus. *J Virol*. 2015;89:688-702.
132. Burns MB, Lackey L, Carpenter MA, Rathore A, Land AM, Leonard B, Refsland EW, Kotandeniya D, Tretyakova N, Nikas JB, Yee D, Temiz NA, Donohue DE, McDougle RM, Brown WL, Law EK, Harris RS. APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature*. 2013;494:366-70.
133. Harris RS, Dudley JP. APOBECs and virus restriction. *Virology*. 2015;479-480:131-45.
134. Zhu B, Xiao Y, Yeager M, Clifford G, Wentzensen N, Cullen M, Boland JF, Bass S, Steinberg MK, Raine-Bennett T, Lee D, Burk RD, Pinheiro M, Song L, Dean M, Nelson CW, Burdett L, Yu K, Roberson D, Lorey T, et al. Mutations in the HPV16 genome induced by APOBEC3 are associated with viral clearance. *Nat Commun*. 2020;11:886.
135. Harris RS, Liddament MT. Retroviral restriction by APOBEC proteins. *Nature Reviews Immunology*. 2004;4:868.
136. Turelli P, Mangeat B, Jost S, Vianin S, Trono D. Inhibition of Hepatitis B Virus Replication by APOBEC3G. *Science*. 2004;303:1829-.
137. Suspène R, Aynaud M-M, Koch S, Pasdeloup D, Labetoulle M, Gaertner B, Vartanian J-P, Meyerhans A, Wain-Hobson S. Genetic editing of Herpes Simplex 1 and Epstein Barr herpesvirus genomes by human APOBEC-3 cytidine deaminases in culture and In Vivo. *J Virol*. 2011.
138. Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, Kiezun A, Kryukov GV, Carter SL, Saksena G, Harris S, Shah RR, Resnick MA, Getz G, Gordenin DA. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat Genet*. 2013;45:970-6.
139. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies

HR, Desmedt C, Eils R, Eyfjord JE, Foekens JA, Greaves M, et al. Signatures of mutational processes in human cancer. *Nature*. 2013;500:415-21.
140. Chu D, Wei L. Nonsynonymous, synonymous and nonsense mutations in human cancer-related genes undergo stronger purifying selections than expectation. *BMC Cancer*. 2019;19:359.
141. Nelson CW, Hughes AL. Within-host nucleotide diversity of virus populations: insights from next-generation sequencing. *Infect Genet Evol*. 2015;30:1-7.
142. Mugal CF, Wolf JB, Kaj I. Why time matters: codon evolution and the temporal dynamics of dN/dS. *Mol Biol Evol*. 2014;31:212-31.
143. Kryazhimskiy S, Plotkin JB. The population genetics of dN/dS. *PLoS Genet*. 2008;4:e1000304.
144. Bonde J, Ejegod DM, Cuschieri K, Dillner J, Heideman DAM, Quint W, Pavon Ribas MA, Padalko E, Christiansen IK, Xu L, Arbyn M. The Valgent4 protocol: Robust analytical and clinical validation of 11 HPV assays with genotyping on cervical samples collected in SurePath medium. *J Clin Virol*. 2018;108:64-71.
145. Poljak M, Kocjan BJ, Ostrbenk A, Seme K. Commercially available molecular tests for human papillomaviruses (HPV): 2015 update. *J Clin Virol*. 2016;76 Suppl 1:S3-S13.
146. Meijer CJ, Berkhof J, Castle PE, Hesselink AT, Franco EL, Ronco G, Arbyn M, Bosch FX, Cuzick J, Dillner J, Heideman DA, Snijders PJ. Guidelines for human papillomavirus DNA test requirements for primary cervical cancer screening in women 30 years and older. *Int J Cancer*. 2009;124:516-20.
147. The Norwegian Cervical Cancer Screening Programme. Krav til HPV-tester. Available on https://www.kreftregisteret.no/screening/livmorhalsprogrammet/Helsepersonell/screeningstrategi-og-nasjonale-retningslinjer/krav-til-hpv-tester/ on February 21, 2020.
148. Arbyn M, Depuydt C, Benoy I, Bogers J, Cuschieri K, Schmitt M, Pawlita M, Geraets D, Heard I, Gheit T, Tommasino M, Poljak M, Bonde J, Quint W. VALGENT: A protocol for clinical validation of human papillomavirus assays. *J Clin Virol*. 2016;76 Suppl 1:S14-S21.
149. Tsakogiannis D, Gartzonika C, Levidiotou-Stefanou S, Markoulatos P. Molecular approaches for HPV genotyping and HPV-DNA physical status. *Expert Rev Mol Med*. 2017;19:e1.
150. Abreu AL, Souza RP, Gimenes F, Consolaro ME. A review of methods for detect human Papillomavirus infection. *Virol J*. 2012;9:262.
151. Karlsen F, Kalantari M, Jenkins A, Pettersen E, Kristensen G, Holm R, Johansson B, Hagmar B. Use of multiple PCR primer sets for optimal detection of human papillomavirus. *J Clin Microbiol*. 1996;34:2095-100.
152. Gravitt PE, Peyton CL, Alessi TQ, Wheeler CM, Coutlee F, Hildesheim A, Schiffman MH, Scott DR, Apple RJ. Improved amplification of genital human papillomaviruses. *J Clin Microbiol*. 2000;38:357-61.
153. Kleter B, van Doorn LJ, ter Schegget J, Schrauwen L, van Krimpen K, Burger M, ter Harmsel B, Quint W. Novel short-fragment PCR assay for highly sensitive broad-spectrum detection of anogenital human papillomaviruses. *The American journal of pathology*. 1998;153:1731-9.
154. Soderlund-Strand A, Carlson J, Dillner J. Modified general primer PCR system for sensitive detection of multiple types of oncogenic human papillomavirus. *J Clin Microbiol*. 2009;47:541-6.
155. Depuydt CE, Boulet GA, Horvath CA, Benoy IH, Vereecken AJ, Bogers JJ. Comparison of MY09/11 consensus PCR and type-specific PCRs in the detection of oncogenic HPV types. *J Cell Mol Med*. 2007;11:881-91.

156. Coutlee F, Gravitt P, Kornegay J, Hankins C, Richardson H, Lapointe N, Voyer H, Franco E. Use of PGMY Primers in L1 Consensus PCR Improves Detection of Human Papillomavirus DNA in Genital Samples. *J Clin Microbiol.* 2002;40:902-7.
157. Sotlar K, Diemer D, Dethleffs A, Hack Y, Stubner A, Vollmer N, Menton S, Menton M, Dietz K, Wallwiener D, Kandolf R, Bultmann B. Detection and typing of human papillomavirus by e6 nested multiplex PCR. *J Clin Microbiol.* 2004;42:3176-84.
158. Lin CY, Chao A, Yang YC, Chou HH, Ho CM, Lin RW, Chang TC, Chiou JY, Chao FY, Wang KL, Chien TY, Hsueh S, Huang CC, Chen CJ, Lai CH. Human papillomavirus typing with a polymerase chain reaction-based genotyping array compared with type-specific PCR. *J Clin Virol.* 2008;42:361-7.
159. Morris BJ. Cervical human papillomavirus screening by PCR: advantages of targeting the E6/E7 region. *Clin Chem Lab Med.* 2005;43:1171-7.
160. Luft F, Klaes R, Nees M, Durst M, Heilmann V, Melsheimer P, von Knebel Doeberitz M. Detection of integrated papillomavirus sequences by ligation-mediated PCR (DIPS-PCR) and molecular characterization in cervical cancer cells. *Int J Cancer.* 2001;92:9-17.
161. Klaes R, Woerner SM, Ridder R, Wentzensen N, Duerst M, Schneider A, Lotz B, Melsheimer P, von Knebel Doeberitz M. Detection of high-risk cervical intraepithelial neoplasia and cervical cancer by amplification of transcripts derived from integrated papillomavirus oncogenes. *Cancer Res.* 1999;59:6132-6.
162. Hudelist G, Manavi M, Pischinger KI, Watkins-Riedel T, Singer CF, Kubista E, Czerwenka KF. Physical state and expression of HPV DNA in benign and dysplastic cervical tissue: different levels of viral integration are correlated with lesion grade. *Gynecol Oncol.* 2004;92:873-80.
163. Arias-Pulido H, Peyton CL, Joste NE, Vargas H, Wheeler CM. Human papillomavirus type 16 integration in cervical carcinoma in situ and in invasive cervical cancer. *J Clin Microbiol.* 2006;44:1755-62.
164. Peitsaro P, Johansson B, Syrjanen S. Integrated human papillomavirus type 16 is frequently found in cervical cancer precursors as demonstrated by a novel quantitative real-time PCR technique. *J Clin Microbiol.* 2002;40:886-91.
165. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A.* 1977;74:5463-7.
166. Petrackova A, Vasinek M, Sedlarikova L, Dyskova T, Schneiderova P, Novosad T, Papajik T, Kriegova E. Standardization of Sequencing Coverage Depth in NGS: Recommendation for Detection of Clonal and Subclonal Mutations in Cancer Diagnostics. *Front Oncol.* 2019;9:851.
167. Tuna M, Amos CI. Next generation sequencing and its applications in HPV-associated cancers. *Oncotarget.* 2017;8:8877-89.
168. Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. Comparison of next-generation sequencing systems. *J Biomed Biotechnol.* 2012;2012:251364.
169. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J, Trajanoski Z. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform.* 2014;15:256-78.
170. Levy SE, Myers RM. Advancements in Next-Generation Sequencing. *Annu Rev Genomics Hum Genet.* 2016;17:95-115.
171. Illumina. DNA Sequencing Methods Collection. An overview of recent DNA-seq publications featuring Illumina technology. 2017. Available on https://www.illumina.com/content/dam/illumina-marketing/documents/products/research_reviews/dna-sequencing-methods-review-web.pdf on January 19, 2020.
172. Illumina. RNA Sequencing Sethods Collection. An overview of recent RNA-Seq publications featuring Illumina® technology. 2017. Available on

https://www.illumina.com/content/dam/illumina-marketing/documents/products/research_reviews/rna-sequencing-methods-review-web.pdf on January 19, 2020.
173. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*. 2016;17:333-51.
174. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, Altaf-Ul-Amin M, Ogasawara N, Kanaya S. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res*. 2011;39:e90.
175. Meacham F, Boffelli D, Dhahbi J, Martin DI, Singer M, Pachter L. Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics*. 2011;12:451.
176. Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res*. 2015;43:e37.
177. Brockman W, Alvarez P, Young S, Garber M, Giannoukos G, Lee WL, Russ C, Lander ES, Nusbaum C, Jaffe DB. Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res*. 2008;18:763-70.
178. Escobar-Escamilla N, Ramirez-Gonzalez JE, Castro-Escarpulli G, Diaz-Quinonez JA. Utility of high-throughput DNA sequencing in the study of the human papillomaviruses. *Virus Genes*. 2018;54:17-24.
179. Arroyo LS, Smelov V, Bzhalava D, Eklund C, Hultin E, Dillner J. Next generation sequencing for human papillomavirus genotyping. *J Clin Virol*. 2013;58:437-42.
180. Barzon L, Militello V, Lavezzo E, Franchin E, Peta E, Squarzon L, Trevisan M, Pagni S, Dal Bello F, Toppo S, Palu G. Human papillomavirus genotyping by 454 next generation sequencing technology. *J Clin Virol*. 2011;52:93-7.
181. 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. Development and validation of a new HPV genotyping assay based on next-generation sequencing. *Am J Clin Pathol*. 2014;141:796-804.
182. Dube Mandishora RS, Christiansen IK, Chin'ombe N, Duri K, Ngara B, Rounge TB, Meisal R, Ambur OH, Palefsky JM, Stray-Pedersen B, Chirenje ZM. Genotypic diversity of anogenital human papillomavirus in women attending cervical cancer screening in Harare, Zimbabwe. *J Med Virol*. 2017;89:1671-7.
183. Meisal R, Rounge TB, Christiansen IK, Eieland AK, Worren MM, Molden TF, Kommedal O, Hovig E, Leegaard TM, Ambur OH. HPV Genotyping of Modified General Primer-Amplicons Is More Analytically Sensitive and Specific by Sequencing than by Hybridization. *PLoS One*. 2017;12:e0169074.
184. Wagner S, Roberson D, Boland J, Yeager M, Cullen M, Mirabello L, Dunn ST, Walker J, Zuna R, Schiffman M, Wentzensen N. Development of the TypeSeq Assay for Detection of 51 Human Papillomavirus Genotypes by Next-Generation Sequencing. *J Clin Microbiol*. 2019;57.
185. Meiring TL, Salimo AT, Coetzee B, Maree HJ, Moodley J, Hitzeroth, II, Freeborough MJ, Rybicki EP, Williamson AL. Next-generation sequencing of cervical DNA detects human papillomavirus types not detected by commercial kits. *Virol J*. 2012;9:164.
186. Liu Y, Lu Z, Xu R, Ke Y. Comprehensive mapping of the human papillomavirus (HPV) DNA integration sites in cervical carcinomas by HPV capture technology. *Oncotarget*. 2016;7:5852-64.
187. Li T, Unger ER, Batra D, Sheth M, Steinau M, Jasinski J, Jones J, Rajeevan MS. Universal Human Papillomavirus Typing Assay: Whole-Genome Sequencing following Target Enrichment. *J Clin Microbiol*. 2017;55:811-23.
188. Kukimoto I, Maehama T, Sekizuka T, Ogasawara Y, Kondo K, Kusumoto-Matsuo R, Mori S, Ishii Y, Takeuchi T, Yamaji T, Takeuchi F, Hanada K, Kuroda M. Genetic variation of

human papillomavirus type 16 in individual clinical specimens revealed by deep sequencing. *PLoS One*. 2013;8:e80583.
189. Arroyo-Muhr LS, Lagheden C, Hultin E, Eklund C, Adami HO, Dillner J, Sundstrom K. The HPV16 Genome Is Stable in Women Who Progress to In Situ or Invasive Cervical Cancer: A Prospective Population-Based Study. *Cancer Res*. 2019;79:4532-8.
190. Lavezzo E, Masi G, Toppo S, Franchin E, Gazzola V, Sinigaglia A, Masiero S, Trevisan M, Pagni S, Palu G, Barzon L. Characterization of Intra-Type Variants of Oncogenic Human Papillomaviruses by Next-Generation Deep Sequencing of the E6/E7 Region. *Viruses*. 2016;8:79.
191. Chung TK, Van Hummelen P, Chan PK, Cheung TH, Yim SF, Yu MY, Ducar MD, Thorner AR, MacConaill LE, Doran G, Pedamallu CS, Ojesina AI, Wong RR, Wang VW, Freeman SS, Lau TS, Kwong J, Chan LK, Fromer M, May T, et al. Genomic aberrations in cervical adenocarcinomas in Hong Kong Chinese women. *Int J Cancer*. 2015;137:776-83.
192. Xu B, Chotewutmontri S, Wolf S, Klos U, Schmitz M, Durst M, Schwarz E. Multiplex Identification of Human Papillomavirus 16 DNA Integration Sites in Cervical Carcinomas. *PLoS One*. 2013;8:e66693.
193. Holmes A, Lameiras S, Jeannot E, Marie Y, Castera L, Sastre-Garau X, Nicolas A. Mechanistic signatures of HPV insertions in cervical carcinomas. *npj Genomic Medicine*. 2016;1.
194. Brant AC, Menezes AN, Felix SP, de Almeida LM, Sammeth M, Moreira MAM. Characterization of HPV integration, viral gene expression and E6E7 alternative transcripts by RNA-Seq: A descriptive study in invasive cervical cancer. *Genomics*. 2019;111:1853-61.
195. Abnizova I, Boekhorst Rt, Orlov YL. Computational Errors and Biases in Short Read Next Generation Sequencing. *Journal of Proteomics & Bioinformatics*. 2017;10.
196. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res*. 2010;38:1767-71.
197. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*. 1998;8:186-94.
198. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*. 1998;8:175-85.
199. Kelley DR, Schatz MC, Salzberg SL. Quake: quality-aware detection and correction of sequencing errors. *Genome Biol*. 2010;11:R116.
200. Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM. An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PLoS One*. 2013;8:e85024.
201. Fonseca NA, Rung J, Brazma A, Marioni JC. Tools for mapping high-throughput sequencing data. *Bioinformatics*. 2012;28:3169-77.
202. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215:403-10.
203. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet*. 2014;15:121-32.
204. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2010;26:589-95.
205. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10:R25.
206. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25:2078-9.
207. Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW, Schupp JM, Keim P, Morrow JB, Salit ML, Zook JM. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Front Genet*. 2015;6:235.

208. Kojima K, Nariai N, Mimori T, Takahashi M, Yamaguchi-Kabata Y, Sato Y, Nagasaki M. A statistical variant calling approach from pedigree information and local haplotyping with phase informative reads. *Bioinformatics*. 2013;29:2835-43.
209. Tattini L, D'Aurizio R, Magi A. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Front Bioeng Biotechnol*. 2015;3:92.
210. Qi Y, Liu X, Liu CG, Wang B, Hess KR, Symmans WF, Shi W, Pusztai L. Reproducibility of Variant Calls in Replicate Next Generation Sequencing Experiments. *PLoS One*. 2015;10:e0119230.
211. Ebbert MT, Wadsworth ME, Staley LA, Hoyt KL, Pickett B, Miller J, Duce J, Alzheimer's Disease Neuroimaging I, Kauwe JS, Ridge PG. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics*. 2016;17 Suppl 7:239.
212. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38:e164-e.
213. Nielsen CB, Cantor M, Dubchak I, Gordon D, Wang T. Visualizing genomes: techniques and challenges. *Nat Methods*. 2010;7:S5-S15.
214. van der Weele P, van Logchem E, Wolffs P, van den Broek I, Feltkamp M, de Melker H, Meijer CJ, Boot H, King AJ. Correlation between viral load, multiplicity of infection, and persistence of HPV16 and HPV18 infection in a Dutch cohort of young women. *J Clin Virol*. 2016;83:6-11.
215. Mollers M, Boot Hein J, Vriend Henrike J, King Audrey J, van den Broek Ingrid VF, van Bergen Jan EA, Brink Antoinette AT, Wolffs Petra FG, Hoebe Christian JP, Meijer Chris JL, van der Sande Marianne AB, de Melker Hester E. Prevalence, incidence and persistence of genital HPV infections in a large cohort of sexually active young women in the Netherlands. *Vaccine*. 2013;31:394-401.
216. van den Broek IV, Brouwers EE, Gotz HM, van Bergen JE, Op de Coul EL, Fennema JS, Koekenbier RH, Pars LL, van Ravesteijn SM, Hoebe CJ. Systematic selection of screening participants by risk score in a Chlamydia screening programme is feasible and effective. *Sex Transm Infect*. 2012;88:205-11.
217. van den Broek IV, Hoebe CJ, van Bergen JE, Brouwers EE, de Feijter EM, Fennema JS, Gotz HM, Koekenbier RH, van Ravesteijn SM, de Coul EL. 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*. 2010;10:89.
218. Trope A, Sjoborg K, Eskild A, Cuschieri K, Eriksen T, Thoresen S, Steinbakk M, Laurak V, Jonassen CM, Westerhagen U, Jacobsen MB, Lie AK. Performance of human papillomavirus DNA and mRNA testing strategies for women with and without cervical neoplasia. *J Clin Microbiol*. 2009;47:2458-64.
219. Trope A, Sjoborg KD, Nygard M, Roysland K, Campbell S, Alfsen GC, Jonassen CM. Cytology and human papillomavirus testing 6 to 12 months after ASCUS or LSIL cytology in organized screening to predict high-grade cervical neoplasia between screening rounds. *J Clin Microbiol*. 2012;50:1927-35.
220. Beaudenon S, Kremsdorf D, Croissant O, Jablonska S, Wain-Hobson S, Orth G. A novel type of human papillomavirus associated with genital neoplasias. *Nature*. 1986;321:246-9.
221. Schmitt M, Bravo IG, Snijders PJ, Gissmann L, Pawlita M, Waterboer T. Bead-based multiplex genotyping of human papillomaviruses. *J Clin Microbiol*. 2006;44:504-12.
222. Kleter B, van Doorn LJ, Schrauwen L, Molijn A, Sastrowijoto S, ter Schegget J, Lindeman J, ter Harmsel B, Burger M, Quint W. 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*. 1999;37:2508-17.

223. Seaman WT, Andrews E, Couch M, Kojic EM, Cu-Uvin S, Palefsky J, Deal AM, Webster-Cyriaque J. Detection and quantitation of HPV in genital and oral tissues and fluids by real time PCR. *Virol J*. 2010;7:194.
224. Van Doorslaer K, Tan Q, Xirasagar S, Bandaru S, Gopalan V, Mohamoud Y, Huyen Y, McBride AA. The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. *Nucleic Acids Res*. 2013;41:D571-8.
225. Van Doorslaer K, Li Z, Xirasagar S, Maes P, Kaminsky D, Liou D, Sun Q, Kaur R, Huyen Y, McBride AA. The Papillomavirus Episteme: a major update to the papillomavirus sequence database. *Nucleic Acids Res*. 2017;45:D499-D506.
226. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, Thompson JD, Higgins DG. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*. 2011;7:539.
227. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. Primer3--new capabilities and interfaces. *Nucleic Acids Res*. 2012;40:e115.
228. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Environ Microbiol*. 2013;79:5112-20.
229. Koster J, Rahmann S. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics*. 2012;28:2520-2.
230. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011*. 2011;17.
231. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*. 2015;12:357-60.
232. Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. *Genome Res*. 2011;21:487-93.
233. Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Res*. 2002;12:656-64.
234. Anchordoquy TJ, Molina MC. Preservation of DNA. *Cell Preservation Technology*. 2007;5:180-8.
235. Lamble S, Batty E, Attar M, Buck D, Bowden R, Lunter G, Crook D, El-Fahmawi B, Piazza P. Improved workflows for high throughput library preparation using the transposome-based Nextera system. *BMC Biotechnol*. 2013;13:104.
236. Snijders PJ, Hogewoning CJ, Hesselink AT, Berkhof J, Voorhorst FJ, Bleeker MC, Meijer CJ. Determination of viral load thresholds in cervical scrapings to rule out CIN 3 in HPV 16, 18, 31 and 33-positive women with normal cytology. *Int J Cancer*. 2006;119:1102-7.
237. Liu Y, Zhang C, Gao W, Wang L, Pan Y, Gao Y, Lu Z, Ke Y. Genome-wide profiling of the human papillomavirus DNA integration in cervical intraepithelial neoplasia and normal cervical epithelium by HPV capture technology. *Sci Rep*. 2016;6:35427.
238. Garcia-Garcia G, Baux D, Faugere V, Moclyn M, Koenig M, Claustres M, Roux AF. Assessment of the latest NGS enrichment capture methods in clinical context. *Sci Rep*. 2016;6:20948.
239. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. Target-enrichment strategies for next-generation sequencing. *Nature Methods*. 2010;7:111-8.
240. Kebschull JM, Zador AM. Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res*. 2015;43:e143.
241. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. Characterizing and measuring bias in sequence data. *Genome Biol*. 2013;14:R51.
242. Eckert KA, Kunkel TA. DNA polymerase fidelity and the polymerase chain reaction. *PCR Methods Appl*. 1991;1:17-24.
243. Potapov V, Ong JL. Examining Sources of Error in PCR by Single-Molecule Sequencing. *PLoS One*. 2017;12:e0169774.

244. Tindall KR, Kunkel TA. Fidelity of DNA synthesis by the Thermus aquaticus DNA polymerase. *Biochemistry*. 1988;27:6008-13.
245. McInerney P, Adams P, Hadi MZ. Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase. *Mol Biol Int*. 2014;2014:287430.
246. Thankaswamy-Kosalai S, Sen P, Nookaew I. Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. *Genomics*. 2017;109:186-91.
247. Gampawar P, Saba Y, Werner U, Schmidt R, Müller-Myhsok B, Schmidt H. Evaluation of the Performance of AmpliSeq and SureSelect Exome Sequencing Libraries for Ion Proton. *Frontiers in Genetics*. 2019;10.
248. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci Rep*. 2015;5:17875.
249. Chen Y, Yao H, Thompson EJ, Tannir NM, Weinstein JN, Su X. VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. *Bioinformatics*. 2013;29:266-7.
250. Forster M, Szymczak S, Ellinghaus D, Hemmrich G, Ruhlemann M, Kraemer L, Mucha S, Wienbrandt L, Stanulla M, Group UFOSCwI-BS, Franke A. Vy-PER: eliminating false positive detection of virus integration events in next generation sequencing data. *Sci Rep*. 2015;5:11534.
251. Ho DW, Sze KM, Ng IO. Virus-Clip: a fast and memory-efficient viral integration site detection tool at single-base resolution with annotation capability. *Oncotarget*. 2015;6:20959-63.
252. Jennings LJ, Arcila ME, Corless C, Kamel-Reid S, Lubin IM, Pfeifer J, Temple-Smolkin RL, Voelkerding KV, Nikiforova MN. Guidelines for Validation of Next-Generation Sequencing-Based Oncology Panels: A Joint Consensus Recommendation of the Association for Molecular Pathology and College of American Pathologists. *J Mol Diagn*. 2017;19:341-65.
253. McDonald JH. Handbook of Biological Statistics. 3rd ed. Baltimore, Maryland: Sparky House Publishing, 2014.
254. Bellack AS, Hersen M. Comprehensive clinical psychology. Amsterdam; New York: Pergamon, 1998.
255. Kadam P, Bhalerao S. Sample size calculation. *Int J Ayurveda Res*. 2010;1:55-7.
256. Hirose Y, Onuki M, Tenjimbayashi Y, Yamaguchi-Naka M, Mori S, Tasaka N, Satoh T, Morisada T, Iwata T, Kiyono T, Mimura T, Sekizawa A, Matsumoto K, Kukimoto I. Whole-Genome Analysis of Human Papillomavirus Type 16 Prevalent in Japanese Women with or without Cervical Lesions. *Viruses*. 2019;11.
257. Liu Y, Pan Y, Gao W, Ke Y, Lu Z. Whole-Genome Analysis of Human Papillomavirus Types 16, 18, and 58 Isolated from Cervical Precancer and Cancer Samples in Chinese Women. *Sci Rep*. 2017;7:263.
258. Shen-Gunther J, Wang Y, Lai Z, Poage GM, Perez L, Huang TH. Deep sequencing of HPV E6/E7 genes reveals loss of genotypic diversity and gain of clonal dominance in high-grade intraepithelial lesions of the cervix. *BMC Genomics*. 2017;18:231.
259. Mariaggi AA, Pere H, Perrier M, Visseaux B, Collin G, Veyer D, Le Hingrat Q, Ferre VM, Joly V, Couvelard A, Bucau M, Davitian C, Descamps D, Abramowitz L, Charpentier C. Presence of Human Papillomavirus (HPV) Apolipoprotein B Messenger RNA Editing, Catalytic Polypeptide-Like 3 (APOBEC)-Related Minority Variants in HPV-16 Genomes From Anal and Cervical Samples but Not in HPV-52 and HPV-58. *J Infect Dis*. 2018;218:1027-36.
260. Li H, Yang Y, Zhang R, Cai Y, Yang X, Wang Z, Li Y, Cheng X, Ye X, Xiang Y, Zhu B. Preferential sites for the integration and disruption of human papillomavirus 16 in cervical lesions. *J Clin Virol*. 2013;56:342-7.

261. Jaisamrarn U, Castellsague X, Garland SM, Naud P, Palmroth J, Del Rosario-Raymundo MR, Wheeler CM, Salmeron J, Chow SN, Apter D, Teixeira JC, Skinner SR, Hedrick J, Szarewski A, Romanowski B, Aoki FY, Schwarz TF, Poppe WA, Bosch FX, de Carvalho NS, et al. Natural history of progression of HPV infection to cervical lesion or clearance: analysis of the control arm of the large, randomised PATRICIA study. *PLoS One*. 2013;8:e79260.
262. Skinner SR, Wheeler CM, Romanowski B, Castellsague X, Lazcano-Ponce E, Del Rosario-Raymundo MR, Vallejos C, Minkina G, Pereira Da Silva D, McNeil S, Prilepskaya V, Gogotadze I, Money D, Garland SM, Romanenko V, Harper DM, Levin MJ, Chatterjee A, Geeraerts B, Struyf F, et al. Progression of HPV infection to detectable cervical lesions or clearance in adult women: Analysis of the control arm of the VIVIANE study. *Int J Cancer*. 2016;138:2428-38.
263. Simanaviciene V, Popendikyte V, Gudleviciene Z, Zvirbliene A. Different DNA methylation pattern of HPV16, HPV18 and HPV51 genomes in asymptomatic HPV infection as compared to cervical neoplasia. *Virology*. 2015;484:227-33.
264. Olmedo-Nieva L, Munoz-Bello JO, Contreras-Paredes A, Lizano M. The Role of E6 Spliced Isoforms (E6*) in Human Papillomavirus-Induced Carcinogenesis. *Viruses*. 2018;10.
265. Raybould R, Fiander A, Wilkinson GW, Hibbitts S. HPV integration detection in CaSki and SiHa using detection of integrated papillomavirus sequences and restriction-site PCR. *J Virol Methods*. 2014;206:51-4.
266. Mincheva A, Gissmann L, zur Hausen H. Chromosomal integration sites of human papillomavirus DNA in three cervical cancer cell lines mapped by in situ hybridization. *Med Microbiol Immunol*. 1987;176:245-56.
267. Geisbill J, Osmers U, Durst M. Detection and characterization of human papillomavirus type 45 DNA in the cervical carcinoma cell line MS751. *J Gen Virol*. 1997;78 (Pt 3):655-8.
268. Yee C, Krishnan-Hewlett I, Baker CC, Schlegel R, Howley PM. Presence and expression of human papillomavirus sequences in human cervical carcinoma cell lines. *Am J Pathol*. 1985;119:361-6.
269. Ferreira D, Adega F, Chaves R. The Importance of Cancer Cell Lines as in vitro Models in Cancer Methylome Analysis and Anticancer Drugs Testing. 2013. p. 139-66.
270. Kulkarni P, Frommolt P. Challenges in the Setup of Large-scale Next-Generation Sequencing Analysis Workflows. *Comput Struct Biotechnol J*. 2017;15:471-7.
271. Gargis AS, Kalman L, Berry MW, Bick DP, Dimmock DP, Hambuch T, Lu F, Lyon E, Voelkerding KV, Zehnbauer BA, Agarwala R, Bennett SF, Chen B, Chin EL, Compton JG, Das S, Farkas DH, Ferber MJ, Funke BH, Furtado MR, et al. Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nat Biotechnol*. 2012;30:1033-6.
272. Chassang G. The impact of the EU general data protection regulation on scientific research. *Ecancermedicalscience*. 2017;11:709-.
273. Thorogood A, Dalpe G, Knoppers BM. Return of individual genomic research results: are laws and policies keeping step? *Eur J Hum Genet*. 2019;27:535-46.
274. Schickhardt C, Fleischer H, Winkler EC. Do patients and research subjects have a right to receive their genomic raw data? An ethical and legal analysis. *BMC Med Ethics*. 2020;21:7.
275. Norwegian Industrial Property Office. 2018. Available on https://www.patentstyret.no/en/services/patents/good-reasons-for-applying-for-a-patent/ on February 25, 2020.

8 APPENDIX

8.1 Appendix 1: Patent application

Rijksdienst voor Ondernemend Nederland

Ontvangstbevestiging

Ontvangstbevestiging van uw verzoek om een aanvrage om een octrooi bij Octrooicentrum Nederland:

Behandelingsnummer	500231999	
(voorlopig) aanvraagnummer	N2022043	
Datum ontvangst	21 november 2018	
Ontvangend bureau	Octrooicentrum Nederland, Den Haag	
Uw referentie	76654NL8	
Aanvrager	Akershus Universitetssykehus HF	
Aantal aanvragers	4	
Land van herkomst	NO	
Korte aanduiding	Tagmentation-Associated Multiplex PCR Enrichment Sequencing	
Meegestuurde documenten	package-data.xml	nl-request.xml
	application-body.xml	nl-request.pdf (3 p.)
	BESCHR.pdf\76654NL8 Description.pdf (43 p.)	CONCL.pdf\76654NL8 Claims.pdf (5 p.)
	UITTR.pdf\76654NL8 Abstract.pdf (1 p.)	TEK.pdf\76654NL8 Drawings.pdf (6 p.)
	OLF-ARCHIVE.zip\76654NL8 Non-converted documents.zip	4001UITV-1.pdf (1 p.)
	4001UITV-2.pdf (1 p.)	4001UITV-3.pdf (1 p.)
	4001UITV-4.pdf (1 p.)	4001UITV-5.pdf (1 p.)
	4001UITV-6.pdf (1 p.)	4001UITV-7.pdf (1 p.)
Ingediend door	CN=Debora Thie 62455	
Methode van indiening	Online	
Datum en tijd aanmaak ontvangstbevestiging	21 november 2018, 16:01:37 (CET)	
Unieke reeks tekens	E5:42:5C:54:91:B7:3F:61:6A:33:73:F9:5F:A2:FE:A2:6F:CC:3D:F2	

Aanvrage om octrooi

Referentie aanvrager: 76654NL8
Aanvraagnummer:

0	**Dit gedeelte wordt door het Octrooicentrum ingevuld**	
	Aanvraagnummer:	
	Datum ontvangst aanvrage:	
	Indieningsdatum:	
1	**Verzoek**	
	Verzoek om verlening van octrooi volgens de bepalingen van de Rijksoctrooiwet	
1-2	Nieuwheidsonderzoek van internationaal type:	✓
2-1	**Aanvrager 1**	
	Naam:	Akershus Universitetssykehus HF
	Relatienummer:	
	Adres:	Postboks 1000
		1478 LØRENSKOG
		Norway
	Bedrijfsgrootte:	Categorie onbekend
2-2	**Aanvrager 2**	
	Naam:	Kreftregisteret
	Relatienummer:	
	Adres:	Kreftregisteret
		Postboks 5313 Majorstuen
		0304 OSLO
		Norway
	Bedrijfsgrootte:	Categorie onbekend
2-3	**Aanvrager 3**	
	Naam:	University of Helsinki
	Relatienummer:	
	Adres:	University of Helsinki
		P. box 33
		00014 HELSINKI
		Finland
	Bedrijfsgrootte:	Categorie onbekend
2-4	**Aanvrager 4**	
	Naam:	OsloMet - storbyuniversitetet
	Relatienummer:	
	Adres:	OsloMet - storbyuniversitetet
		Postboks 4 St. Olavs plass
		0130 OSLO
		Norway
	Bedrijfsgrootte:	Categorie onbekend
3-1	**Gemachtigde 1**	
	Naam:	VAN EEKELEN Markus Johannus Leonardus
	Relatienummer:	
	Adres:	Algemeen Octrooi- en Merkenbureau B.V.

Referentie aanvrager: 76654NL8
Aanvraagnummer:

		P.O. Box 645	
		5600 AP EINDHOVEN	
		Nederland	
	Telefoonnummer:	0031 40 2433715	
	Faxnummer:	0031 40 2434557	
	e-mailadres:	mail@aomb.nl	
4	**Uitvinder(s)** Uitvindergegevens worden apart ingediend		
5	**Korte aanduiding**	Tagmentation-Associated Multiplex PCR Enrichment Sequencing	
6	**Beroep op voorrang (GB=gebruiksmodel OC=octrooiaanvrage)**		
7	**De aanvrage omvat het gebruik van een micro-organisme** art-18 van de Uitvoeringsbepalingen 1995		
8	**Nucleotide en Amino zuur sequenties**		
9	**Documenten**	Details:	Elektronisch bestand:
9-1	Verzoek		als nl-request.pdf
9-2	1. Uitvinder	1. Uitvinder	als 4001UITV-1.pdf
9-3	2. Uitvinder	2. Uitvinder	als 4001UITV-2.pdf
9-4	3. Uitvinder	3. Uitvinder	als 4001UITV-3.pdf
9-5	4. Uitvinder	4. Uitvinder	als 4001UITV-4.pdf
9-6	5. Uitvinder	5. Uitvinder	als 4001UITV-5.pdf
9-7	6. Uitvinder	6. Uitvinder	als 4001UITV-6.pdf
9-8	7. Uitvinder	7. Uitvinder	als 4001UITV-7.pdf
9-9	Taks betaling		
9-10	Validatie verslag		
10	**Technische documenten**	Details:	Elektronisch bestand:
10-1	Beschrijving		BESCHR.pdf
10-2	Conclusies	45 conclusies	CONCL.pdf
10-3	Tekeningen	12 figuren	TEK.pdf
10-4	Uittreksel		UITTR.pdf
10-5	Vooraf conversie archief ?? (pre_conversion_archive)		OLF-ARCHIVE.zip
11	**Overige documenten**	Details:	Elektronisch bestand:
12	**Betalingen**		
12-1	**Wijze van betalen**	Depot	
	Het octrooicentrum wordt hierbij toestemming verleend de verschuldigde taks(en) van onderstaand depotnummer af te schrijven.	✓	
	Valuta:	EURO	
	Depotnummer:	1030	
	Rekeninghouder:	Algemeen Octrooi- en Merkenbureau B.V.	

		Aantal	Taksschaal	Te betalen bedrag
13	**Taksen**			
13-1	001 Betaalopdracht Indieningstaks	1	80.00	80.00
13-2	003 Betaalopdracht VNO Internationaal	1	794.00	794.00
	Totaal:		**EUR**	**874.00**
14	**Annotaties**			

Referentie aanvrager: 76654NL8
Aanvraagnummer:

15 Handtekening(en)

Plaats:	**Eindhoven**
Datum:	**21 november 2018**
Getekend door:	**/Markus Johannus Leonardus VAN EEKELEN/**
Association	**Algemeen Octrooi- en Merkenbureau B.V.**
Hoedanigheid:	**(Vertegenwoordiger)**

Uitvindergegevens

Gebruikers referentie: 76654NL8
Aanvraagnummer:

	Uitvinder	
	Naam:	ROUNGE , Mw. Trine
	Adres:	[redacted]

Uitvindergegevens

Gebruikers referentie: 76654NL8
Aanvraagnummer:

	Uitvinder	
	Naam:	KRAUS CHRISTIANSEN, Mw. Irene
	Adres:	[redacted]

Uitvindergegevens

Gebruikers referentie: 76654NL8
Aanvraagnummer:

	Uitvinder	
	Naam:	AMBUR , Dhr. Ole Herman
	Adres:	[redacted]

Uitvindergegevens

Gebruikers referentie: 76654NL8
Aanvraagnummer:

	Uitvinder	
	Naam:	LAGSTRÖM , Mw. Sonja
	Adres:	[redacted]

Uitvindergegevens

Gebruikers referentie: 76654NL8
Aanvraagnummer:

	Uitvinder	
	Naam:	MEISAL, Dr. Roger
	Adres:	[redacted]

Uitvindergegevens

Gebruikers referentie: 76654NL8
Aanvraagnummer:

	Uitvinder	
	Naam:	ELLONEN , Dhr. Pekka
	Adres:	[redacted]

Uitvindergegevens

Gebruikers referentie: 76654NL8
Aanvraagnummer:

	Uitvinder	
	Naam:	LEPISTÖ , Mw. Maja
	Adres:	[redacted]

A B S T R A C T

The present invention is related to methods for parallel sequencings of nucleic acid target sequences of interest, and in particular to massively parallel sequencing of nucleic acid sequences such as viral sequences that may have been integrated into a genome. For example, the methods, systems and kits provided herein may be used to enrich and sequence viral DNA sequences such as HPV and HIV sequences.

9 PAPERS I-III

I

SCIENTIFIC REPORTS

OPEN

TaME-seq: An efficient sequencing approach for characterisation of HPV genomic variability and chromosomal integration

Received: 29 May 2018
Accepted: 26 November 2018
Published online: 24 January 2019

Sonja Lagström[1,2], Sinan Uğur Umu[2], Maija Lepistö[3], Pekka Ellonen[3], Roger Meisal[1], Irene Kraus Christiansen[1,4], Ole Herman Ambur[5] & Trine B. Rounge[2]

HPV genomic variability and chromosomal integration are important in the HPV-induced carcinogenic process. To uncover these genomic events in an HPV infection, we have developed an innovative and cost-effective sequencing approach named TaME-seq (tagmentation-assisted multiplex PCR enrichment sequencing). TaME-seq combines tagmentation and multiplex PCR enrichment for simultaneous analysis of HPV variation and chromosomal integration, and it can also be adapted to other viruses. For method validation, cell lines (n = 4), plasmids (n = 3), and HPV16, 18, 31, 33 and 45 positive clinical samples (n = 21) were analysed. Our results showed deep HPV genome-wide sequencing coverage. Chromosomal integration breakpoints and large deletions were identified in HPV positive cell lines and in one clinical sample. HPV genomic variability was observed in all samples allowing identification of low frequency variants. In contrast to other approaches, TaME-seq proved to be highly efficient in HPV target enrichment, leading to reduced sequencing costs. Comprehensive studies on HPV intra-host variability generated during a persistent infection will improve our understanding of viral carcinogenesis. Efficient identification of both HPV variability and integration sites will be important for the study of HPV evolution and adaptability and may be an important tool for use in cervical cancer diagnostics.

Human papillomavirus (HPV) is the main cause of cervical cancer[1], one of the most common cancers in women worldwide, causing more than 200,000 deaths each year[2,3]. A persistent infection with HPV high-risk genotypes is recognised as a necessary cause of cancer development[4]. Of the 13 carcinogenic high-risk types, HPV16 and 18 are associated with about 70% of all cervical cancers[5,6]. HPV infection is also associated with cancer in penis, vulva, vagina, anus, and head and neck[7]. However, only a small fraction of HPV infections at any site will progress to cancer[8]. This indicates that in addition to HPV infection, additional factors such as HPV genomic variability and integration, could contribute to the HPV-induced carcinogenic process. An appropriate sequencing approach is needed to uncover these genomic events during a persistent HPV infection.

HPV contains an approximately 7.9 kb circular double-stranded DNA genome, consisting of early region (E1, E2, E4-7) genes, late region (L1, L2) genes and an upstream regulatory region (URR)[9]. To date, more than 200 HPV types have been identified[10]. Each individual HPV type shares at least 90% sequence identity in the conserved L1 open reading frame (ORF) nucleotide sequence. Isolates of the same HPV types that differ by 1–10% or 0.5–1% across the genome are referred to as variant lineages or sublineages, respectively[11,12].

Despite phylogenetic relatedness, HPV variant lineages can differ in their carcinogenic potential[13–16]. Traditionally, studies have focused on cancer risk of main variants. However, recent studies have revealed variability below the level of variant lineages that may be evidence of intra-host viral evolution and adaptation[17–20]. In contrast to a limited number of studies on HPV variability, HPV integration into the host genome has been more

[1]Department of Microbiology and Infection Control, Akershus University Hospital, Lørenskog, Norway. [2]Department of Research, Cancer Registry of Norway, Oslo, Norway. [3]Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland. [4]Clinical Molecular Biology (EpiGen), Medical Division, Akershus University Hospital and Institute of Clinical Medicine, University of Oslo, Norway. [5]Faculty of Health Sciences, OsloMet - Oslo Metropolitan University, Oslo, Norway. Correspondence and requests for materials should be addressed to T.B.R. (email: trine.rounge@kreftregisteret.no)

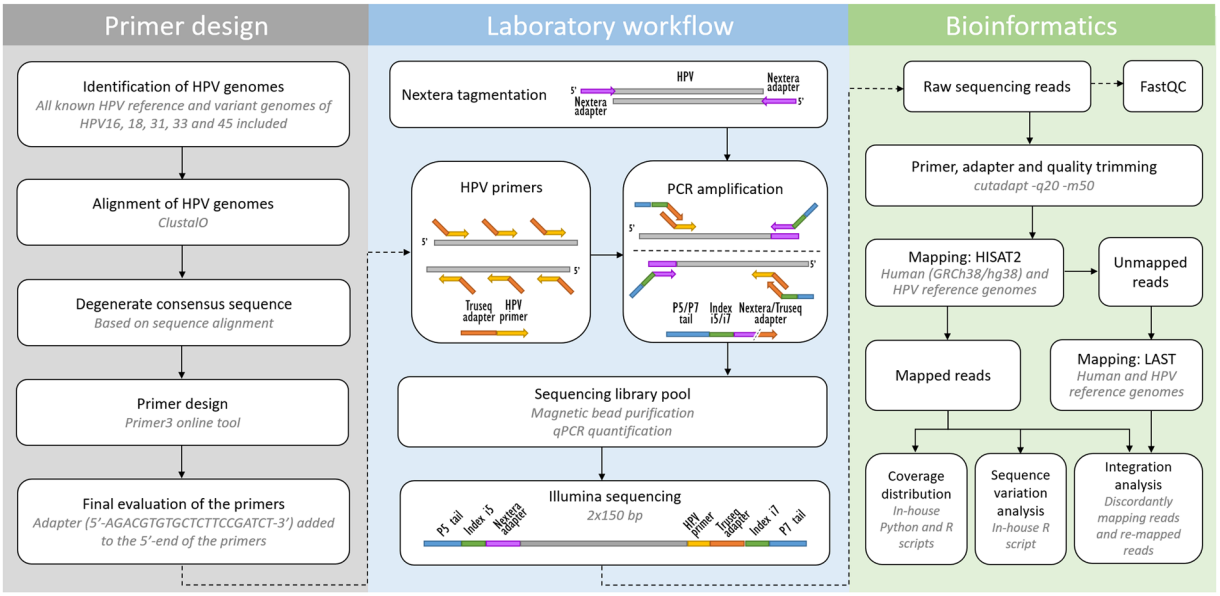

Figure 1. Primer design, laboratory and bioinformatics workflows of the TaME-seq method.

widely studied and is regarded as a determining event in cervical carcinogenesis[21–23]. Upon integration, disruption or complete deletion of the E1 or E2 gene is often observed, resulting in constitutive expression of the E6 and E7 oncogenes[24–26], inactivation of cell cycle checkpoints and genetic instability[23]. Viral integration may also lead to modified expression of cellular genes nearby, disruption of genes, as well as genomic amplifications that may promote oncogenesis[23,27]. The finding of certain chromosomal clusters of integration in precancerous lesions and cancers[28] also suggests a selective advantage of specific HPV integrations. Still, several important questions remain for HPV integration and more comprehensive analyses of integration sites are needed in order to expand our understanding of HPV pathogenesis.

The development of next generation sequencing (NGS) technologies has provided new tools for viral genomic research. During the recent years, a few studies have described different NGS based approaches to study HPV variability and integration in the human genome. The most common approaches used in HPV genomic analyses are based on target enrichment using highly multiplexed degenerate primers[29], enrichment by multiplex PCR using HPV16 forward primers[30], bead-based target capture[31–33], and rolling circle amplification[34] followed by NGS. These methods are however designed to detect either HPV integration or HPV variability. In addition, target capture methods poorly enrich HPV and remain expensive due to high probe cost and off-target sequencing.

In order to contribute to the understanding of the role of intra-host HPV genomic variability and chromosomal integration in carcinogenesis, we have developed an innovative library preparation strategy followed by an in-house bioinformatics pipeline named TaME-seq (tagmentation-assisted multiplex PCR enrichment sequencing). TaME-seq combines tagmentation and multiplex PCR enrichment, allowing simultaneous HPV genomic variability and integration analysis (Fig. 1). TaME-seq, with highly efficient target enrichment and reduced sequencing cost, enables deep sequencing analysis in order to find low frequency variants and rare integration events. Here, we present the results of HPV integration and genomic variability analysis in HPV16, 18, 31, 33 and 45 positive clinical samples and cell lines. The method described here provides an important tool for comprehensive studies of HPV genomic variability and chromosomal integration, and it can also be adapted to studies on other viruses such as retroviruses, adeno-associated viruses and integrating human herpesviruses.

Results

Read mapping analysis and genome coverage. Table 1 summarises liquid-based cytology (LBC) samples (n = 21), cell lines (n = 4) and plasmid samples (n = 3) included in the analysis. The samples generated 154.8 million raw reads of which 72.5 million reads (47%) mapped to the target HPV reference genomes. Only a small fraction (0.08%) of the reads mapped to other HPV types than those reported positive by HPV genotyping. The mean coverage ranged from 303 to 273898, while the fraction of the genome covered by minimum 10 × ranged from 0.35 to 1, and the fraction of the genome covered by minimum 100 × ranged from 0.33 to 1 (Table 1). HPV genome sequencing coverage aligned to the target HPV genomes with the location of HPV genomic regions and primers is visualised for CaSki, HeLa, LBC34, LBC11 and MS751 (Fig. 2). Overall, the samples showed varying HPV genome coverage profiles (Supplementary Figs S1–S5). Totally, 10 HPV positive samples were excluded from further analysis due to poor sequencing coverage (Supplementary Table S1). Sequencing of the HPV negative control samples resulted in no or negligible amount (<500) of reads mapped to target HPV genomes (Supplementary Table S2). The MS751 cell line was confirmed not to contain HPV18 sequences (Supplementary Table S1)[35].

Deletions in HPV genomes. The method enables identification of regions covered with very few or no sequencing reads, interpreted as large HPV genomic deletions. Cell lines HeLa and MS751 are known to contain partial HPV genomes due to deletions of 2.5 kb and 5 kb, respectively[35,36], which was confirmed by our method

Sample	Sample type	Raw reads	Trimmed reads	Reads mapped to target HPV	% Reads mapped to target HPV	Mean coverage	Fraction of genome covered by minimum	
							10×	100×
HPV16								
CaSki	Cell line	16138790[b]	12944262	12634651	78%	184716	1.00	1.00
SiHa	Cell line	151168[b]	133360	67496	45%	1018	0.96	0.83
SiHa-1	Cell line	5948008[c]	3735936	1249594	21%	17561	0.93	0.90
SiHa-1	Cell line	844178[b]	532874	181199	21%	2554	0.92	0.78
SiHa-2	Cell line	1405886[c]	789664	420774	30%	5609	0.91	0.85
SiHa-2	Cell line	158672[b]	90150	48412	31%	646	0.84	0.52
WHO std HPV16	Plasmid	359638[b]	304002	278987	78%	4104	0.99	0.96
LBC1[a]	LBC	128008[b]	108756	75323	59%	1124	0.96	0.88
LBC7[a]	LBC	62246[b]	51590	25567	41%	384	0.94	0.66
HPV18								
HeLa	Cell line	1433248[b]	1120824	394420	28%	5897	0.68	0.62
WHO std HPV18	Plasmid	2021206[b]	1358182	1098783	54%	15447	0.99	0.96
LBC103[a]	LBC	1477706[b]	1209564	74358	5%	1056	0.93	0.83
LBC105[a]	LBC	190664[b]	160450	32695	17%	484	0.51	0.34
LBC107	LBC	2180284[b]	1881868	978435	45%	14663	1.00	0.99
LBC108[a]	LBC	5407154[b]	3773986	3360463	62%	46691	1.00	0.98
LBC48[a]	LBC	641378[b]	433884	72589	11%	988	0.95	0.83
HPV31								
LBC16	LBC	276994[b]	191290	74465	27%	1065	0.94	0.80
LBC24[a]	LBC	471666[b]	348416	24197	5%	355	0.96	0.69
LBC32	LBC	2446832[b]	1523572	1319939	54%	18983	0.99	0.98
LBC34	LBC	3285680[b]	1841812	1723631	52%	23790	0.99	0.96
HPV33								
HPV33 plasmid	Plasmid	13824396[b]	5202718	5230090	38%	61527	1.00	1.00
LBC11	LBC	2852262[b]	1052512	986936	35%	12038	0.99	0.98
LBC30	LBC	77128[b]	51682	21431	28%	303	0.93	0.63
LBC31[a]	LBC	4276740[c]	2831408	44917	1.1%	544	0.76	0.60
LBC52	LBC	154936[b]	86990	34390	22%	439	0.95	0.62
LBC65[a]	LBC	368260[b]	248142	144022	39%	1993	1.00	0.91
HPV45								
MS751	Cell line	1221694[b]	1047286	56291	5%	845	0.35	0.33
LBC13[a]	LBC	496370[b]	389306	58293	12%	849	0.96	0.78
LBC29	LBC	211052[b]	122502	45925	22%	614	0.91	0.69
LBC36[a]	LBC	2412532[b]	1822912	1579570	65%	22093	1.00	0.97
LBC54	LBC	50169422[c]	26385910	20570184	41%	256857	1.00	1.00
LBC64[a]	LBC	5121416[c]	3040714	307476	6%	3943	0.95	0.88

Table 1. Read counts and sequencing coverage of HPV positive cell lines, plasmids and LBC samples. [a]Sample has multiple HPV infections. [b]Sequenced on MiSeq sequencing platform. [c]Sequenced on HiSeq 2500 sequencing platform.

(Fig. 2). A large deletion of 4.8 kb was revealed in the clinical sample LBC105, indicating partial or complete deletion of HPV18 genes E1, E2, E4, E5, L1 and L2 (Supplementary Fig. S2).

HPV-human integration sites. A two-step strategy was applied to detect possible integration sites (Fig. 3). A total of 27 integration sites were detected in cell lines CaSki, SiHa, HeLa and MS571 (Table 2). For CaSki, 16 previously reported integration sites[30,32,37] were confirmed. In addition, three novel sites were identified. These mapped to HPV16 E6, E2 and L1 genes. One was located in an intronic region of the gene *BRSK1*; two were located more than 50 kb from annotated genes (Table 2). Three sites, including one previously reported site as a control[30,37], were subjected to Sanger sequencing to confirm the integration sites (Supplementary Table S3). Integration sites identified in SiHa, HeLa and MS751 were consistent with previous studies[31,35–39] and were not subjected to validation by Sanger sequencing. Additionally, two integration sites were detected in the clinical sample LBC105 (Table 2). The integration breakpoints were mapped to the HPV E1 and L1 genes flanking the deleted region (Supplementary Fig. S2) and they were located in intronic regions of the gene *GTF2IRD1* (Table 2). Both integration sites were confirmed by Sanger sequencing (Supplementary Table S3).

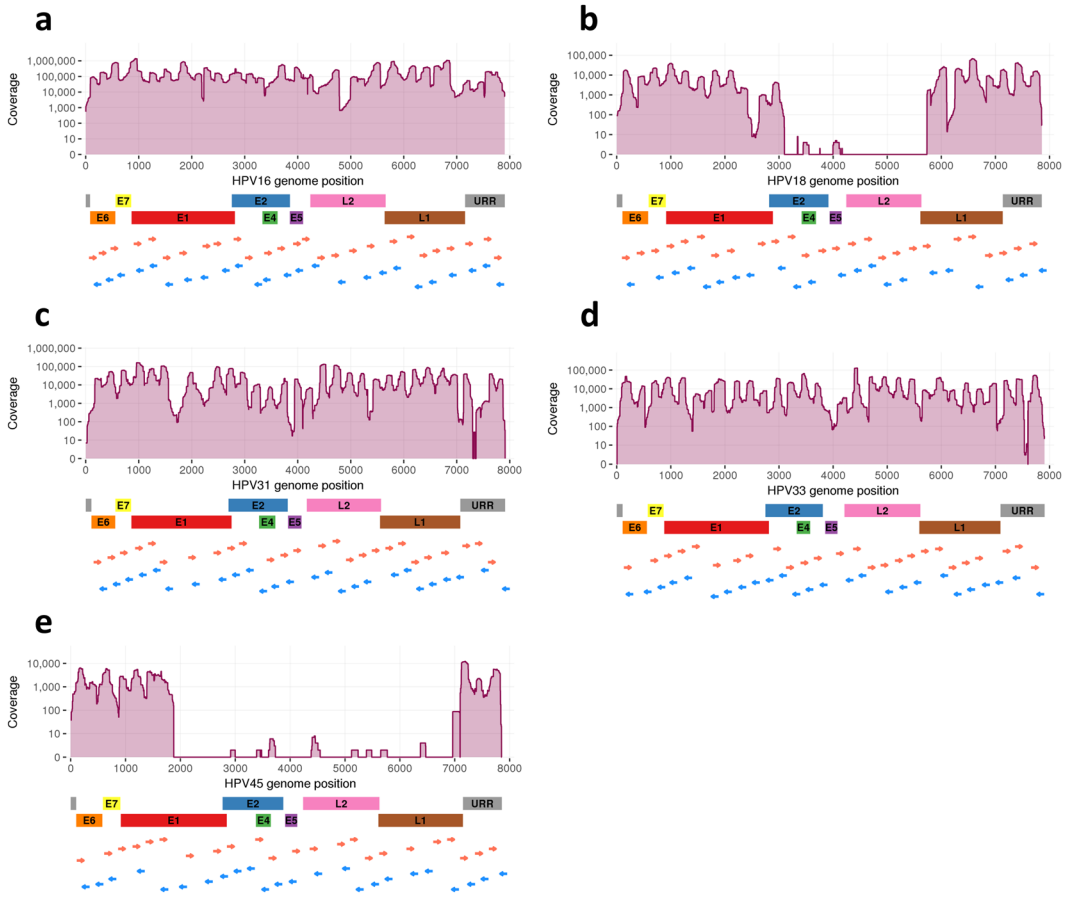

Figure 2. HPV genome sequencing coverage in HPV positive samples. The coverage plots of (**a**) CaSki, (**b**) HeLa, (**c**) LBC34, (**d**) LBC11, and (**e**) MS751 are aligned to the respective target HPV genomes. The location of early (E1, E2, E4-7), late (L1, L2) genes, URR, and forward (red arrows) and reverse (blue arrows) HPV primers is indicated below the genomic positions.

Evaluation of variant calling using SiHa technical replicates. Sequencing libraries of the SiHa cell line served as technical replicates to assess the variant calling performance. In both SiHa-1 and SiHa-2, more variable sites were detected with higher mean coverage (Fig. 4). Number of variable sites in SiHa-1 ranged from 477 to 809 and mean coverage ranged from 2554 to 17561. Number of variable sites in SiHa-2 ranged from 257 to 522 and mean coverage ranged from 646 to 5609 (Fig. 4; Supplementary Table S4). First, reproducibility of variant calling was assessed within the same SiHa sequencing library. Concordance rate of variable sites was calculated using HiSeq 2500 result as the reference value. The concordance rates varied from 92% (HiSeq down-sampled 90%) to 45% (MiSeq) in SiHa-1 and from 89% (HiSeq downsampled 90%) to 27% (MiSeq) in SiHa-2 (Supplementary Table S4). Concordance rates of variants, including low frequency variation, between replicates (different library, same sequencing platform) were calculated to evaluate the effect of library preparation steps on the number of variable sites found in each sample. Concordance rates were 21% and 19% in SiHa-1 and SiHa-2, respectively (Supplementary Table S5).

HPV genomic variability. Variability was analysed in cell lines and LBC samples. Samples had variable sites (variant allele frequency $>0.2\%$ and coverage $\geq 100\times$) in all genes with the exception of regions that were deleted or had low sequencing coverage. The number of variable sites was normalised by the length of each HPV genomic region. Genomic regions had varying percentages of variable sites (0–28%) in each of the samples. Overall, there were samples within each HPV type that had $>15\%$ variable sites in at least one HPV gene (Fig. 5). Principally, samples with higher mean coverage had more variable sites (Supplementary Table S6), which is in line with the results from the variant analysis done on SiHa replicates (Fig. 4). CaSki had most variable sites (1017) of the cell lines and LBC54 had most variable sites (1641) of the clinical samples (Supplementary Table S6). A variant profile with variable site positions and variant allele frequency (VAF) is shown for CaSki and LBC54 (Fig. 6). Overall, the results show considerable variability in the samples throughout the HPV genome (Fig. 5, Supplementary Figs S6–S10).

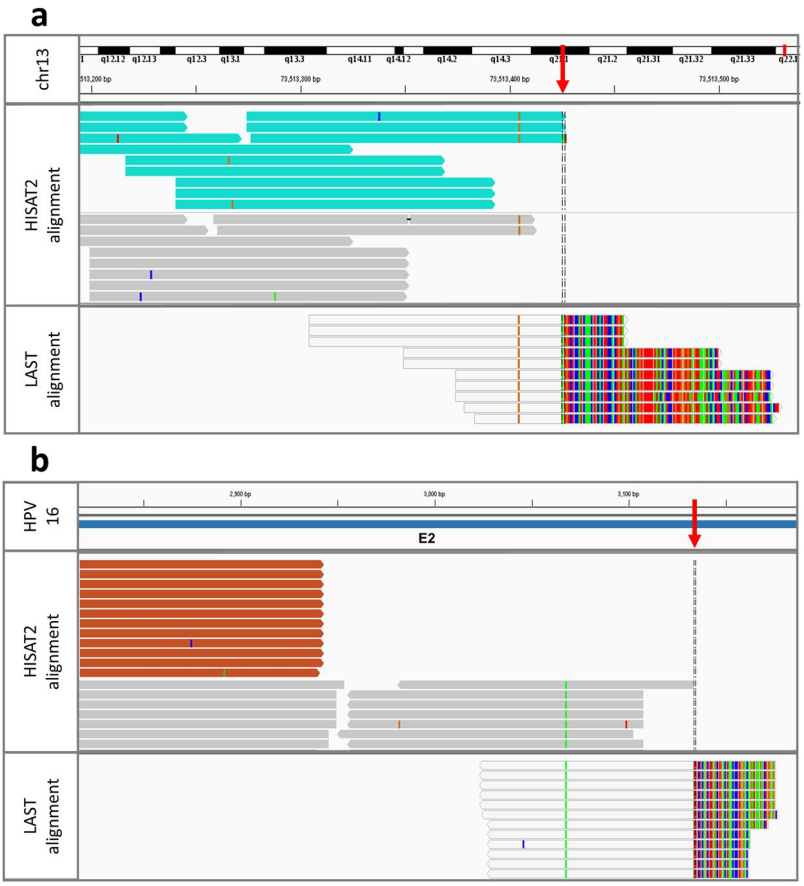

Figure 3. An IGV visualisation of HISAT2 and LAST alignments to find HPV-human integration breakpoints. All the reads were first mapped with HISAT2 and then the unmapped reads were remapped with LAST. (**a**) SiHa reads mapping to chromosome 13 (GRCh38/hg38). Light blue HISAT2 reads have pairs mapping to HPV16 reference genome. Multi-coloured parts of the LAST reads are mismatched bases that map to HPV16 (not visualised). (**b**) SiHa reads mapping to HPV16 reference genome. Orange HISAT2 reads have pairs mapping to chromosome 13 (GRCh38/hg38). Multi-coloured parts of the LAST reads are mismatched bases that map to chromosome 13 (not visualised). Red arrows point to the exact breakpoint positions.

Discussion

Here, we present a novel cost-efficient approach, TaME-seq, for the simultaneous analysis of HPV variation and chromosomal integration. Previous methods have been less effective and/or limited to either one of the two analyses[29–34]. To demonstrate the performance of TaME-seq, we employed HPV16, 18, 31, 33 and 45 positive clinical samples, HPV positive cell lines and HPV plasmids. With 47% of the total of 154.8 million raw reads mapped on the target HPV reference genomes, TaME-seq proved to be highly efficient in HPV target enrichment. Other approaches for HPV target enrichment have reported much lower HPV mapping ratios[32,40], requiring more sequencing and therefore at a higher sequencing cost. TaME-seq currently covers HPV16, 18, 31, 33 and 45, being the most common HPV genotypes in cervical cancer[5]. TaME-seq can be extended to cover additional HPV types, as well as other viruses, by implementing new primers to the method.

The ability of TaME-seq to detect chromosomal integration sites has been shown for the HPV positive cervical cancer cell lines CaSki, SiHa, HeLa and MS751. CaSki cells contain a high copy number (~600 copies/cell) of integrated full-length HPV16 arranged in concatemers[41,42]. SiHa (1–2 HPV16 copies/cell)[39,41] and HeLa (10–50 HPV18 copies/cell)[43] cells harbour integrated HPV genomes. MS751 cells contains integrated HPV45[35], but in contrast to the product specification sheet (ATCC, Manassas, VA) no HPV18, which was verified in our analyses. For CaSki, 16 previously reported integration sites[30,32,37] were detected by our method. In addition, three novel integration sites were identified. Known integration sites in SiHa[31,37,39], HeLa[31,36] and MS751[35], as well as large deletions demonstrated in HeLa[36] and MS751[35], were confirmed by the TaME-seq method. Of the 21 LBC samples, HPV integration sites could only be detected in one sample, being in line with previous studies reporting no or few HPV integration events in LSIL/ASC-US samples[44,45]. However, other studies report integration events also in LSIL samples[32,46]. The detection of integrated forms of the virus is also dependent on the amount of episomes in the sample; low copy integration sites may remain undetected against a high background of episomal HPV.

The high sequencing coverage throughout the HPV genome enables detection of low frequency variants. Variant calling was evaluated using SiHa replicates to set the variant calling threshold. Previous studies have used variant calling thresholds of 0.5% or 1%[17,34]. With the high coverage provided by the TaME-seq method there is

Sample	HPV		Human (GRCh38/hg38)		# Unique discordant read pairs	# Unique junction reads
	Breakpoint	ORF	Chromosomal locus	Breakpoint		
HPV16						
CaSki	273	E6	20p11.1	chr20:26276796	19	0[e]
	494[a]	E6	20p11.1	chr20:26341342[b]	7	0[e]
	582	E7	19q13.42	chr19:55310208	0	15
	975	E1	Xq27.3	chrX:145696778	0	7
	1398	E1	2p23.3	chr2:27135968	6	0[e]
	1793	E1	10p14	chr10:11700197	4	0[e]
	2987	E2	Xq27.3	chrX:145708231	3	8
	3239	E2	7p22.1	chr7:6925283	5	0[e]
	3631[a]	E2	19q13.42	chr19:55310043[c]	3	0[e]
	3729	E2	6p21.1	chr6:45691388	0	11
	4654	L2	11p15.4	chr11:6741077	11	0[e]
	5432	L2	11q22.1	chr11:100766632	2	0[e]
	5698	L1	10p14	chr10:11700617	20	0[e]
	5698	L1	5p11	chr5:46292081	2	0[e]
	5762	L1	11q22.1	chr11:100771699	4	0[e]
	6572	L1	19q13.42	chr19:55307445	3	0[e]
	7123[a]	L1	20p11.1	chr20:26357640[b]	20	0[e]
	7733	URR	11p15.4	chr11:6740842	2	0[e]
	7733	URR	2p23.3	chr2:27137265	6	0[e]
SiHa	3133	E2	13q22.1	chr13:73513425	7	7
	3385	E2/E4	13q22.1	chr13:73214729	3	0[e]
HPV18						
HeLa	2066	E1	8q24.21	chr8:127229053	2	0[e]
	2887	E2	8q24.21	chr8:127221122	13	0[e]
	5730	L1	8q24.21	chr8:127218384	11	89
	7655	URR	8q24.21	chr8:127221804	3	0[e]
LBC105	1561	E1	7q11.23	chr7:74525628[d]	0	10
LBC105	6528	L1	7q11.23	chr7:74515883[d]	2	0[e]
HPV45						
MS751	1646	E1	18q11.2	chr18:23024744	10	0[e]
	7120	L1	18q11.2	chr18:23021388	15	0[e]

Table 2. Chromosomal integration sites detected by TaME-seq. [a]Novel breakpoint in CaSki cell line. [b]No annotated genes within 50 kb from the breakpoint. [c]Intronic region in gene *BRSK1*. [d]Intronic region in gene *GTF2IRD1*. [e]When number of unique junction reads is 0, the breakpoint coordinates are not exact.

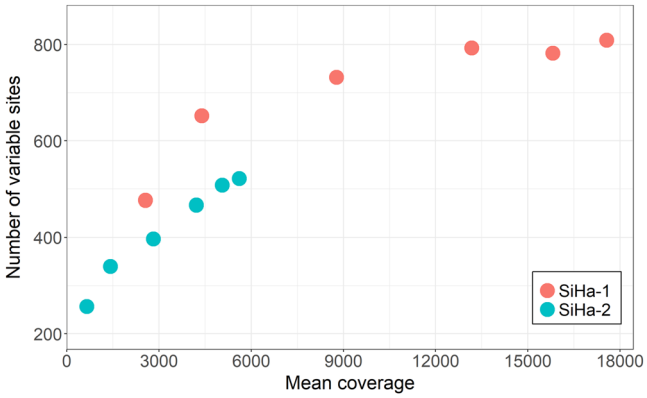

Figure 4. Number of variable sites in SiHa replicates. SiHa-1 (red dots) and SiHa-2 (blue dots) served as technical replicates to assess the variant calling performance. In SiHa libraries, sequenced on MiSeq and HiSeq 2500 platforms, increasing number of variable sites were detected with higher mean coverage.

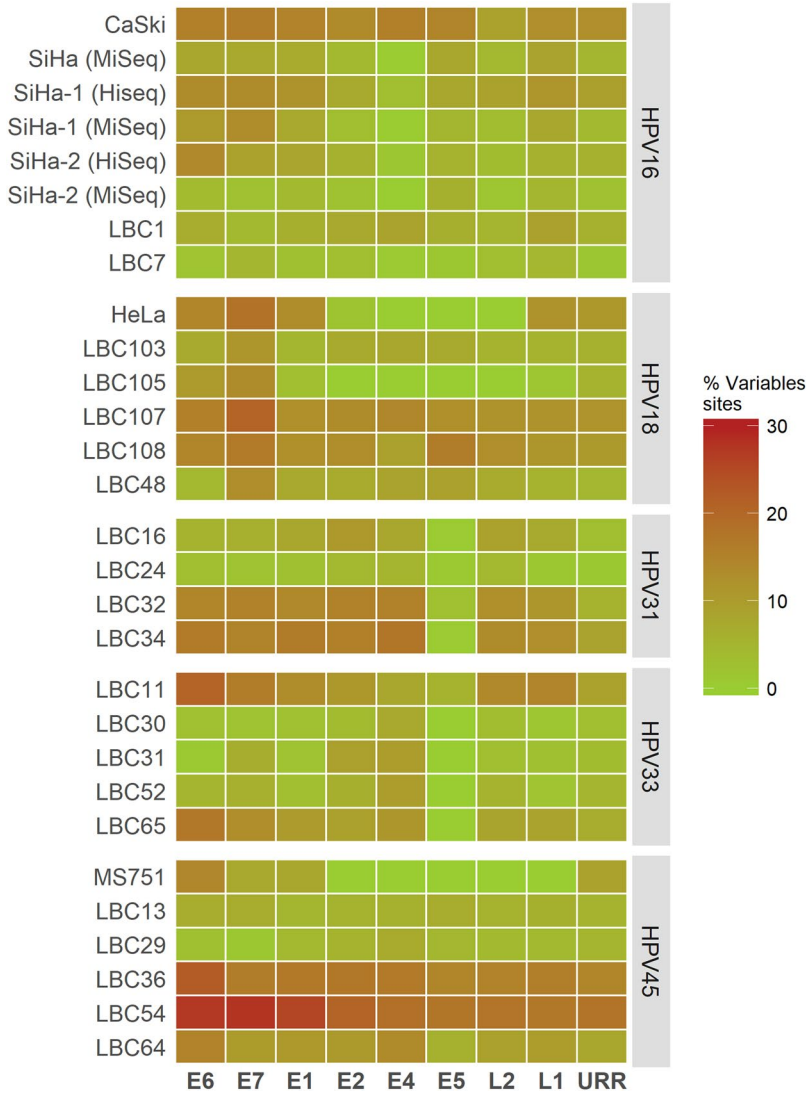

Figure 5. Proportion of variable sites in HPV genes in HPV positive samples. The number of variable sites was normalised by the length of each HPV gene. Gradient green (0% variable sites) to red (30% variable sites) color-coding of the results is shown to present the considerable variability in the samples throughout the HPV genome.

potential for detecting very low frequency variation. We have therefore analysed the variation using 0.2% as the variant calling threshold. Multiple and stringent filtering steps was included to filter out non-reliable variants, as we are approaching the inherent error rate profile of the PCR amplification and Illumina sequencing[47]. However, the threshold for variant calling is dependent on experimental and analytical basis and must be set according to the study aims.

The results from the SiHa analysis indicate that calling ultra-low frequency variants is dependent on the sequencing coverage. Lower sequencing coverage results in the detection of fewer variants and less concordance between sample replicates. In order to find ultra-low frequency variants, high sequencing coverage is required. Figure 4 shows that at the mean coverage of 12000×, the number of variants in SiHa-1 is approaching saturation. This indicates that more variants are not likely to be found even with higher sequencing coverage. Finally, differences in sequencing coverage affect the number of variable sites found, but also experimental approaches due to stochastic sampling and variant calling can fail to reveal low frequency variants. Overall, our results uncover low frequency variants in the samples, potentially introduced by DNA repair mechanisms and APOBEC enzyme mediated DNA editing[48–50], although some bias may be introduced by PCR and sequencing. Variable sites are present in all genes of the studied HPV types. Traditionally, studies have focused on sequence variation on a viral sublineage level[13–16] or the high variability has been interpreted as HPV variant co-infections[29]. The development of NGS technologies has provided comprehensive tools for the study of HPV genomic variability. Recent studies have reported high HPV variability that may be evidence of intra-host viral evolution and adaptation generated during a chronic HPV infection[17–20].

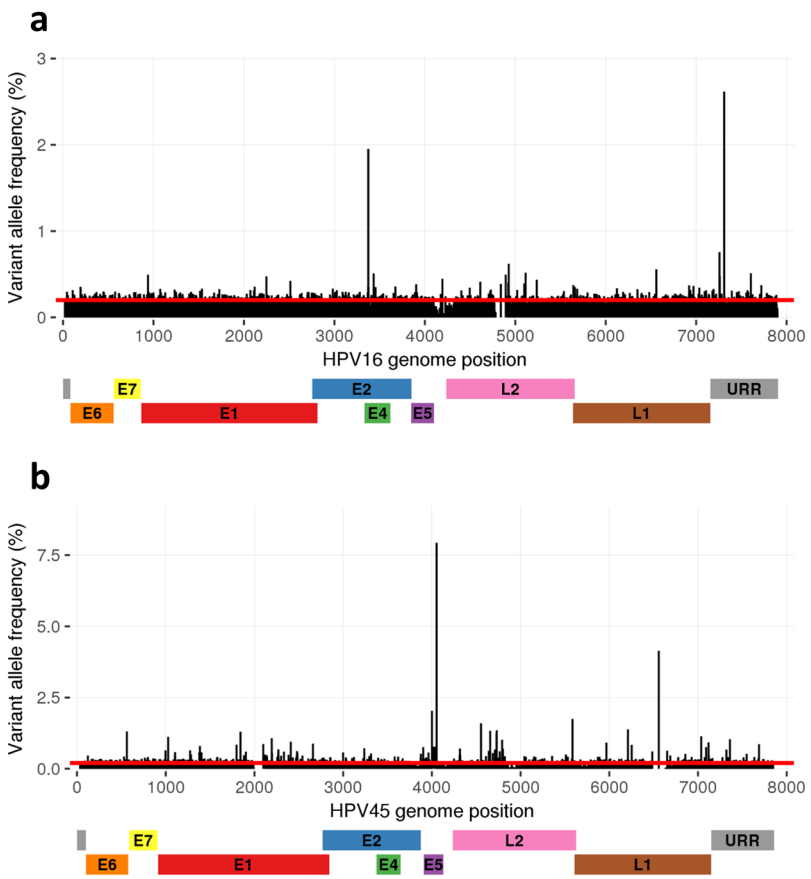

Figure 6. HPV nucleotide variation observed in two samples. The plots showing variable sites and variant allele frequency (%) in (**a**) CaSki, and (**b**) LBC54 are aligned to the respective target HPV genomes. The location of genes and URR is indicated below the genomic positions. The red line indicates the variant calling threshold value of 0.2%.

Our study has some limitations. Firstly, TaME-seq is not intended for determining HPV genotypes and we recommend it for analyses of HPV variability and integration events in samples with known HPV status. Secondly, due to variation in amplification efficacy, an uneven coverage is seen for different genomic regions. Sudden drops in the coverage, that are not genomic deletions, may be due to suboptimal primer performance or poor alignment against the reference genomes. This issue can be solved partly by designing new primers covering these regions and optimising the primer performance. Also, the read alignment step can be further optimised. Alternatively, alignment could be performed by *de novo* assembly to create consensus sequences for the alignment. Thirdly, enough viral DNA and good dsDNA quality are important for achieving consistent tagmentation results in the Nextera protocol[51]. Sample preparation of the excluded LBC samples failed likely due to very low viral load in the samples, which was not quantified separately.

In summary, we have developed a NGS approach that allows the simultaneous study of HPV genomic variability and chromosomal integration. TaME-seq is applicable to large sample cohorts due to its highly efficient target enrichment, leading to less off-target sequences and therefore reduced sequencing cost. Comprehensive studies on HPV intra-host variability generated during a persistent infection will improve our understanding of viral carcinogenesis. Efficient identification of HPV genomic variability and integration sites will be important both for the study of HPV evolution, adaptability and may be a useful tool for cervical cancer diagnostics.

Methods

Samples. Anonymised LBC samples from routine cervical cancer screening were included in the study, comprising cases of atypical squamous cells of undetermined significance (ASC-US) and low-grade squamous intraepithelial lesions (LSIL). HPV positive samples with the cobas 4800 HPV test (Roche Molecular Diagnostics, Pleasanton, CA) were extracted for DNA using the automated system NucliSENS easyMAG (BioMerieux Inc., France) with off-board lysis. The samples were HPV genotyped using the modified GP5+/6+ PCR protocol (MGP)[52], followed by HPV type-specific hybridisation using Luminex suspension array technology[53] or the Anyplex™ II HPV28 assay (Seegene, Inc., Seoul, Korea). LBC samples (n = 31) were positive for HPV16, 18, 31, 33 or 45 alone, or had multiple infections including at least one of the five types. DNA extracted from the HPV positive cervical carcinoma cell lines CaSki, SiHa, HeLa and MS751 (ATCC, Manassas, VA) served as positive controls. WHO international standards for HPV 16 (1st WHO International Standard for Human Papillomavirus Type 16 DNA, NIBSC code: 06/202) and 18 (1st WHO International Standard for Human Papillomavirus Type 18

DNA, NIBSC code: 06/206)(NIBSC, Potters Bar, Hertfordshire, UK) and a plasmid containing the strain HPV33[54] were used as additional positive controls. Laboratory-grade water and DNA from an HPV negative human sample were included as negative controls. DNA was quantified by the fluorescence-based Qubit dsDNA HS assay (Thermo Fisher Scientific Inc.,Waltham, MA, USA).

Primer design. HPV16, 18, 31, 33, and 45 whole genome reference and variant sequences were obtained from the PapillomaVirus Episteme (PaVE) database[55]. All the available reference and variant sequences within an HPV type were aligned using the multiple sequence alignment tool ClustalO[56]. The sequence alignment was converted to a consensus sequence for each HPV type in CLC Sequence viewer version 7.7.1 (QIAGEN Aarhus A/S). TaME-seq HPV primers were designed using Primer3[57] and HPV consensus sequences as the source sequence. Finally, primers were modified by adding an Illumina TruSeq-compatible adapter tail (5′-AGACGTGTGCTCTTCCGATCT-3′) to the 5′-end and then synthesised by Thermo Fisher Scientific, Inc. (Waltham, MA).

Library preparation and sequencing. Primer pools for each HPV type were prepared by combining primers separately in equal volumes. Samples were subjected to tagmentation using Nextera DNA library prep kit (Illumina, Inc., San Diego, CA). Tagmented DNA was purified using DNA Clean & Concentrator™-5 columns (Zymo Research, Irvine, CA) according the manufacturer's instructions or ZR-96 DNA Clean & Concentrator™-5 plates (Zymo Research, Irvine, CA) according to the Nextera® DNA Library Prep Reference Guide (15027987 v01) before PCR amplification for target enrichment. Amplification was performed using Qiagen Multiplex PCR Master mix (Qiagen, Hilden, Germany) according to the manufacturer's instructions. For each sample, two PCR reactions were performed separately with 0.75 µM of HPV primer pools, 0.5 µM of i7 index primers (adapted from Kozich *et al.*[58]) and 1 µl of i5 index primers from the Nextera index kit (Illumina, Inc., San Diego, CA). The cycling conditions were as follows: initial denaturation and hot start at 95 °C for 5 minutes; 30 cycles at 95 °C for 30 seconds, at 58 °C for 90 seconds and at 72 °C for 20 seconds; final extension at 68 °C for 10 minutes. Following amplification, libraries were pooled in equal volumes and the final sample pool was purified with Agencourt® AMPure® XP beads (Beckman Coulter, Brea, CA). The quality and quantity of the pooled libraries were assessed on Agilent 2100 Bioanalyzer using Agilent High Sensitivity DNA Kit (Agilent Technologies Inc., Santa Clara, CA) and by qPCR using KAPA DNA library quantification kit (Kapa Biosystems, Wilmington, MA). Sequencing was performed on the MiSeq platform (Illumina, Inc., San Diego, CA) or on the HiSeq 2500 platform (Illumina, Inc., San Diego, CA). Samples were sequenced as 151 bp paired-end reads and two 8 bp index reads.

Sequence alignment. Raw paired-end reads were trimmed for adapters, HPV primers, quality (-q 20) and finally for minimum length (-m 50) using cutadapt (v1.10)[59]. Trimmed reads were mapped to human (GRCh38/hg38) and HPV16, 18, 31, 33 and 45 reference genomes obtained from the PaVE database[55] using HISAT2 (v2.1.0)[60]. Mapping statistics and sequencing coverage were calculated using the Pysam package[61] with an in-house Python (v3.5.4) script. Downstream analysis was performed using an in-house R (v3.4.4) script. Results from both reactions of the same sample were combined and method performance was then evaluated based on the percentage of obtained reads mapped to the HPV reference genome, mean sequencing coverage and percentage of HPV reference genome coverage for each sample. Further analysis was performed when a sample had >20000 reads mapped to the target HPV reference genome. The target HPV genomes correspond to the HPV types for which the samples were reported positive by HPV genotyping.

Detecting HPV-human integration sites. The paired-end reads that mapped (HISAT2) with one end to a human chromosome and the other end to the target HPV reference genome were identified as discordant read pairs. If a specific position had ≥2 read pairs with unique start or end coordinates, it was considered as a potential integration site. To determine the exact position of HPV-human integration breakpoints, previously unmapped reads were remapped to human and HPV reference genomes (as above) using the LAST (v876) aligner (options -M -C2)[62]. Positions covered by ≥3 junction reads, with unique start or end coordinates, were considered as potential integration breakpoints. Integration site detection was not based on reads sharing the same start and end coordinates as these reads were considered as potential PCR duplicates. Selected HPV integration breakpoints were confirmed by PCR amplification and Sanger sequencing.

Sequence variation analysis. Mapped nucleotide counts over HPV reference genomes and average mapping quality values of each nucleotide were retrieved from BAM files and variant calling was performed using an in-house R script. To reduce the effects of PCR amplification and sequencing artefacts in the variation analysis, filtering was applied before the variant calling. Nucleotides seen ≤2 times in each position and nucleotides with mean Phred quality score of <20 were filtered out. Nucleotide counts from both reactions of the same sample were combined and variant allele frequencies (VAF) of the three minor alleles in each position were calculated. If results from either of the reaction showed >5 times larger VAF with <20% of the total coverage, it was discarded from variant calling. Finally, variants were called if VAF was >0.2% and coverage was ≥100×.

Two sequencing libraries of SiHa cell line served as technical replicates to assess the variant calling performance. The technical replicates were sequenced on the MiSeq platform or on the HiSeq 2500 platform. In addition, HiSeq raw sequencing data was downsampled randomly and defined portions (90%, 75%, 50% and 25%) of the original reads were further analysed. Reproducibility of calling variants in the replicates was assessed by calculating concordance rate. The concordance rate (R_c) between duplicates was defined as follows:

$$R_c = \frac{N_c}{mean(N_1, N_2)}$$

where N_c was the number of concordant variants between a pair of replicate samples, and N_1 and N_2 were the total number of variants detected in each of the duplicated sample.

Ethical approval. This study was approved by the regional committee for medical and health research ethics, Oslo, Norway [2017/447] and we confirm that all experiments were performed in accordance with the committee's guidelines and regulations.

Data Availability

Sequence data from cell lines will be available at European Nucleotide Archive (ENA) accession number ERP111061. Plasmids are third party property and requests must be made to International Human Papillomavirus Reference Center and Institut Pasteur. Sequencing data from clinical samples will be available from the authors upon request with obtained ethical approval. Clinical sequence data may be deposited at the European Genome-phenome Archive (EGA) (ethical and legal assessments are on-going).

References

1. Walboomers, J. M. *et al*. Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *J. Pathol.* **189**, 12–19, 10.1002/(sici)1096-9896(199909)189:1<12::aid-path431>3.0.co;2-f (1999).
2. Ferlay, J. *et al*. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **136**, E359–386, https://doi.org/10.1002/ijc.29210 (2015).
3. Fitzmaurice, C. *et al*. The Global Burden of Cancer 2013. *JAMA Oncol* **1**, 505–527, https://doi.org/10.1001/jamaoncol.2015.0735 (2015).
4. Bosch, F. X., Lorincz, A., Munoz, N., Meijer, C. J. & Shah, K. V. The causal relation between human papillomavirus and cervical cancer. *J. Clin. Pathol.* **55**, 244–265 (2002).
5. de Sanjose, S. *et al*. Human papillomavirus genotype attribution in invasive cervical cancer: a retrospective cross-sectional worldwide study. *The Lancet Oncology* **11**, 1048–1056, https://doi.org/10.1016/s1470-2045(10)70230-8 (2010).
6. Crosbie, E. J., Einstein, M. H., Franceschi, S. & Kitchener, H. C. Human papillomavirus and cervical cancer. *The Lancet* **382**, 889–899, https://doi.org/10.1016/s0140-6736(13)60022-7 (2013).
7. Forman, D. *et al*. Global burden of human papillomavirus and related diseases. *Vaccine* **30**(Suppl 5), F12–23, https://doi.org/10.1016/j.vaccine.2012.07.055 (2012).
8. Moscicki, A. B. *et al*. Updating the natural history of human papillomavirus and anogenital cancers. *Vaccine* **30**(Suppl 5), F24–33, https://doi.org/10.1016/j.vaccine.2012.05.089 (2012).
9. Bernard, H. U. Taxonomy and phylogeny of papillomaviruses: an overview and recent developments. *Infect. Genet. Evol.* **18**, 357–361, https://doi.org/10.1016/j.meegid.2013.03.011 (2013).
10. Bzhalava, D., Eklund, C. & Dillner, J. International standardization and classification of human papillomavirus types. *Virology* **476**, 341–344, https://doi.org/10.1016/j.virol.2014.12.028 (2015).
11. Bernard, H. U. *et al*. Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology* **401**, 70–79, https://doi.org/10.1016/j.virol.2010.02.002 (2010).
12. Burk, R. D., Harari, A. & Chen, Z. Human papillomavirus genome variants. *Virology* **445**, 232–243, https://doi.org/10.1016/j.virol.2013.07.018 (2013).
13. Cornet, I. *et al*. HPV16 genetic variation and the development of cervical cancer worldwide. *Br. J. Cancer* **108**, 240–244, https://doi.org/10.1038/bjc.2012.508 (2013).
14. Mirabello, L. *et al*. HPV16 Sublineage Associations With Histology-Specific Cancer Risk Using HPV Whole-Genome Sequences in 3200 Women. *J. Natl. Cancer Inst.* **108**, https://doi.org/10.1093/jnci/djw100 (2016).
15. Chan, P. K. *et al*. Geographical distribution and oncogenic risk association of human papillomavirus type 58 E6 and E7 sequence variations. *Int. J. Cancer* **132**, 2528–2536, https://doi.org/10.1002/ijc.27932 (2013).
16. Chen, A. A., Gheit, T., Franceschi, S., Tommasino, M. & Clifford, G. M. Human Papillomavirus 18 Genetic Variation and Cervical Cancer Risk Worldwide. *J. Virol.* **89**, 10680–10687, https://doi.org/10.1128/jvi.01747-15 (2015).
17. de Oliveira, C. M. *et al*. High-level of viral genomic diversity in cervical cancers: A Brazilian study on human papillomavirus type 16. *Infect. Genet. Evol.* **34**, 44–51, https://doi.org/10.1016/j.meegid.2015.07.002 (2015).
18. Mirabello, L. *et al*. HPV16 E7 Genetic Conservation Is Critical to Carcinogenesis. *Cell* **170**, 1164–1174 e1166, https://doi.org/10.1016/j.cell.2017.08.001 (2017).
19. Hirose, Y. *et al*. Within-Host Variations of Human Papillomavirus Reveal APOBEC-Signature Mutagenesis in the Viral Genome. *J. Virol.* https://doi.org/10.1128/jvi.00017-18 (2018).
20. Dube Mandishora, R. S. *et al*. Intra-host sequence variability in human papillomavirus. *Papillomavirus Res*, https://doi.org/10.1016/j.pvr.2018.04.006 (2018).
21. Zur Hausen, H. Papillomaviruses and cancer: from basic studies to clinical application. *Nat. Rev. Cancer* **2**, 342–350, https://doi.org/10.1038/nrc798 (2002).
22. Pett, M. & Coleman, N. Integration of high-risk human papillomavirus: a key event in cervical carcinogenesis? *J. Pathol.* **212**, 356–367, https://doi.org/10.1002/path.2192 (2007).
23. McBride, A. A. & Warburton, A. The role of integration in oncogenic progression of HPV-associated cancers. *PLoS Pathog.* **13**, e1006211, https://doi.org/10.1371/journal.ppat.1006211 (2017).
24. Jeon, S., Allen-Hoffmann, B. L. & Lambert, P. F. Integration of human papillomavirus type 16 into the human genome correlates with a selective growth advantage of cells. *J. Virol.* **69**, 2989–2997 (1995).
25. Doorbar, J., Egawa, N., Griffin, H., Kranjec, C. & Murakami, I. Human papillomavirus molecular biology and disease association. *Rev. Med. Virol.* **25**(Suppl 1), 2–23, https://doi.org/10.1002/rmv.1822 (2015).
26. Ziegert, C. *et al*. A comprehensive analysis of HPV integration loci in anogenital lesions combining transcript and genome-based amplification techniques. *Oncogene* **22**, 3977–3984, https://doi.org/10.1038/sj.onc.1206629 (2003).
27. Peter, M. *et al*. Frequent genomic structural alterations at HPV insertion sites in cervical carcinoma. *J. Pathol.* **221**, 320–330, https://doi.org/10.1002/path.2713 (2010).
28. Kraus, I. *et al*. The Majority of Viral-Cellular Fusion Transcripts in Cervical Carcinomas Cotranscribe Cellular Sequences of Known or Predicted Genes. *Cancer Res.* **68**, 2514–2522, https://doi.org/10.1158/0008-5472.Can-07-2776 (2008).
29. Cullen, M. *et al*. 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, https://doi.org/10.1016/j.pvr.2015.05.004 (2015).
30. Xu, B. *et al*. Multiplex Identification of Human Papillomavirus 16 DNA Integration Sites in Cervical Carcinomas. *PLoS One* **8**, e66693, https://doi.org/10.1371/journal.pone.0066693 (2013).
31. Liu, Y., Lu, Z., Xu, R. & Ke, Y. Comprehensive mapping of the human papillomavirus (HPV) DNA integration sites in cervical carcinomas by HPV capture technology. *Oncotarget* **7**, 5852–5864, https://doi.org/10.18632/oncotarget.6809 (2016).

32. Hu, Z. *et al*. Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism. *Nat. Genet.* **47**, 158–163, https://doi.org/10.1038/ng.3178 (2015).
33. Holmes, A. *et al*. Mechanistic signatures of HPV insertions in cervical carcinomas. *npj Genomic Medicine* **1**, https://doi.org/10.1038/npjgenmed.2016.4 (2016).
34. Kukimoto, I. *et al*. Genetic variation of human papillomavirus type 16 in individual clinical specimens revealed by deep sequencing. *PLoS One* **8**, e80583, https://doi.org/10.1371/journal.pone.0080583 (2013).
35. Geisbill, J., Osmers, U. & Durst, M. Detection and characterization of human papillomavirus type 45 DNA in the cervical carcinoma cell line MS751. *J. Gen. Virol.* **78**(Pt 3), 655–658, https://doi.org/10.1099/0022-1317-78-3-655 (1997).
36. Adey, A. *et al*. The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* **500**, 207–211, https://doi.org/10.1038/nature12064 (2013).
37. Akagi, K. *et al*. Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. *Genome Res.* **24**, 185–199, https://doi.org/10.1101/gr.164806.113 (2014).
38. Mincheva, A., Gissmann, L. & zur Hausen, H. Chromosomal integration sites of human papillomavirus DNA in three cervical cancer cell lines mapped by *in situ* hybridization. *Med. Microbiol. Immunol.* **176**, 245–256 (1987).
39. el Awady, M. K., Kaplan, J. B., O'Brien, S. J. & Burk, R. D. Molecular analysis of integrated human papillomavirus 16 sequences in the cervical cancer cell line SiHa. *Virology* **159**, 389–398 (1987).
40. Li, T. *et al*. Universal Human Papillomavirus Typing Assay: Whole-Genome Sequencing following Target Enrichment. *J. Clin. Microbiol.* **55**, 811–823, https://doi.org/10.1128/JCM.02132-16 (2017).
41. Baker, C. C. *et al*. Structural and transcriptional analysis of human papillomavirus type 16 sequences in cervical carcinoma cell lines. *J. Virol.* **61**, 962–971 (1987).
42. Yee, C., Krishnan-Hewlett, I., Baker, C. C., Schlegel, R. & Howley, P. M. Presence and expression of human papillomavirus sequences in human cervical carcinoma cell lines. *Am. J. Pathol.* **119**, 361–366 (1985).
43. Meissner, J. D. Nucleotide sequences and further characterization of human papillomavirus DNA present in the CaSki, SiHa and HeLa cervical carcinoma cell lines. *J. Gen. Virol.* **80**(Pt 7), 1725–1733, https://doi.org/10.1099/0022-1317-80-7-1725 (1999).
44. Hudelist, G. *et al*. Physical state and expression of HPV DNA in benign and dysplastic cervical tissue: different levels of viral integration are correlated with lesion grade. *Gynecol. Oncol.* **92**, 873–880, https://doi.org/10.1016/j.ygyno.2003.11.035 (2004).
45. Liu, Y. *et al*. Genome-wide profiling of the human papillomavirus DNA integration in cervical intraepithelial neoplasia and normal cervical epithelium by HPV capture technology. *Sci. Rep.* **6**, 35427, https://doi.org/10.1038/srep35427 (2016).
46. Li, H. *et al*. Preferential sites for the integration and disruption of human papillomavirus 16 in cervical lesions. *J. Clin. Virol.* **56**, 342–347, https://doi.org/10.1016/j.jcv.2012.12.014 (2013).
47. Schirmer, M., D'Amore, R., Ijaz, U. Z., Hall, N. & Quince, C. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics* **17**, 125, https://doi.org/10.1186/s12859-016-0976-y (2016).
48. Warren, C. J. *et al*. APOBEC3A functions as a restriction factor of human papillomavirus. *J. Virol.* **89**, 688–702, https://doi.org/10.1128/JVI.02383-14 (2015).
49. Kukimoto, I. *et al*. Hypermutation in the E2 gene of human papillomavirus type 16 in cervical intraepithelial neoplasia. *J. Med. Virol.* **87**, 1754–1760, https://doi.org/10.1002/jmv.24215 (2015).
50. Chen, J. & Furano, A. V. Breaking bad: The mutagenic effect of DNA repair. *DNA Repair (Amst)* **32**, 43–51, https://doi.org/10.1016/j.dnarep.2015.04.012 (2015).
51. Lamble, S. *et al*. Improved workflows for high throughput library preparation using the transposome-based Nextera system. *BMC Biotechnol.* **13**, 104, https://doi.org/10.1186/1472-6750-13-104 (2013).
52. Soderlund-Strand, A., Carlson, J. & Dillner, J. Modified general primer PCR system for sensitive detection of multiple types of oncogenic human papillomavirus. *J. Clin. Microbiol.* **47**, 541–546, https://doi.org/10.1128/JCM.02007-08 (2009).
53. Schmitt, M. *et al*. Bead-based multiplex genotyping of human papillomaviruses. *J. Clin. Microbiol.* **44**, 504–512, https://doi.org/10.1128/JCM.44.2.504-512.2006 (2006).
54. Beaudenon, S. *et al*. A novel type of human papillomavirus associated with genital neoplasias. *Nature* **321**, 246–249, https://doi.org/10.1038/321246a0 (1986).
55. Van Doorslaer, K. *et al*. The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. *Nucleic Acids Res.* **41**, D571–578, https://doi.org/10.1093/nar/gks984 (2013).
56. Sievers, F. *et al*. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539, https://doi.org/10.1038/msb.2011.75 (2011).
57. Untergasser, A. *et al*. Primer3–new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115, https://doi.org/10.1093/nar/gks596 (2012).
58. Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. 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**, 5112–5120, https://doi.org/10.1128/AEM.01043-13 (2013).
59. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. **2011** 17, https://doi.org/10.14806/ej.17.1.200 (2011).
60. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357–360, https://doi.org/10.1038/nmeth.3317 (2015).
61. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079, https://doi.org/10.1093/bioinformatics/btp352 (2009).
62. Kielbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**, 487–493, https://doi.org/10.1101/gr.113985.110 (2011).

Acknowledgements

We thank Mona Hansen and Hanne Kristiansen-Haugland for DNA sample extraction and HPV genotyping, and Tobias Neidel for primer design for HPV31, 33 and 45. This work was funded by a grant from South-Eastern Norway Regional Health Authority (project number 2016020).

Author Contributions

S.L. designed primers, performed the experiments, analysed the results and drafted the manuscript text. S.U.U. contributed to the data analysis. M.L. and P.E. performed the pilot experiments and P.E. designed the initial TaME-seq assay concept. R.M. contributed to the primer design process and designed primers. I.K.C. and O.H.A. contributed to study design and result interpretation. T.B.R. contributed to the study design, data analysis and result interpretation. All authors contributed to writing, reading and approving the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-36669-6.

Competing Interests: S.L., M.L., P.E., R.M., I.K.C., O.H.A. and T.B.R. and their corresponding institutions have filed a patent application at the technology transfer company Inven2, Oslo, Norway on the protocol described here.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

II

Journal of Clinical Virology 119 (2019) 24–30

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 $> 100\times$. 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),

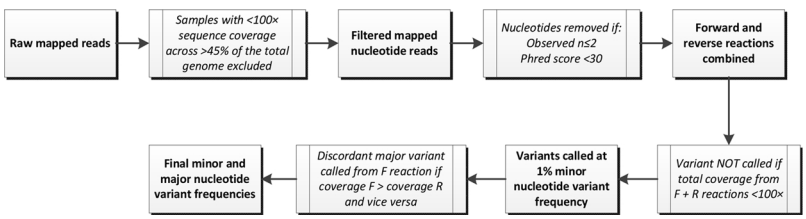

Fig. 1. Schematic representation of the nucleotide variant calling.

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

A total of 1717 HPV16 MNVs (variant frequency > 1% and coverage ≥100×) 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

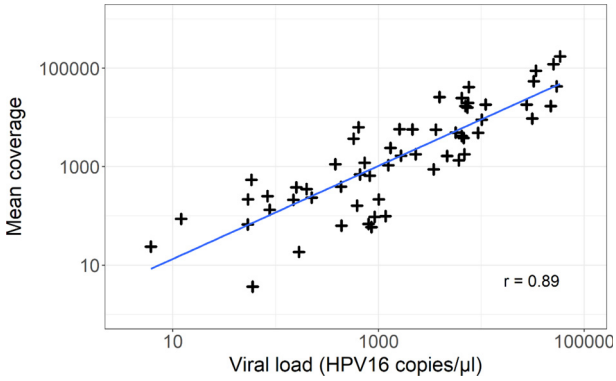

Fig. 3. Correlation between mean sequencing coverage and viral load (HPV16 copies/µL) in each sample.

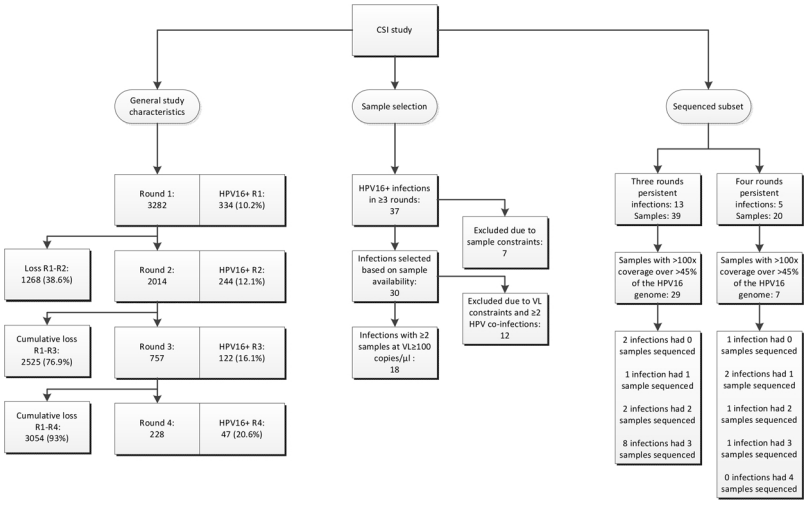

Fig. 2. Study flowchart describing selected samples and sequencing outcome. VL = viral load.

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

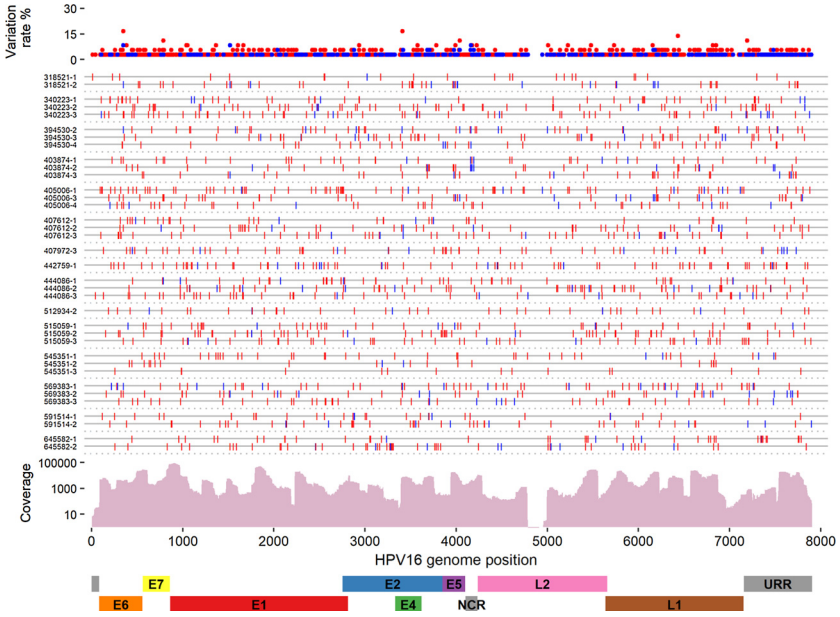

Fig. 4. Variable sites (n = 1717) and mean sequencing coverage in the 36 samples from 15 individuals. Variation rate (top) shows the amount of samples (in %) carrying a minor nucleotide variant in each position. Each horizontal line represents an individual sample, which is named according to case number 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.

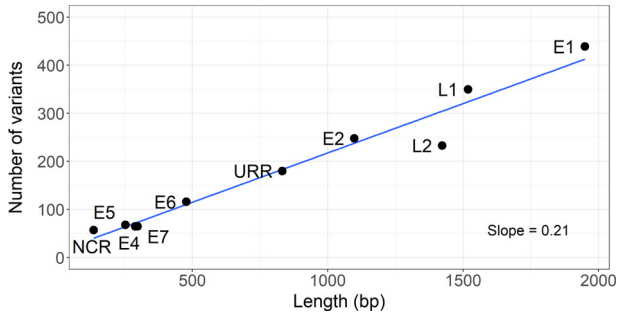

Fig. 5. Correlation between the total number of minor nucleotide variants (MNVs) and the length of viral gene regions including URR and NCR. Since certain genes overlap, MNVs can be counted more than once.

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

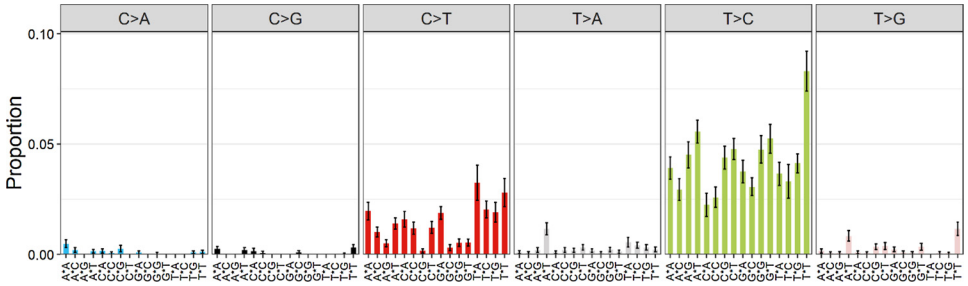

Fig. 6. Overall mutational signatures of 1717 minor nucleotide variants (MNVs) in 36 samples. Mutations are classified into six base substitutions and further into 96 trinucleotide substitution types. Mean proportion of each 96 mutational signature was calculated in the samples. Error bars represent 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 $\geq 100\times$ 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.

Journal of Clinical Virology 124 (2020) 104286

Contents lists available at ScienceDirect

Journal of Clinical Virology

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

Erratum

Erratum to "HPV16 whole genome minority variants in persistent infections from young Dutch women" [J. Clin. Virol. 119 (2019) 24–30]

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

The publisher regrets that there was a typing error in the second paragraph of the section entitled "3.3. HPV16 minor nucleotide variations".

The correct text should read:

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 1.67 times (STDEV ± 0.19) more non-synonymous than synonymous mutations. No genomic region could be singled out as notably different from other regions.

The incorrect text had previously read:

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 publisher would like to apologise for any inconvenience caused.

DOI of original article: https://doi.org/10.1016/j.jcv.2019.08.003
* 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.2020.104286

ERRATA LIST

Name of candidate: Sonja Lagström
Title of thesis: Characterisation of human papillomavirus genomic variation and chromosomal integration in cervical samples

Abbreviations: Cor – correction of language
Cpltf – change of page layout or text format
Date: 23 September 2020

Page	Line	Original text	Correction type	Corrected text
3	5	polypetide-like	Cor	polypeptide-like
12	8	(Figure 4)	Cpltf	Font format
12	19	[56]m	Cor	[56]
14	16	Most high-risk	Cor	High-risk
16	17	surface of epithelial	Cor	surface of epithelium
34	15	<45% genome	Cor	<45% of the genome
37	20	(Appendix 1: Patent application)	Cpltf	Font format
41	8	the sequencing depth	Cor	the mean sequencing depth
47	3	The Kruskal-Wallis is	Cor	The Kruskal-Wallis test is
48	23	minor variation	Cor	minor nucleotide variation

III

12 (2021) 200221

Contents lists available at ScienceDirect

Tumour Virus Research

journal homepage: www.journals.elsevier.com/tumour-virus-research

HPV16 and HPV18 type-specific APOBEC3 and integration profiles in different diagnostic categories of cervical samples

Sonja Lagström [a,b,c], Alexander Hesselberg Løvestad [d], Sinan Uğur Umu [b], Ole Herman Ambur [d], Mari Nygård [b], Trine B. Rounge [b,e,**,1], Irene Kraus Christiansen [a,f,*,1]

[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] Faculty of Health Sciences, OsloMet, Oslo Metropolitan University, Oslo, Norway
[e] Department of Informatics, University of Oslo, Oslo, Norway
[f] Department of Clinical Molecular Biology (EpiGen), Division of Medicine, Akershus University Hospital and University of Oslo, Lørenskog, Norway

ARTICLE INFO

Keywords:
Human papillomavirus
Minor nucleotide variation
APOBEC3
Chromosomal integration
Viral genomic deletion

ABSTRACT

Human papillomavirus (HPV) 16 and 18 are the most predominant types in cervical cancer. Only a small fraction of HPV infections progress to cancer, indicating that additional factors and genomic events contribute to the carcinogenesis, such as minor nucleotide variation caused by APOBEC3 and chromosomal integration.

We analysed intra-host minor nucleotide variants (MNVs) and integration in HPV16 and HPV18 positive cervical samples with different morphology. Samples were sequenced using an HPV whole genome sequencing protocol TaME-seq. A total of 80 HPV16 and 51 HPV18 positive samples passed the sequencing depth criteria of 300× reads, showing the following distribution: non-progressive disease (HPV16 n = 21, HPV18 n = 12); cervical intraepithelial neoplasia (CIN) grade 2 (HPV16 n = 27, HPV18 n = 9); CIN3/adenocarcinoma *in situ* (AIS) (HPV16 n = 27, HPV18 n = 30); cervical cancer (HPV16 n = 5).

Similar numbers of MNVs in HPV16 and HPV18 samples were observed for most viral genes, with the exception of HPV18 E4 with higher numbers across clinical categories. APOBEC3 signatures were observed in HPV16 lesions, while similar mutation patterns were not detected for HPV18. The proportion of samples with integration was 13% for HPV16 and 59% for HPV18 positive samples, with a noticeable portion located within or close to cancer-related genes.

1. Introduction

A persistent infection with one of the carcinogenic HPV genotypes is accepted as a necessary cause of cervical cancer development [1]. Of the 12 carcinogenic types [2], HPV16 and HPV18 are associated with about 70% of all cervical cancers [3]. HPV16 is predominantly associated with squamous cell carcinomas (SCC), while HPV18 is more often detected in adenocarcinomas [3], suggesting that these HPV types differ in their target cell specificity [4]. Nevertheless, only a small fraction of HPV infections will persist and progress to cancer [5], indicating that additional factors and genomic events are necessary for the HPV-induced carcinogenic process.

The 7.9 kb double stranded HPV DNA genome consists of early region (E1, E2, E4-7) genes, late region (L1, L2) genes, an upstream regulatory region (URR) and a short non-coding region (NCR) between the genes E5 and L2 [6,7]. To date, more than 200 HPV genotypes have been identified, based on at least 10% difference within the conserved L1 gene sequence [8]. HPV types harbouring minor genetic variation are grouped into lineages (1–10% whole genome nucleotide difference) and sublineages (0.5–1.0% difference) [9]. HPV evolve slowly partly since the HPV genome replication is dependent on host cell high-fidelity polymerases [10]. However, recent studies have revealed variability below the level of HPV sublineages. These are non-lineage genetic variants, which may at low frequencies indicate intra-host viral diversification and evolution [11–13].

* Corresponding author. Department of Microbiology and Infection Control, Akershus University Hospital, Lørenskog, Norway.
** Corresponding author. Department of Research, Cancer Registry of Norway, Oslo, Norway.
E-mail addresses: trine.rounge@kreftregisteret.no (T.B. Rounge), irene.kraus.christiansen@ahus.no (I.K. Christiansen).
[1] Equal contribution.

https://doi.org/10.1016/j.tvr.2021.200221
Received 19 November 2020; Received in revised form 9 April 2021; Accepted 7 June 2021
Available online 25 June 2021

Abbreviations

AID	activation-induced cytidine deaminase
AIS	adenocarcinoma *in situ*
ASC-US	atypical squamous cells of undetermined significance
CIN	cervical intraepithelial neoplasia
dN/dS	ratio of non-synonymous to synonymous substitutions
HPV	human papillomavirus
LSIL	low-grade squamous intraepithelial lesion
MNV	minor nucleotide variant
NCR	non-coding region
ncRNA	non-coding RNA
NGS	next-generation sequencing
SCC	squamous cell carcinoma
URR	upstream regulatory region
UTR	untranslated region

The generation of viral genetic variants is caused by various stochastic or targeted mutagenic processes [14]. One of the targeted mechanisms suggested to cause MNVs and impact HPV mutational drift involves the anti-viral host-defence enzyme apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like 3 (APOBEC3) proteins [15]. APOBEC3 proteins are cytidine deaminases causing deoxycytidine (C) to deoxythymidine (T) mutations during viral replication. The mutations can lead to defects in viral genome replication necessary for the viral life cycle [16]. APOBEC3 mutational signatures have been found in the human genome in cervical cancers [17], as well as in HPV genomes in cervical pre-cancerous and cancer samples [11,18,19], and has recently been associated with viral clearance [20]. APOBEC3A may function as a HPV restriction factor [15] and APOBEC3B has been shown to be upregulated by HPV [21]. The two enzymes APOBEC3A and APOBEC3B display preference for the motifs YTCA (Y = pyrimidine) and RTCA (R = purine), respectively [22]. Findings of hypovariability of the E7 gene suggest negative selection opposite of APOBEC3-related editing and an essential gene conservation for progression to cancer [23,24].

HPV integration into the host genome is regarded as a driving event in cervical carcinogenesis and is observed in >80% of HPV-induced cancers [25]. Integrations causing disruption or complete deletion of the E1 or E2 gene result in constitutive expression of the viral E6 and E7 oncogenes [26], leading to inactivation of cell cycle checkpoints and genomic instability [27]. Integration may also lead to disruption of host genes, such as tumour-suppressor genes or negative regulators of oncogenes, modified expression of adjacent genes, as well as other genomic alterations, which may promote HPV-induced carcinogenesis [28–30]. In high-grade lesions and cancers, integrations in certain chromosomal loci, including loci 3q28, 8q24.21 and 13q22.1, have been reported more often than in other loci [31], suggesting selective growth advantages for cells with site-specific integrations in e.g. important regulatory genes. Increasing integration frequencies have been reported upon comparison of cervical precancerous and cancer lesions [32,33].

Recently, we developed a novel next-generation sequencing (NGS) strategy TaME-seq for simultaneous analysis of HPV genomic variability and chromosomal integration [34]. Employing the TaME-seq method, we have explored HPV16 and HPV18 intra-host genomic variability and integration in HPV positive cervical samples with different morphologies. Differences in HPV variability between the diagnostic categories may shed light on intra-host viral genome dynamics and evolution processes in cervical carcinogenesis. In addition, integration analysis will contribute to a better understanding of this event during HPV-induced carcinogenesis.

2. Material and methods

2.1. Sample selection

Cervical cell samples have previously been collected from women attending the cervical cancer screening program in Norway between January 2005 and April 2008. Samples were collected in ThinPrep PreservCyt solution (Hologic, Marlborough, MA) and pelleted before storage at −80 °C. The samples were stored in a research biobank at Akershus University Hospital, consisting of both the cell material and extracted DNA. Recruitment criteria and HPV detection and genotyping have been described previously [35,36]. Cytology samples were previously analysed for HPV using the Amplicor HPV DNA test (Roche Diagnostics, Switzerland) followed by genotyping by Linear Array (Roche Diagnostics, Switzerland) and PreTect HPV-Proofer (PreTect AS, Norway).

In this study, primarily DNA was used for downstream analyses; for some samples, DNA extraction had to be performed from the cell material. DNA extraction was performed using the automated NucliSENS easyMag platform (BioMerieux Inc., France) with off-board lysis. All samples in the biobank that were positive for HPV16 and/or HPV18, alone or together with other HPV types, by one or both of the genotyping methods were included in the study, with the exception of HPV16 CIN3 samples for which a random selection of 50 samples were included. In total, 157 HPV16 positive samples and 75 HPV18 positive samples were subjected to sequencing (Table 1). All samples were allocated to mutually exclusive categories based on the HPV type and the diagnostic categories of non-progressive disease, histologically confirmed cervical intraepithelial neoplasia (CIN) grade 2 (CIN2), CIN3/adenocarcinoma *in situ* (AIS) and cancer. The non-progressive disease category included samples from women with normal cytology also having normal cytology the preceding two years and with no previous history of treatment for cervical neoplasia (HPV16 n = 24, HPV18 n = 3), and samples from women with atypical squamous cells of undetermined significance (ASC-US) or low-grade squamous intraepithelial lesions (LSIL) with no follow-up diagnosis within four years subsequent to the diagnosis (HPV16 n = 31, HPV18 n = 13). For the CIN2, CIN3/AIS and cancer categories, sequencing was performed on cell samples taken at the time of conisation; cytological examination of these samples was not performed. The cancer category included SCC (n = 4) and adenocarcinoma (n = 1) samples.

2.2. Library preparation and sequencing

Library preparation was performed using the TaME-seq method as described previously [34]. In brief, samples were subjected to tagmentation using Nextera DNA library prep kit (Illumina, Inc., San Diego, CA), following target enrichment performed by multiplex PCR using HPV primers and a combination of i7 index primers [37] and i5 index primers from the Nextera index kit (Illumina, Inc., San Diego, CA). Sequencing was performed on the HiSeq2500 platform with 125 bp paired-end reads.

2.3. Sequence alignment

Data was analysed by an in-house bioinformatics pipeline as described previously [34]. Reads were mapped to human genome (GRCh38/hg38) using HISAT2 (v2.1.0) [38]. HPV16 and HPV18 reference genomes were obtained from the PaVE database (https://pave.niaid.nih.gov). Mapping statistics and sequencing coverage were calculated using the Pysam package [39] with an in-house Python (v3.5.4) script. Downstream analysis was performed using an in-house R (v3.5.1) script. Samples with a mean sequencing depth of <300× were excluded from the further analysis.

Table 1
Number of samples and mean mappings statistics in each HPV16 and HPV18 diagnostic category.

Diagnostic category	Sequenced samples	Analysed samples		Mean age	Mean numbers in the analysed samples			
					Raw reads	Reads mapped to target HPV	Mean coverage	Fraction of genome covered by min. 100×
HPV16								
Normal[a]	24	2[e]	21	49 (32–68)	1.4 M	1.1 M	13516	0.78
ASC-US/LSIL[b]	31	19[e]		33 (19–54)				
CIN2[c]	47		27	31 (17–61)	0.6 M	0.4 M	4711	0.69
CIN3/AIS[c]	50		27	34 (22–54)	1.0 M	0.8 M	9616	0.76
Cancer[c,d]	5		5	30 (25–39)	2.4 M	1.7 M	20850	0.67
Total	157		80					
HPV18								
Normal[a]	3	1[e]	12	49 (47–52)	38.8 M	23.4 M	292143	0.86
ASC-US/LSIL[b]	13	11[e]		33 (20–49)				
CIN2[c]	13		9	34 (20–44)	77.1 M	36.5 M	431649	0.86
CIN3/AIS[c]	46		30	34 (24–54)	25.5 M	12.2 M	147747	0.82
Cancer	0		–	–				
Total	75		51					

[a] By cytology.
[b] By cytology; no cell abnormalities within 4-year follow-up.
[c] Cytology taken at the time of conisation, with the histological diagnosis presented.
[d] Includes cases of SCC (n = 4) and adenocarcinoma (n = 1).
[e] Non-progressive category, samples combined for analysis.

2.4. Sequence variation analysis

Mapped nucleotide counts over the HPV genomes and average mapping quality values for each nucleotide were retrieved from the HISAT sequence alignment. Variant calling was performed using an in-house R (v3.5.1) script. Nucleotides seen ≤2 times in each position and nucleotides with mean Phred quality score of <20 were filtered out. Since the analysis focused on the intra-host MNVs, the variant calling was performed independent of the reference genome; the most frequent base in each position was called as the major nucleotide and the second most abundant base as the MNV. Both F and R nucleotide counts from the same sample, obtained independently from separate amplification reactions, were combined and variant allele frequencies were calculated for each genomic position. If MNVs called from the two separate reactions were discordant, the highest covered MNV was used. Genomic positions covered with <100× were filtered out. MNVs were called if the MNV frequency was >1%. HPV16 and HPV18 have homopolymeric T tracts in NCR (HPV16:4156–4173, HPV16:4183–4212, HPV18:4198–4234); these regions may be prone to polymerase or sequencing errors and were filtered out.

The ratio of non-synonymous to synonymous substitutions (dN/dS) was calculated to indicate potential positive (new MNVs favoured) or negative (new MNVs eliminated) selection affecting protein-coding genes. For mutational signature analysis, all nucleotide substitutions were classified into six base substitutions, C > A, C > G, C > T, T > A, T > C, and T > G, and further into 96 trinucleotide substitution types, including information on the bases immediately 5′ and 3’ of the mutated base. To differentiate APOBEC3A and APOBEC3B activity, an extended mutational signature analysis was conducted on mutations in the genomic context YTCA and RTCA, respectively. Analysis was performed using an in-house R (v3.5.1) script.

2.5. Detection of chromosomal integration

Integration site detection was performed as described previously [34]. In brief, a two-step analysis strategy was employed to identify read pairs spanning integration sites. First, read pairs with one read mapped to HPV and the other to the human chromosome were identified using HISAT2. Second, unmapped reads were re-mapped using the LAST (v876) aligner (options -M -C2) [40] to increase detections of the above mentioned read pairs. Reads sharing the same start and end coordinates were considered as potential PCR duplicates and were excluded. Selected integration sites were confirmed by PCR amplification and Sanger sequencing on the ABI® 3130xl/3100 Genetic Analyzer 16-Capillary Array (Thermo Fisher Scientific Inc., Waltham, MA) using BigDye™ Terminator v1.1 cycle sequencing kit (Thermo Fisher Scientific Inc., Waltham, MA). Samples with a mean depth of >1000× and <85% of the genome covered by minimum 100× were manually inspected using IGV (v2.3.90) to detect HPV genomic deletions.

2.6. Functional annotation of genes within or close to integration sites

Nearest gene, with a transcription start site within 100 kb from the integration site, was identified using Ensembl. Gene2function (http://www.gene2function.org) and Genecards (https://www.genecards.org) were used to annotate the molecular function and disease phenotype of each gene. SNP associations in the GWAS Catalog [41] were retrieved from Genecards. Genes involved in cell cycle regulation, cell proliferation, apoptosis, tumour suppressor mechanisms, cancer-related pathways, or genes interacting with these pathways, or genes with direct cancer-related SNP associations, were termed as cancer-related genes. The integration sites were manually inspected using Geneious Prime (v.2019.0.4) to investigate whether the integration site was located in exons, introns or UTRs. Information regarding regulatory elements, including promoters, promoter flanking regions, enhancers and CTCF-binding sites, was retrieved from Ensembl regulatory build [42]. Integration sites in retained introns, ncRNA and antisense RNA were reported if they had a transcript support level of 1 or 2.

2.7. Statistical analysis

Statistical analyses were done in R (v3.5.1). The Kruskal-Wallis test was used to examine differences in numbers and frequencies of MNVs and integrations between the groups. A p-value of <0.05 was considered statistically significant.

2.8. Ethical considerations

This study was approved by the Regional Committee for Medical and Health Research Ethics, Oslo, Norway (REK 2017/447). Written informed consent has been obtained from all study participants.

3. Results

3.1. Characteristics and sequencing statistics

This study included 232 HPV16 and HPV18 positive cervical cell samples which were categorised according to cytology or histology diagnosis. A total of 80 HPV16 positive samples and 51 HPV18 positive samples, allocated to diagnostic categories of non-progressive disease, CIN2, CIN3/AIS and cancer, passed the strict sequencing depth criteria necessary for further analyses of minor nucleotide variation and integration. In total, 1.05 billion read pairs were analysed. The mean sequencing coverage per sample in the different categories ranged from 4711 (CIN2) to 20850 (cancer) for HPV16 positive samples and from 147747 (CIN3/AIS) to 431649 (CIN2) for HPV18 positive samples. On average, the samples had 77.7% of the genome covered with a minimum depth of 100× (Table 1).

3.2. Minor nucleotide variation profiles similar for HPV16 and HPV18

Overall, the number of MNVs was similar in HPV16 and HPV18 positive samples, and between the diagnostic categories. In total, 3669 MNVs were found in all 131 samples. In HPV16 positive samples, the mean number of MNVs found in the non-progressive category was 36 per sample, 29 in the CIN2 category, 27 in the CIN3/AIS category, and 24 in the cancer category. Corresponding numbers for HPV18 positive samples were 24, 20, and 27 for the non-progressive, CIN2 and CIN3/AIS categories, respectively (Fig. 1A). HPV16 positive samples had mean MNV frequencies of 2.8% for non-progressive, 2.9% for CIN2, 3.3% for CIN3/AIS and 3.0% for cancer categories. For HPV18 positive samples, the mean MNV frequencies were 3.1% for non-progressive, 2.6% for CIN2 and 5.0% for CIN3/AIS categories (Fig. 1B). Statistical analysis was performed; the mean numbers and MNV frequencies were not statistically different between the HPV types or the diagnostic groups within an HPV type.

3.3. Different level of variation in HPV16 and HPV18 genes

HPV MNVs occurred throughout all HPV genes (Fig. 2A). A higher degree of variation was observed in the HPV18 E4 gene throughout the different diagnostic categories. The dN/dS patterns for HPV16 showed mostly nonsynonymous variants (dN/dS > 1), while a considerable part of HPV18 genes had equal amounts of nonsynonymous and synonymous variants (dN/dS ≈ 1) (Fig. 2B). Strikingly, several HPV16 genes showed signs of positive selection, i.e. a preference for non-synonymous mutations (dN) over synonymous mutations (dS). HPV16 E6 had the most pronounced dN/dS ratio of 6. In contrast, the E7 gene in the same samples had a dN/dS ratio of 0.4, indicating neutral or negative selection. Over all, diagnostic categories and in both HPV types, the E2 gene displayed the highest dN/dS ratio, which for HPV18 were consistently >2. For the other HPV18 genes, the dN/dS ratio was close to 1 across diagnostic categories.

3.4. APOBEC3-related mutational signatures identified in non-progressive and CIN2 samples

Among nucleotide substitutions, predominantly C > T and T > C substitutions were observed across all diagnostic categories (Supplementary Figure S1). The APOBEC3-related C > T substitutions were compared between the different categories and HPV types (Fig. 3). C > T substitutions in the trinucleotide context TCW (W is A or T), a preferred target sequence for the APOBEC3 proteins [43] and a more stringent motif than TCN (N is any nucleotide [44], was the most prevalent mutational signature type in HPV16 non-progressive samples and to a slightly less extent in HPV16 CIN2 samples. HPV16 CIN3/AIS and cancer samples did not show any preferred signature patterns. Interestingly, HPV18 samples showed different C > T trinucleotide substitution patterns compared to HPV16 samples. In all HPV18 diagnostic categories, C > T substitutions in the trinucleotide context ACA was predominantly observed, while C > T substitutions in the trinucleotide context GCA was

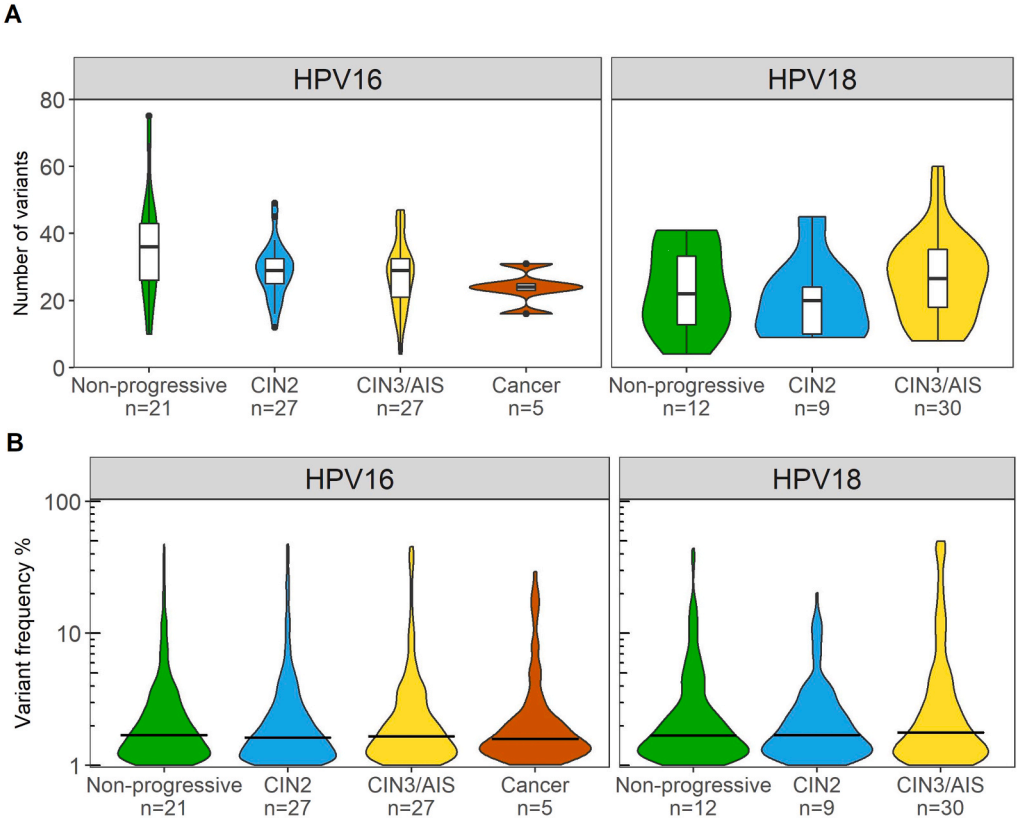

Fig. 1. Number of variants and variant frequencies in HPV16 and HPV18 positive samples. A) Number of variants presented as violin plots across the different diagnostic categories shown on x-axis. Violin plot shows the probability density of the data, using kernel density estimation. Box-and-whisker plots are added to show the median number (horizontal line), 25% and 75% percentiles (box), minimum and maximum values (whiskers). Black dots represent outliers. B) Variant frequencies (%) of detected minor variants shown as violin plots across the different diagnostic categories shown on x-axis. The horizontal bar indicates the median variant frequency.

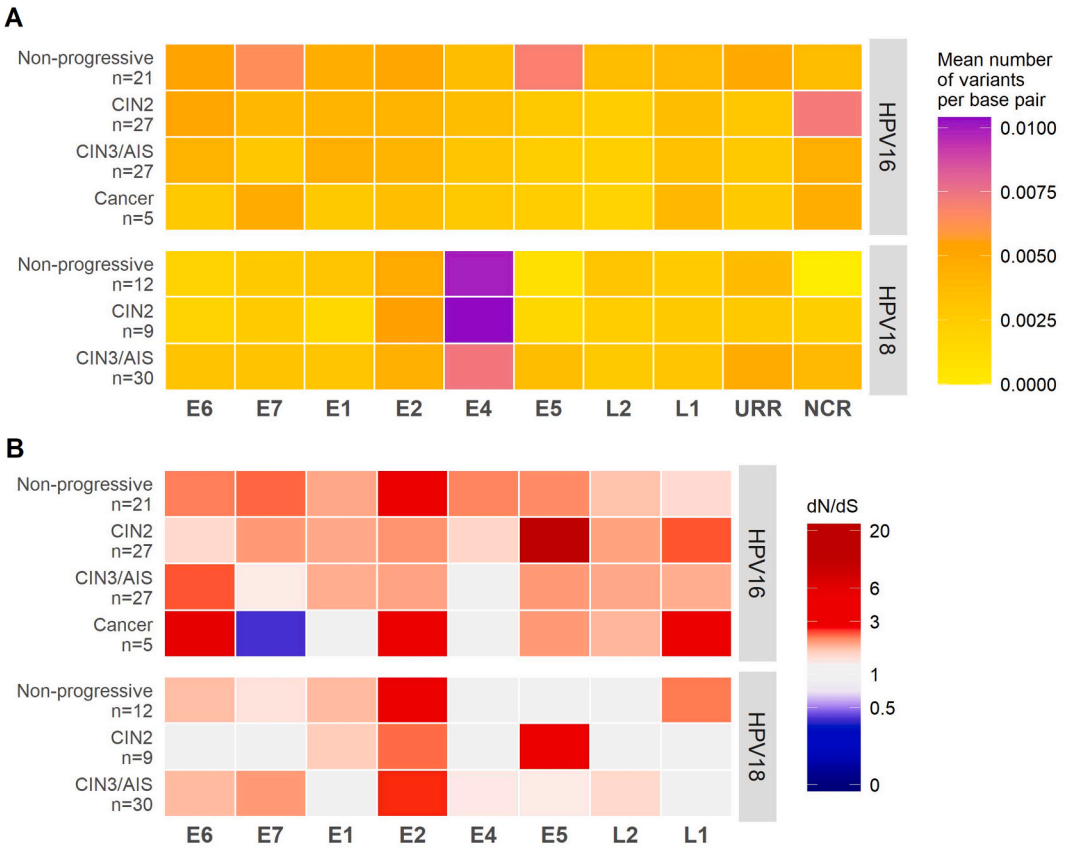

Fig. 2. Number of variants, nonsynonymous and synonymous variations in the different HPV genes. A) Heat map with yellow-orange-purple gradient colour-coding representing mean number of variants per sample in HPV16 and HPV18 genomic regions. Number of variants is normalised by the gene length and stratified by the diagnostic category. B) Heat map with blue-white-red gradient colour-coding representing the ratio of non-synonymous to synonymous substitutions (dN/dS) in HPV16 and HPV18 genomic regions across the different diagnostic categories.

the second most prevalent in non-progressive and CIN2 samples. For the extended signature mutational analysis, there were only 15 instances of mutations in the YTCA context in 8 samples while mutations in the RTCA context were not found in any samples in the dataset.

3.5. Higher HPV integration frequencies in HPV18 than in HPV16 positive samples

The proportion of samples with integration was 13% (10/80) for HPV16 and 59% (30/51) for HPV18 positive samples (Table 2). The integration frequency was higher in all HPV18 positive diagnostic categories compared to the HPV16 categories. Of the HPV16 positive samples, HPV integration was detected in 4%, 7% and 60% in CIN2, CIN3/AIS and cancer samples, respectively. Corresponding numbers in HPV18 samples were 78% and 53% for CIN2 and CIN3/AIS categories, respectively. The total number of integration sites found in each diagnostic category was in general higher for HPV18 positive samples, ranging from 22 (CIN2) to 60 (CIN3/AIS), while for HPV16 samples, a total of 17 integration sites were identified (Table 2).

In Fig. 4A, the difference between HPV16 and HPV18 positive samples in terms of number of integration sites is illustrated, stratified by diagnostic category. Combined for all diagnostic groups, HPV18 samples had significantly more integration sites than HPV16 samples (p-value < 0.001). The mean numbers of integration sites per HPV18 positive sample were 3.4, 3.1 and 3.8 for the non-progressive, CIN2 and CIN3/AIS categories, respectively. The mean numbers of integration sites per HPV16 positive sample with observed integration, were 1.3, 2, 1.5 and 2.3 for the non-progressive, CIN2, CIN3/AIS and cancer categories, respectively (Fig. 4A). In total, six HPV16 positive samples and 18 HPV18 positive samples had more than one integration site observed (Supplementary Table S1).

The validation rates of integration sites using Sanger sequencing (good quality chromatograms produced) was 44% (7/16 samples) (Supplementary Table S1, Supplementary Table S2). A PCR product or a smear was identified on agarose gel but no clean chromatogram was seen in additional 44% (7/16) of the reactions (Supplementary Figure S2). Two integration sites, one in HPV16 and one in HPV18 positive sample, both in the non-progressive category, could not be confirmed (Supplementary Table S1).

3.6. Break points and deletions in the HPV genome

For HPV16, integration-associated break points in the viral genome were detected in all genes except E4 and E7. Notably, NCR between the E5 and L2 genes, harboured two break points in one cancer sample (Fig. 4B, Supplementary Table S1). In the HPV18 positive samples, break points were located in all HPV genomic regions except NCR. Expected number of break points in each gene relative to gene lengths was estimated with regard to randomness by dividing the total number of break points within a HPV type by the length of the gene. Based on this, breaks were more frequently observed in E1 and NCR in HPV16 samples and in E2, E4 and L2 in HPV18 samples, while L1 and URR were less prone to break (Fig. 4B). For HPV16 and HPV18 combined, break points were located in E1 or E2 in 38%, 38%, 48%, and 57% of all the breaks in non-progressive, CIN2, CIN3/AIS, and cancer categories, respectively (Supplementary Figure S3). All cancer samples had at least one break point in E1 or E2 (Supplementary Table S1).

HPV genomic regions covered with very few or no sequencing reads were considered as deletions according to previous validations [34]. Such deletions were observed in six samples; one HPV16 positive sample

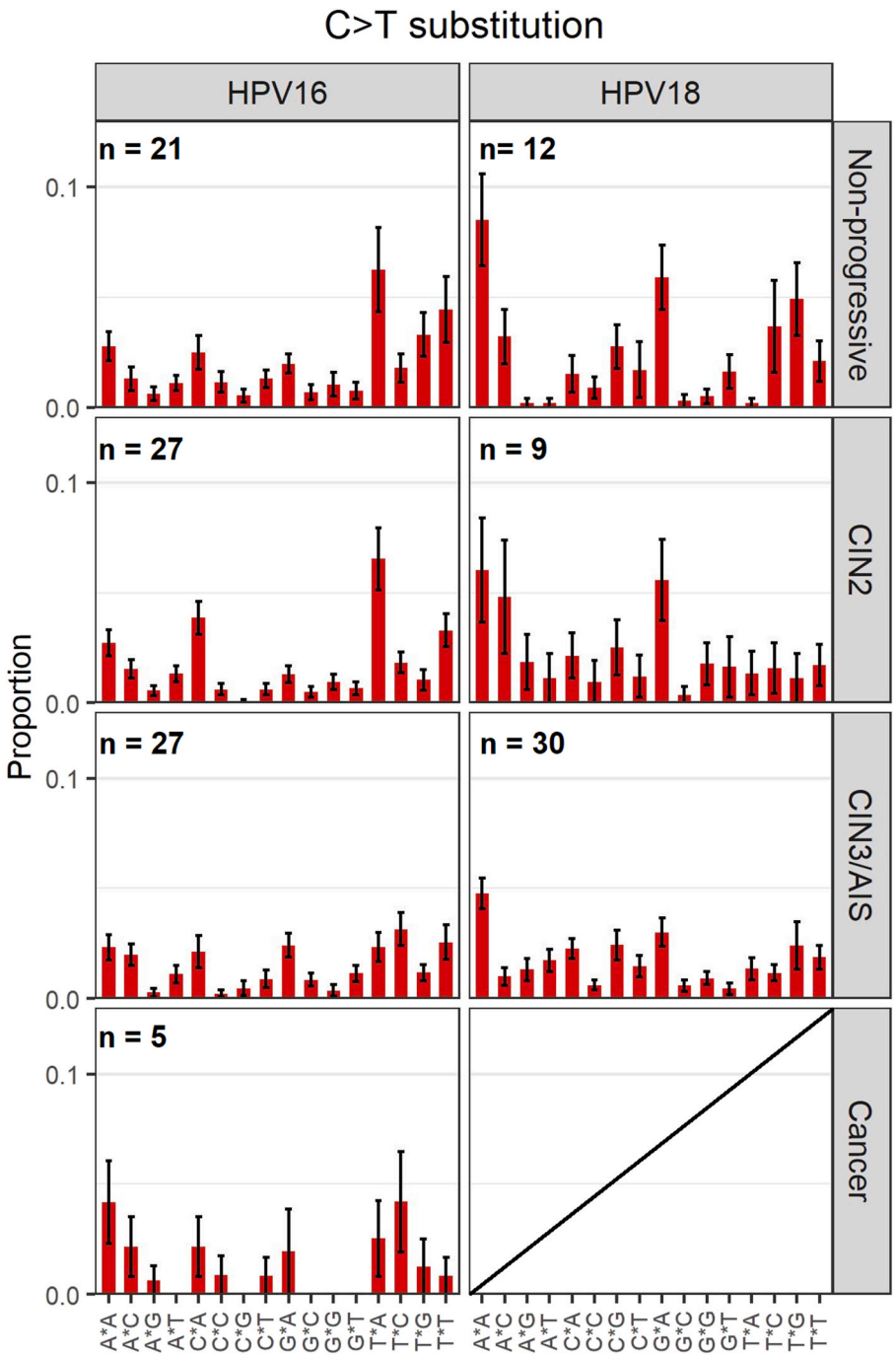

Fig. 3. C > T mutational signatures in HPV16 and HPV18 positive samples. The mean proportion of 16 trinucleotide substitution types is shown below the plots across the different diagnostic categories. Error bars represent the standard error of the mean.

(cancer) and five HPV18 positive samples (Supplementary Figure S4). For these samples, human sequences were detected flanking the deleted regions, indicating chromosomal integration. In all six samples, the genomic deletion encompassed the region between E1/E2 and L2. The deletions were either partial, suggesting the presence of both episomal and integrated HPV DNA, or complete with no reads detected for the deleted region.

3.7. Integration sites in the human genome

In HPV16 positive samples, integration sites (n = 17) were distributed on 10 chromosomes; for the cancer samples, all integration sites (n = 7) were located on chromosomes 1, 8 or 10 (Fig. 4C). Interestingly, the integration sites on chromosome 8 were located in the *PVT1* oncogene, in the chromosomal locus 8q24.21 (Supplementary Table S1), previously being defined as an HPV integration hotspot [31]. For the HPV18 positive samples, integration sites (n = 106) were found in all chromosomes except chromosomes 18 and 21 (Fig. 4C). Most HPV18 integration sites were observed on chromosomes 2 and 4. In HPV18 samples, 36% (4/11) of the integration sites on chromosome 4 were located in the previously defined hotspot locus 4q13.3 [31], all from samples diagnosed with CIN2 or CIN3/AIS.

Due to a low frequency of integration events in HPV16 positive samples, HPV16 and HPV18 samples were combined for reporting HPV integrations affecting different human genetic elements. The frequency of integration sites located in human genes ranged from 50 to 71%, with

Table 2
Number of HPV16 and HPV18 positive samples with integration, stratified by the diagnostic categories.

Diagnostic category	Number of samples with integration (Frequency %)	Total number of integration sites
HPV16		
Non-progressive (n = 21)	4 (19%)	5
CIN2 (n = 27)	1 (4%)	2
CIN3/AIS (n = 27)	2 (7%)	3
Cancer (n = 5)	3 (60%)	7
Total (n = 80)	10 (13%)	17
HPV18		
Non-progressive (n = 12)	7 (58%)	24
CIN2 (n = 9)	7 (78%)	22
CIN3/AIS (n = 30)	16 (53%)	60
Total (n = 51)	30 (59%)	106

the highest frequency observed in cancer samples (Fig. 5A). Integration sites were detected in or close to cancer-related genes (Supplementary Table S3) in 100% (7/7) of cancer samples (n = 3), in 65% (41/63) of CIN3/AIS samples (n = 18), in 38% (9/24) of CIN2 samples (n = 8), and in 34% (10/29) in non-progressive samples (n = 11) (Fig. 5B). In individual samples, the highest numbers of integration sites located in or near cancer-related genes was 13/21 in CIN3/AIS, 3/10 in CIN2, and 5/12 in non-progressive samples, all being HPV18 positive (Supplementary Figure S5).

Integration located in exons, introns, regulatory regions, retained introns, non-coding RNA (ncRNA), antisense RNA and untranslated regions (UTRs) varied between the diagnostic groups (Supplementary Table S1). Integration frequency in exons and regulatory regions decreased with lesion severity, while the integration frequency in introns, retained introns and ncRNA increased with lesion severity. Antisense and UTR showed only few integrations in certain diagnostic groups (Supplementary Figure S6).

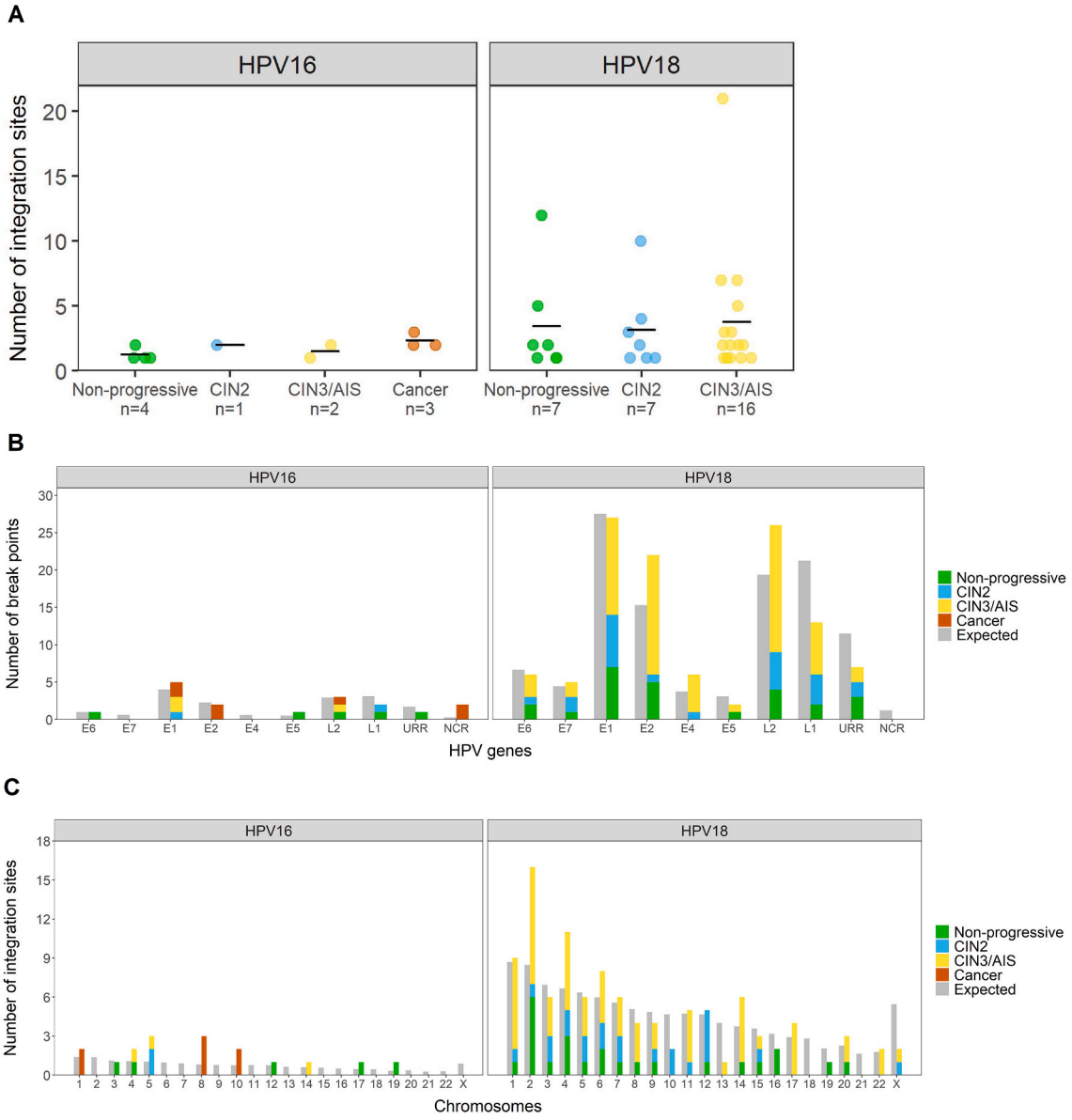

Fig. 4. Chromosomal integration sites and HPV break points in HPV16 and HPV18 positive samples. A) Number of integration sites in samples with observed integration. Each spot in the plot indicates one sample. Total number of samples with integration is specified for each diagnostic category on x-axis. Vertical lines indicate the mean number of integration sites. B) Break points in HPV genes. C) Integration sites in human chromosomes compared to expected number of break points assuming random viral genome integration.

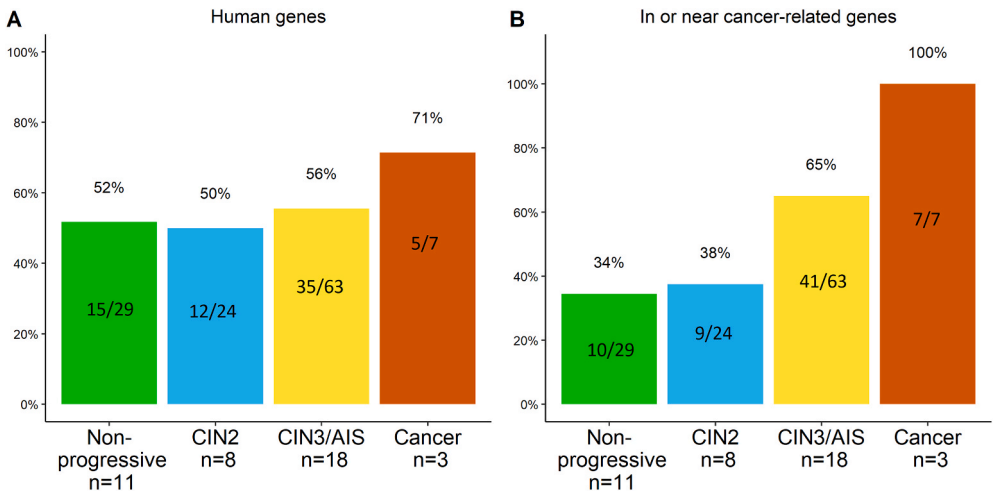

Fig. 5. The frequency of integration sites combined for HPV16 and HPV18 A) in human genes, and B) in or near human cancer-related genes. Number of integration sites is indicated inside the bars and total number of samples with integration (n) for each diagnostic category is specified on x-axis.

4. Discussion

This study compares HPV16- and HPV18-associated genomic events, i.e. MNVs and integrations, in normal/ASC-US/LSIL samples from women with no clinical progression; CIN2, CIN3/AIS and cervical cancer samples. We find that these genomic events are strikingly different between HPV16 and HPV18 positive samples. In line with other studies [11,20], we show decreases in APOBEC3-related nucleotide substitutions in HPV16 positive samples of increasing severity. As previously reported [25,45], HPV18 samples show higher integration frequencies compared to HPV16, while we also found an increase in integration frequencies in or in close proximity to cancer-related genes with increasing lesion severity.

In this study, the number and frequency of intra-host MNVs was similar between HPV genotypes and morphological categories. Recent HPV deep sequencing studies, exploring HPV genomic variation with various PCR-based NGS approaches and different variant calling thresholds, show slightly divergent numbers of MNVs [11,20,34]. We found a total of 3669 MNVs in the 131 samples, being in line with studies reporting a high number of HPV variation at the population level [24,46,47], within infected hosts [11,12,34]. A recent study on HPV16 genome stability analysed possible HPV16 sublineage co-infections and observed 20–38 variants in each sample [48], corresponding to the mean numbers of MNVs in this study. The variation was reported not to be due to co-infections, but interpretation of the nucleotide variation source was not further elaborated [48]. The prevalence of sub-lineage co-infections is expected to be low [49].

When investigating the number of MNVs for each region or gene in the HPV genomes, normalised by the gene length, HPV18 E4 showed a higher degree of variation relative to other genomic regions. This is an interesting observation which should be further examined. For HPV16, a higher degree of variation in the NCR was initially observed in the categories CIN2, CIN3/AIS and cancer. However, when filtering out the homopolymeric T tracts in the NCR, the differences between categories subside. This filtering was done since the T tracts are inherently unstable making it challenging to assign mutations to methodological factors or true biology. Similar variation was not seen for HPV18 positive samples with less homopolymeric tracts. Recent studies document high degrees of variation in HPV16 NCR, but without any biological interpretation [11,23]. The NCR in HPV16 has been characterised to portray a weak promoter activity specific to L2 mRNA expression [50]. Repeat sequences of varying length in NCR have been reported [51] and the NCR has been shown to harbour miRNA binding sites [52]. The loss of miRNA binding sites due to nucleotide variation in NCR was suggested to serve as a novel mechanism to sustain L2 expression, and thereby justify the potential role of L2 in HPV-induced carcinogenesis [52]. However, an opposite finding has also been reported, showing more variation in NCR in clearing than in persistent HPV16 infections [46].

Ratio of nonsynonymous to synonymous variants (dN/dS) is used as indicator of positive or negative selection occurring over generations within hosts [14]. This ratio may indicate non-random occurrence and persistence of minor nucleotide variability in genes. In this study, the observed nucleotide variations in the HPV16 and HPV18 genes were biased toward nonsynonymous substitutions, being in line with previous results showing a high ratio of non-synonymous nucleotide variation [11]. Only HPV16 E7 had a dN/dS ratio of <1, indicating negative selection and conservation of function. Interestingly, two recent studies reported similar results on strict conservation of the HPV16 E7 gene at the population level [23,24]. A potential source of synonymous and non-synonymous substitutions may be APOBEC3 activity creating C > T substitutions [16]. APOBEC3-related mutations have previously been reported in cervical cancer lesions [11,19,20]. Our finding of APOBEC3-related signatures in the HPV16 positive non-progressive samples indicates that this mechanism is active also in an early stage of infection. The relative amount of variants related to APOBEC3 may at a more severe stage of disease disappear, due to an increase in non-APOBEC3 mutations caused by e.g. hampered DNA repair mechanisms in an increasingly cancerous environment [53]. This study was the first to characterise mutational patterns in HPV18 samples, showing mutation patterns in the trinucleotide context RCA (R is A or G), a target motif for the activation-induced cytidine deaminase (AID) that is a member of the APOBEC protein family [54].

HPV-induced carcinogenesis is a multi-step process that may be facilitated through the disruption of host genes and genomic instability caused by viral integration [28–30]. A high number of integrations in a sample may in itself be a sign of genomic instability, which may further accelerate such events. In our dataset, multiple integration sites were observed in 24 samples, with the maximum of 21 integration sites in one HPV18 sample in the CIN3/AIS category, possibly promoting a higher degree of chromosomal instability. Our results showed a higher number of integration events in HPV18 positive samples compared to HPV16 positive samples, being consistent with previous observations [25,45]. Genomic instability as a consequence of multiple integrations, is further strengthened by finding integrations in the E1 and E2 genes, which might result in overexpression of the viral oncogenes E6 and E7. Previous studies using NGS methodology for HPV integration analysis report disruptions mainly in E1 and E2 genes in samples that have progressed to cancer [55,56]. In addition, we found HPV genomic

deletions in one HPV16 positive cancer sample and in five HPV18 positive samples of all categories. In all of these, the genomic deletion always led to partial or complete loss of E1, E2 and L2. Similar results showing HPV genomic deletions have been reported in cervical carcinomas [57] and HPV positive oropharyngeal squamous cell carcinomas [58]. Interestingly, we also observed integration with break points in NCR in one cancer sample. To our knowledge, this is the first study to report break points in NCR.

Due to the low frequency of integration events in HPV16 positive samples, HPV16 and HPV18 integrations were combined for the analysis of integrations in or close to cancer-related genes. We observed an overall increase in the proportion of integration sites within or close to cancer-related genes with increasing lesion severity. All integrations in the cancer samples occurred within or near the cancer-related genes *PVT1*, *WAC* and *miR-205*. The *PVT1* oncogene, a long non-coding RNA gene, has been associated with multiple cancers including cervical cancer [59]. The *PVT1* gene is located in the chromosomal locus 8q24.21, which is one of the regions previously reported to contain integration sites in cervical carcinomas more often than other loci [31]. Transcription of *PVT1* is regulated by the key tumour suppressor protein p53 and PVT1 is implicated in regulating the *MYC* oncogene [60]. The WAC protein regulates the cell-cycle checkpoint activation in response to DNA damage and is a positive regulator of mTOR, which functions as a key player in the regulation of cell growth and metabolism [61]. The miRNA miR-205 has been implicated in many cancers and targets genes involved in DNA repair, cell cycle control and cancer-related pathways [62]. In the CIN2 and CIN3/AIS categories, 38% and 65% of the integration sites were observed in or close to cancer-related genes, respectively. Interestingly, integration sites in or close to cancer-related genes were also observed in the non-progressive disease category. Whether this might represent one of several components for risk stratification remains to be determined. Our results, together with a recent study [63], have shown that viral integrations may also occur in other genetic elements that are involved in regulation of gene expression, such as ncRNA and UTRs.

NGS protocols with comprehensive analyses of whole HPV genomes, their variability and integrations, enable greater understanding of the role of genomic events during cancer development. By comparatively analysing genomic events, we get a broader picture of the dynamic changes in the HPV genome during malignant cell transformation. HPV16 and HPV18 are to a certain degree associated with different types of invasive cervical cancers [3,4] and may utilise different molecular mechanisms to induce carcinogenesis. Firstly, HPV18 is suggested to cause more genomic instability [4,45] and HPV18 lesions are more aggressively progressing from CIN3 to cancer than HPV16 positive lesions [4]. Furthermore, previously reported results show different DNA methylation patterns [64] and mechanistic signatures of integrations [57] for HPV16 and HPV18, which strengthens the hypothesis of different underlying mechanisms for HPV16- and HPV18-induced cervical carcinogenesis.

Despite the large sample number in total, the sample size in certain diagnostic categories was low, limiting us from performing statistical analyses and drawing conclusions from the given part of the dataset. Some samples, mainly in the non-progressive category, had low sequencing coverage for the HPV genome. This is most likely explained by low viral load, which was not measured in the samples. Low viral load has previously been observed to affect the sequencing yield [13]. Two integration sites in non-progressive samples were not confirmed by Sanger sequencing. This may be explained by sub-optimal PCR primers, PCR conditions, low viral load or may reflect repeated integrations or other genomic structures affecting the PCR reaction. Still, since the NGS data showed clear results, both integration sites were included in the analysis.

5. Conclusions

To summarise, we have in this study analysed intra-host HPV minor nucleotide variation, chromosomal integration and genomic deletions in cervical cell samples with different morphology by utilising the TaME-seq protocol [34]. The results show a high number of low-frequency variation, distinct variation patterns and integration frequencies, providing initial insight into dissimilar genomic alterations between HPV16 and HPV18, possibly reflecting differences in the mechanisms of cell transformation induced by the two genotypes. In addition, the study adds to the growing evidence of within-host HPV genomic variability. Cancer registry data with information on future cervical disease or longitudinal studies including patient outcome, preferably with a larger sample size for all diagnostic categories, are needed for further interpretation of different HPV whole genome MNV signatures and to validate the role and importance of viral integrations.

CRediT authorship contribution statement

Sonja Lagström: Writing – original draft, Formal analysis, designed and performed the experiments, analysed the results and drafted the manuscript text. . All authors contributed to writing and approved the final version of the manuscript. **Alexander Hesselberg Løvestad:** Writing – original draft, Formal analysis, analysed the results and contributed to drafting the manuscript. . All authors contributed to writing and approved the final version of the manuscript. **Sinan Uğur Umu:** Writing – original draft, Data curation, Formal analysis, contributed to the data analysis. . All authors contributed to writing and approved the final version of the manuscript. **Ole Herman Ambur:** Writing – original draft, contributed to the study design and result interpretation. . All authors contributed to writing and approved the final version of the manuscript. **Mari Nygård:** Writing – original draft, contributed to the clinical interpretation of the results. . All authors contributed to writing and approved the final version of the manuscript. **Trine B. Rounge:** Writing – original draft, Data curation, Formal analysis, contributed to the study design, data analysis and result interpretation. . All authors contributed to writing and approved the final version of the manuscript. **Irene Kraus Christiansen:** Writing – original draft, managed the sample material, contributed to the study design and result interpretation. All authors contributed to writing and approved the final version of the manuscript.

Acknowledgements

We thank Mona Hansen for DNA extraction, Hanne Kristiansen for sequencing library preparation, Karin Helmersen for Sanger sequencing, and Marcin W. Wojewodzic for his help with gene annotation of the chromosomal integration sites.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.tvr.2021.200221.

Funding

This work was funded by a grant from South-Eastern Norway Regional Health Authority (project number 2016020).

Data statement

The data presented in this article are not readily available because of the principles and conditions set out in the General Data Protection Regulation (GDPR), with additional national legal basis as per the Regulations on population-based health surveys and ethical approval from the Norwegian Regional Committee for Medical and Health

Research Ethics (REC). Requests to access the data should be directed to the corresponding authors.

Authors' contributions

SL designed and performed the experiments, analysed the results and drafted the manuscript text. AHL analysed the results and contributed to drafting the manuscript. SUU contributed to the data analysis. OHA contributed to the study design and result interpretation. MN contributed to the clinical interpretation of the results. TBR contributed to the study design, data analysis and result interpretation. IKC managed the sample material, contributed to the study design and result interpretation. All authors contributed to writing and approved the final version of the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] F.X. Bosch, et al., The causal relation between human papillomavirus and cervical cancer, J. Clin. Pathol. 55 (2002) 244–265.
[2] IARC working group on the evaluation of carcinogenic risks to humans, biological agents. Volume 100 B. A review of human carcinogens, IARC Monogr. Eval. Carcinog. Risks Hum. 100 (2012) 1–441.
[3] S. de Sanjose, et al., Human papillomavirus genotype attribution in invasive cervical cancer: a retrospective cross-sectional worldwide study, Lancet Oncol. 11 (2010) 1048–1056, https://doi.org/10.1016/s1470-2045(10)70230-8.
[4] W.A. Tjalma, et al., Differences in human papillomavirus type distribution in high-grade cervical intraepithelial neoplasia and invasive cervical cancer in Europe, Int. J. Canc. 132 (2013) 854–867, https://doi.org/10.1002/ijc.27713.
[5] H. zur Hausen, Papillomaviruses and cancer: from basic studies to clinical application, Nat. Rev. Canc. 2 (2002) 342–350, https://doi.org/10.1038/nrc798.
[6] H.U. Bernard, Taxonomy and phylogeny of papillomaviruses: an overview and recent developments, Infect. Genet. Evol. 18 (2013) 357–361, https://doi.org/10.1016/j.meegid.2013.03.011.
[7] B. Smith, et al., Sequence imputation of HPV16 genomes for genetic association studies, PloS One 6 (2011), e21375, https://doi.org/10.1371/journal.pone.0021375.
[8] D. Bzhalava, C. Eklund, J. Dillner, International standardization and classification of human papillomavirus types, Virology 476 (2015) 341–344, https://doi.org/10.1016/j.virol.2014.12.028.
[9] R.D. Burk, A. Harari, Z. Chen, Human papillomavirus genome variants, Virology 445 (2013) 232–243, https://doi.org/10.1016/j.virol.2013.07.018.
[10] K. Van Doorslaer, Evolution of the papillomaviridae, Virology 445 (2013) 11–20, https://doi.org/10.1016/j.virol.2013.05.012.
[11] Y. Hirose, et al., Within-host variations of human papillomavirus reveal APOBEC-signature mutagenesis in the viral genome, J. Virol. (2018), https://doi.org/10.1128/jvi.00017-18.
[12] R.S. Dube Mandishora, et al., Intra-host sequence variability in human papillomavirus, Papillomavirus Res (2018), https://doi.org/10.1016/j.pvr.2018.04.006.
[13] S. Lagstrom, et al., HPV16 whole genome minority variants in persistent infections from young Dutch women, J. Clin. Virol. 119 (2019) 24–30, https://doi.org/10.1016/j.jcv.2019.08.003.
[14] E. Domingo, J. Sheldon, C. Perales, Viral quasispecies evolution, Microbiol. Mol. Biol. Rev. 76 (2012) 159–216, https://doi.org/10.1128/MMBR.05023-11.
[15] C.J. Warren, et al., APOBEC3A functions as a restriction factor of human papillomavirus, J. Virol. 89 (2015) 688–702, https://doi.org/10.1128/JVI.02383-14.
[16] R.S. Harris, J.P. Dudley, APOBECs and virus restriction, Virology 479–480 (2015) 131–145, https://doi.org/10.1016/j.virol.2015.03.012.
[17] L.B. Alexandrov, et al., Signatures of mutational processes in human cancer, Nature 500 (2013) 415–421, https://doi.org/10.1038/nature12477.
[18] J.P. Vartanian, et al., Evidence for editing of human papillomavirus DNA by APOBEC3 in benign and precancerous lesions, Science 320 (2008) 230–233, https://doi.org/10.1126/science.1153201.
[19] A.A. Mariaggi, et al., Presence of human papillomavirus (HPV) apolipoprotein B messenger RNA editing, catalytic polypeptide-like 3 (APOBEC)-Related minority variants in HPV-16 genomes from anal and cervical samples but not in HPV-52 and HPV-58, J. Infect. Dis. 218 (2018) 1027–1036, https://doi.org/10.1093/infdis/jiy287.
[20] B. Zhu, et al., Mutations in the HPV16 genome induced by APOBEC3 are associated with viral clearance, Nat. Commun. 11 (2020) 886, https://doi.org/10.1038/s41467-020-14730-1.
[21] V.C. Vieira, et al., Human papillomavirus E6 triggers upregulation of the antiviral and cancer genomic DNA deaminase APOBEC3B, mBio 5 (2014), https://doi.org/10.1128/mBio.02234-14.
[22] K. Chan, et al., An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers, Nat. Genet. 47 (2015) 1067–1072, https://doi.org/10.1038/ng.3378.
[23] L.S. Arroyo-Muhr, et al., Human papillomavirus type 16 genomic variation in women with subsequent in situ or invasive cervical cancer: prospective population-based study, Br. J. Canc. 119 (2018) 1163–1168, https://doi.org/10.1038/s41416-018-0311-7.
[24] L. Mirabello, et al., HPV16 E7 genetic conservation is critical to carcinogenesis, Cell 170 (2017) 1164–1174, https://doi.org/10.1016/j.cell.2017.08.001, e6.
[25] Cancer Genome Atlas Research Network, Integrated genomic and molecular characterization of cervical cancer, Nature 543 (2017) 378–384, https://doi.org/10.1038/nature21386.
[26] J. Doorbar, et al., Human papillomavirus molecular biology and disease association, Rev. Med. Virol. 25 (Suppl 1) (2015) 2–23, https://doi.org/10.1002/rmv.1822.
[27] A.A. McBride, A. Warburton, The role of integration in oncogenic progression of HPV-associated cancers, PLoS Pathog. 13 (2017), e1006211, https://doi.org/10.1371/journal.ppat.1006211.
[28] K. Akagi, et al., Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability, Genome Res. 24 (2014) 185–199, https://doi.org/10.1101/gr.164806.113.
[29] C. Bodelon, et al., Genomic characterization of viral integration sites in HPV-related cancers, Int. J. Canc. 139 (2016) 2001–2011, https://doi.org/10.1002/ijc.30243.
[30] M. Peter, et al., Frequent genomic structural alterations at HPV insertion sites in cervical carcinoma, J. Pathol. 221 (2010) 320–330, https://doi.org/10.1002/path.2713.
[31] I. Kraus, et al., The majority of viral-cellular fusion transcripts in cervical carcinomas cotranscribe cellular sequences of known or predicted genes, Canc. Res. 68 (2008) 2514–2522, https://doi.org/10.1158/0008-5472.Can-07-2776.
[32] Y. Liu, et al., Genome-wide profiling of the human papillomavirus DNA integration in cervical intraepithelial neoplasia and normal cervical epithelium by HPV capture technology, Sci. Rep. 6 (2016) 35427, https://doi.org/10.1038/srep35427.
[33] J. Huang, et al., Comprehensive genomic variation profiling of cervical intraepithelial neoplasia and cervical cancer identifies potential targets for cervical cancer early warning, J. Med. Genet. 56 (2019) 186–194, https://doi.org/10.1136/jmedgenet-2018-105745.
[34] S. Lagström, et al., TaME-seq: an efficient sequencing approach for characterisation of HPV genomic variability and chromosomal integration, Sci. Rep. 9 (2019) 524, https://doi.org/10.1038/s41598-018-36669-6.
[35] A. Trope, et al., Performance of human papillomavirus DNA and mRNA testing strategies for women with and without cervical neoplasia, J. Clin. Microbiol. 47 (2009) 2458–2464, https://doi.org/10.1128/JCM.01863-08.
[36] A. Trope, et al., Cytology and human papillomavirus testing 6 to 12 months after ASCUS or LSIL cytology in organized screening to predict high-grade cervical neoplasia between screening rounds, J. Clin. Microbiol. 50 (2012) 1927–1935, https://doi.org/10.1128/JCM.00265-12.
[37] 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 (2013) 5112–5120, https://doi.org/10.1128/AEM.01043-13.
[38] D. Kim, B. Langmead, S.L. Salzberg, HISAT: a fast spliced aligner with low memory requirements, Nat. Methods 12 (2015) 357–360, https://doi.org/10.1038/nmeth.3317.
[39] H. Li, et al., The sequence alignment/map format and SAMtools, Bioinformatics 25 (2009) 2078–2079, https://doi.org/10.1093/bioinformatics/btp352.
[40] S.M. Kielbasa, et al., Adaptive seeds tame genomic sequence comparison, Genome Res. 21 (2011) 487–493, https://doi.org/10.1101/gr.113985.110.
[41] D. Welter, et al., The NHGRI GWAS Catalog, a curated resource of SNP-trait associations, Nucleic Acids Res. 42 (2014) D1001–D1006, https://doi.org/10.1093/nar/gkt1229.
[42] D.R. Zerbino, et al., The ensembl regulatory build, Genome Biol. 16 (2015) 56, https://doi.org/10.1186/s13059-015-0621-5.
[43] S.A. Roberts, et al., An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers, Nat. Genet. 45 (2013) 970–976, https://doi.org/10.1038/ng.2702.
[44] M.B. Burns, et al., APOBEC3B is an enzymatic source of mutation in breast cancer, Nature 494 (2013) 366–370, https://doi.org/10.1038/nature11881.
[45] S. Vinokurova, et al., Type-dependent integration frequency of human papillomavirus genomes in cervical lesions, Canc. Res. 68 (2008) 307–313, https://doi.org/10.1158/0008-5472.CAN-07-2754.
[46] P. van der Weele, C. Meijer, A.J. King, Whole-genome sequencing and variant analysis of human papillomavirus 16 infections, J. Virol. 91 (2017), https://doi.org/10.1128/jvi.00844-17.
[47] P. van der Weele, C. Meijer, A.J. King, High whole-genome sequence diversity of human papillomavirus type 18 isolates, Viruses 10 (2018), https://doi.org/10.3390/v10020068.
[48] L.S. Arroyo-Muhr, et al., The HPV16 genome is stable in women who progress to in situ or invasive cervical cancer: a prospective population-based study, Canc. Res. 79 (2019) 4532–4538, https://doi.org/10.1158/0008-5472.CAN-18-3933.

[49] D.T. Geraets, et al., Long-term follow-up of HPV16-positive women: persistence of the same genetic variant and low prevalence of variant co-infections, PloS One 8 (2013), e80382, https://doi.org/10.1371/journal.pone.0080382.
[50] H. Maki, K. Fujikawa-Adachi, O. Yoshie, Evidence for a promoter-like activity in the short non-coding region of human papillomaviruses, J. Gen. Virol. 77 (Pt 3) (1996) 453–458, https://doi.org/10.1099/0022-1317-77-3-453.
[51] B. Bhattacharjee, et al., Characterization of sequence variations within HPV16 isolates among Indian women: prediction of causal role of rare non-synonymous variations within intact isolates in cervical cancer pathogenesis, Virology 377 (2008) 143–150, https://doi.org/10.1016/j.virol.2008.04.007.
[52] P. Mandal, et al., Differential expression of HPV16 L2 gene in cervical cancers harboring episomal HPV16 genomes: influence of synonymous and non-coding region variations, PloS One 8 (2013), e65647, https://doi.org/10.1371/journal.pone.0065647.
[53] K. McFadden, M.A. Luftig, Interplay between DNA tumor viruses and the host DNA damage response, Curr. Top. Microbiol. Immunol. 371 (2013) 229–257, https://doi.org/10.1007/978-3-642-37765-5_9.
[54] P. Pham, et al., Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation, Nature 424 (2003) 103–107, https://doi.org/10.1038/nature01760.
[55] Z. Hu, et al., Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism, Nat. Genet. 47 (2015) 158–163, https://doi.org/10.1038/ng.3178.
[56] B. Xu, et al., Multiplex identification of human papillomavirus 16 DNA integration sites in cervical carcinomas, PloS One 8 (2013), e66693, https://doi.org/10.1371/journal.pone.0066693.
[57] A. Holmes, et al., Mechanistic signatures of HPV insertions in cervical carcinomas, npj Genomic Medicine 1 (2016), https://doi.org/10.1038/npjgenmed.2016.4.
[58] G. Gao, et al., Whole genome sequencing reveals complexity in both HPV sequences present and HPV integrations in HPV-positive oropharyngeal squamous cell carcinomas, BMC Canc. 19 (2019) 352, https://doi.org/10.1186/s12885-019-5536-1.
[59] M. Iden, et al., The lncRNA PVT1 contributes to the cervical cancer phenotype and associates with poor patient prognosis, PloS One 11 (2016), e0156274, https://doi.org/10.1371/journal.pone.0156274.
[60] S.W. Cho, et al., Promoter of lncRNA gene PVT1 is a tumor-suppressor DNA boundary element, Cell 173 (2018) 1398–1412, https://doi.org/10.1016/j.cell.2018.03.068, e22.
[61] G. David-Morrison, et al., WAC regulates mTOR activity by acting as an adaptor for the TTT and pontin/reptin complexes, Dev. Cell 36 (2016) 139–151, https://doi.org/10.1016/j.devcel.2015.12.019.
[62] A.Y. Qin, et al., MiR-205 in cancer: an angel or a devil? Eur. J. Cell Biol. 92 (2013) 54–60, https://doi.org/10.1016/j.ejcb.2012.11.002.
[63] D. Tang, et al., VISDB: a manually curated database of viral integration sites in the human genome, Nucleic Acids Res. 48 (2020) D633–D641, https://doi.org/10.1093/nar/gkz867.
[64] S.M. Amaro-Filho, et al., HPV DNA methylation at the early promoter and E1/E2 integrity: a comparison between HPV16, HPV18 and HPV45 in cervical cancer, Papillomavirus Res 5 (2018) 172–179, https://doi.org/10.1016/j.pvr.2018.04.002.