

OSLOMET

# Master's Thesis

Master in Biomedicine

May 2021

## HPV chromosomal integration as a biomarker for cancer progression

*Validation and characterization of integration sites in HPV31, 33 and 45 positive cervical samples*

Name: Adina Repesa (Candidate number: 504)  
Course code: MABIO5900

60 ECTS

**Faculty of Health Sciences**  
OSLO METROPOLITAN UNIVERSITY  
STORBYUNIVERSITETET



A thesis submitted for the degree of

Master in Biomedicine, 60 ECT

By

Adina Repesa

## “HPV chromosomal integration as a biomarker for cancer progression”

*Validation and characterization of integration sites in HPV31, 33 and 45 positive cervical samples*

Faculty of Health Sciences

Department of Life Sciences and Health

Performed at Akershus University Hospital

Supervisors: Ole Herman Ambur, Alexander Hesselberg Løvestad, Irene Kraus Christiansen

May 2021

OsloMet – Oslo Metropolitan University

OSLOMET

 Akershus  
universitetssykehus

## Acknowledgements

The work on this thesis was performed at the Department of Research and Development (FoU) at Akershus University Hospital (Ahus) from August 2020 to May 2021. The master thesis is the final part of the educational program in the master program of Biomedicine at Oslo Metropolitan University (OsloMet).

Firstly, I would like to thank all my supervisors, Ole Herman Ambur, Alexander Hesselberg Løvestad, and Irene Kraus Christiansen for the warm welcome into your research group, HPVseq. It has been an honor to be supervised by such brilliant experts. Further, I would like to thank my supervisor Ole Herman Ambur giving me the chance to work on this interesting and challenging project. I would also like to thank my supervisor Alexander Hesselberg Løvestad for all the guidance through the practical and theoretical work, especially the bioinformatics part of the project. I would also like to thank my co-supervisor Irene Kraus Christiansen for all the help in understanding the biological aspect of the project. I would also like to express my gratitude to Milan Stosic for helping me in the writing process.

I also wish to thank my colleagues from FoU for their scientific and nonscientific discussions. It made my work at the laboratory much more enjoyable.

Last but not least I would like to acknowledge family and friends who supported me during a hectic and challenging period, especially during this Corona pandemic.

Oslo, May 2021

Adina Repesa

Adina Repesa

## Abstract

**Background:** Human papillomavirus (HPV) is associated with 4.5% of all human cancers worldwide including cervical cancer. Cytological and/or HPV primary screening is used to uncover cancer precursors. Still, better clinical specificity of screening procedures is warranted to decrease unnecessary follow-up and treatment. However, currently there is no ideal secondary diagnostic biomarker for predicting the risk of cancer progression. Viral integration is one of several reported potential biomarkers. Current HPV integration research often uses NGS approaches. NGS is a revolutionary technology but is prone to generating technical artefacts, which warrants validation of reported integrations using other methods, such as Sanger sequencing. Furthermore, most integration studies have focused on HPV16 and 18 because of their high prevalence in cervical cancer cases, and less in other HR-HPV types, such as 31, 33, and 45. **The aim** was to validate and characterize NGS reported HPV integrations in HPV31, 33, and 45 positive samples. **Materials and methods:** LBC samples were obtained from women with HPV31, 33, or 45 positive infections with a diagnostic category of LSIL/ASCUS, CIN2, CIN3, or cancer. The NGS reported HPV integrations were first investigated for known artefacts and filtered out. Subsequently, DNA templates and primer pairs were designed for the qualified HPV integrations containing a human and HPV-specific sequence. Subsequently, the sequences were Sanger sequenced and the data was processed. Finally, hot-spot and microhomology regions were identified. **Results:** 68% (21/31) of the NGS reported HPV integrations in 14 samples were confirmed with Sanger sequencing, accounting for 3.2% (1/31) HPV31 positive sample, 3.2% (1/31) HPV33 positive sample, and 61% (19/31) HPV45 positive samples. Of the confirmed proportion: 95% (20/21) had a CIN3 diagnostic category and 4.8% (1/21) cervical cancer. 24% (5/21) had integrations reported in hot-spot regions and 24% (5/21) were identified with microhomology regions at the integration breakpoints. Two of the confirmed HPV integrations mapped to the tumor suppressors p63 and Wilms protein, suggesting a role of these specific integrations in driving the cancer progression. **Conclusions:** Integrations in HPV45 positive CIN3 samples were significantly higher compared to HPV31 and 33 CIN3 samples. The confirmed HPV integrations were also found in hot-spots and with microhomology regions at the integration breakpoint, suggesting a non-random distribution of integration sites and a fusion between viral and human DNA through the microhomology-mediated DNA-repair pathway.

**Keywords:** HPV31, HPV33, HPV45, cervical cancer, integrations, p53, pRb, microhomology, hot-spot regions, NGS, Sanger sequencing

## Sammendrag

**Bakgrunn:** Humant papillomavirus (HPV) er assosiert med 4,5% av alle humane kreftypen på verdensbasis, inkludert livmorhalskreft. Cytologi og HPV-testing, sammen eller hver for seg, blir anvendt for å avdekke forstadier til kreft gjennom screening. Det finnes imidlertid ingen ideell sekundær biomarkør for å predikere risiko for krefutvikling. Mangel på klinisk spesifitet i screeningprogram kan medføre unødvendig oppfølging og behandling. HPV integrasjon i det humane genom har blitt foreslått som en potensiell biomarkør. I dag bruker de fleste integrasjonsstudier NGS-metoder. NGS er en revolusjonerende teknologi, men som også produserer tekniske artefakter som må kontrolleres med andre metoder, slik som med Sanger sekvensering. I tillegg har de fleste integrasjonsstudier fokusert på HPV16 og 18 grunnet deres høye forekomst livmorhalskreft og færre studier gjort på andre kreftfremkallende (høyrisiko) HPV-typen slik som 31, 33 og 45. **Mål:** Validere NGS-rapporterte integrasjoner i HPV31, 33 og 45 positive prøver. **Materialer og metoder:** Celleprøver fra kvinner positive for HPV31-, 33- og 45 i følgende diagnostisk kategori: LSIL/ASCUS, CIN2, CIN3 eller kreft, ble inkludert i studien. De NGS-rapporterte HPV integrasjonene ble først undersøkt for kjente artefakter som ble filtrert bort. Templatesekvenser og tilsvarende primer-par bestående av human- og HPV-spesifikk sekvens ble laget for de ufiltrerte HPV integrasjonene. Deretter ble integrasjonene Sanger-sekvensert og dataene prosessert. Hot-spot regioner og mikrohomologi regioner ble også identifisert. **Resultater:** 68% (21/31) av de NGS predikerte HPV integrasjonene fra 14 prøver ble bekreftet, tilsvarende 3,2% (1/31) HPV31- positive prøver, 3,2% (1/31) HPV33- positive prøver og 61% (19/31) HPV45- positive prøver. Blant de 21 integrasjonene var 95% (20/21) funnet i prøver med CIN3 og 4,8% (1/21) i prøver med livmorhalskreft. 24% (5/21) av de bekrefte HPV integrasjonene hadde integrasjoner rapportert i hot-spot regioner og 24% (5/21) ble identifisert med mikrohomologiregioner ved integrasjonsbruddspunktet. To av de bekrefte HPV integrasjonene mappet til tumorsuppressorgener som koder for hhv. p63 og Wilms protein. **Konklusjon:** Integrasjon i HPV45-positive CIN3-prøver var signifikant høyere sammenlignet med HPV31 og 33 positive CIN3-prøver. De bekrefte HPV-integrasjonene var også lokalisert i hot-spotregioner og med mikrohomologi områder ved integrasjonsbruddspunktet, som kan indikere en ikke-tilfeldig fordeling av integrasjoner i det humane genom og en fusjon mellom viralt og human DNA gjennom en mikrohomologi-mediert DNA reparasjonsmekanisme.

**Nøkkelord:** HPV31, HPV33, HPV45, livmorhalskreft, integrasjoner, p53, pRb, mikrohomologi, hot-spot regioner, NGS, Sanger sekvensering

## Abbreviations

2vHPV	Bivalent Human papillomavirus vaccine
4vHPV	Quadrivalent Human papillomavirus vaccine
9vHPV	Nonavalent Human papillomavirus vaccine
AC	Adenocarcinoma
ACIS	Adenocarcinoma In situ
ADC	Adenocarcinoma
Ahus	Akershus University Hospital/ Norwegian abbreviation: Akershus universitetssykehus
APOT	Amplification of Papillomavirus Oncogene Transcripts
ASC-H	Atypical Squamous Cells, cannot exclude High-grade lesion
ASC-US	Atypical Squamous Cells of Undetermined Significance
BAM	Binary Alignment Mapping
BER	Base Excision Repair machinery
BLAST	Basic Local Alignment Search Tool
BLASTn	Nucleotide Basic Local Alignment Search Tool
BLAT	Basic Local Alignment- like Tool
bp	Base pair
CDK	Cyclin-Dependent Kinase inhibitor
Chr	Chromosome
CIN	Cervical Intraepithelial Neoplasia
CIN1	Cervical Intraepithelial Neoplasia grade 1/Low -/Mild grade dysplasia
CIN2	Cervical Intraepithelial Neoplasia grade 2/ Moderate dysplasia
CIN3	Cervical Intraepithelial Neoplasia grade 3/High-grade-/ severe dysplasia
CO-bands	Cut out Bands first time
CObands2	Cut out Bands second time
ddNTP	Dideoxynucleotides triphosphates
DDR	DNA Damage Response
dH <sub>2</sub> O	Distilled Water (H <sub>2</sub> O)
DIPS	Detection of Integrated Papillomavirus Sequences
DNA	Deoxyribonucleic acid
dNTPs	Deoxyribonucleotide triphosphates
dsDNA	Double-stranded deoxyribonucleic acid
E gene	Early gene
EDTA	Ethylenediaminetetraacetic acid
F-primer	Forward primer
FHI	National Institute of Public Health/ Norwegian abbreviation: Folkehelseinstituttet
FISH	Fluorescence in Situ Hybridization
FoSTeS	Fork Stalling and Template Switching
GRCh8/hg38	Human Reference genome 38
Hi-Di	Highly Deionized Formamide

HISAT2	Hierarchical Indexing for Spliced Alignment of Transcripts
HIV	Human Immunodeficiency virus
HPV	Human papillomavirus
HR-HPV	High-risk human papillomavirus
HS	High Sensitivity
HSIL	High-grade Squamous Intraepithelial Lesion
HSV-2	Herpes Simplex Virus type 2
IGV	Integrative Genomics Viewer
IHC	Immunohistochemistry
L gene	Late gene
LAST	Local Alignment Search Tool
LBC	Liquid-Based Cytology
LEEP	Loop Electrosurgical Excision Procedure
LR-HPV	Low risk Human papillomavirus
LSIL	Low-grade Squamous Intraepithelial lesions
Min	Minutes
ML	Molecular weight Ladder/ Molecular weight standard
MMBIR	Microhomology-Mediated Break-Induced Replication
MSIS	Communicable Disease Notification System/ Norwegian abbreviation: Meldingssystem for smittsomme sykdommer
NGS	Next-Generation Sequencing
NHEJ	Non-Homologous End Joining
Nucleotide A	Nucleotide Adenine
Nucleotide C	Nucleotide Cytosine
Nucleotide G	Nucleotide Guanine
Nucleotide T	Nucleotide Thymine
ORF	Open Reading Frame
p16	Protein 16
p21	Protein 21
p53	Protein 53
Pap	Papanicolaou
PaVE	Papilloma Virus Episteme
PCR	Polymerase Chain Reaction
pRb	Protein Retinoblastoma
R-primer	Reverse primer
REK	Regional Committee for medical and health research ethics/ Norwegian abbreviation: Regionale komiteer for medisinsk og helsefaglig forskningssetikk
RNA	Ribonucleic acid
ROS	Reactive Oxygen Species
RS-PCR	Restriction Site Polymerase Chain Reaction
S	Seconds
SAM	Sequence Alignment Mapping
SARS-CoV	Severe Acute Respiratory Syndrome Coronavirus

SCC	Squamous Cell Carcinoma
SH3CT2	SH3 Domain and Tetratricopeptide Repeats 2
SINes	Short Interspersed Nuclear elements
SPSS	Statistical Package for Social Sciences
T-PCR1	Traditional PCR, followed by 2% agarose gel run at 100 Volt for 60 minutes
T-PCR2	Traditional PCR, followed by 2% agarose gel run at 70 Volt for 60 minutes
T-PCR3	Traditional PCR, followed by 2% agarose gel run at 70 Volt for 120 minutes
TAE	Tris-Acetate
TaME-seq	Tagmentation-assisted Multiplex Polymerase chain reaction Enrichment sequencing
TCGA	The Cancer Genome Atlas
TD-PCR	Touch Down Polymerase Chain Reaction
TD-PCR1	Touch Down Polymerase Chain Reaction, followed by 2% agarose gel run at 70 Volt for 60 minutes
TD-PCR2	Touch down Polymerase Chain Reaction, followed by 2% agarose gel run at 70 Volt for 135 minutes
TP63	Tumor protein 63
URR	Upstream Regulatory Region
UV	Ultraviolet
V	Volt
VLPs	Virus-like particles
WHO	World Health Organization

# Table of contents

<b>1. Introduction.....</b>	<b>1</b>
1.1 <i>HPV and cervical cancer</i> .....	1
1.2 <i>HPV Infection, pathology and cancer progression</i> .....	2
1.3 <i>HPV genome structure</i> .....	4
1.3.1 Classifications of Human papillomaviruses (HPVs).....	6
1.4 <i>HPV life cycle</i> .....	8
1.5 <i>Viral oncogenes and human tumor suppressor genes</i> .....	10
1.5.1 HPVE6 and tumor suppressors .....	10
1.5.2 HPVE7 and tumor suppressors .....	10
1.6 <i>Mechanism of HPV integration</i> .....	11
1.6.1 Looping model .....	12
1.6.2 Microhomologies .....	14
1.7 <i>Position of HPV integrations</i> .....	14
1.7.1 Hot-spot regions .....	14
1.7.2 Transcriptionally active regions.....	15
1.7.3 Viral integration in HPV 16, 18, 31, 33 and 45 positive samples.....	15
1.8 <i>Methods for detecting HPV integrations</i> .....	16
1.8.1 Polymerase Chain Reaction (PCR) .....	16
1.8.2 First generation sequencing.....	16
1.8.3 Next-generation sequencing (NGS) .....	17
1.9 <i>Cervical cancer prevention</i> .....	17
1.9.1 HPV vaccination .....	17
1.9.2 Cervical cancer screening .....	18
1.10 <i>Treatment of cervical lesions</i> .....	20
<b>2. Aims of study.....</b>	<b>21</b>
<b>3. Materials and methods .....</b>	<b>22</b>
3.1 <i>Study population and specimen collection</i> .....	22
3.2 <i>Validation of HPV integration sites</i> .....	22
3.2.1 Read alignment and visualization on Integrative Genomics Viewer (IGV) software .....	23
3.2.2 Categorizing potential HPV integrations or potential artefacts?.....	24
3.3. <i>In silico DNA template for primer design</i> .....	28
3.3.1 Discordant reads.....	29
3.3.2 Junction reads.....	29
3.4 <i>Primer design</i> .....	30
3.5 <i>Sample preparation and DNA extraction</i> .....	31
3.6 <i>Measurement of DNA concentration</i> .....	31
3.7 <i>Amplification by Polymerase Chain Reaction (PCR)</i> .....	32
3.8 <i>Analysis of PCR product</i> .....	33
3.9 <i>Purifying DNA fragments from Gel</i> .....	34
3.10 <i>Preparing sequencing PCR</i> .....	35
3.11 <i>Precipitation of PCR sequencing products</i> .....	35

3.12 Sanger sequencing .....	36
3.13 Processing sequencing data.....	36
3.14 Determining microhomology regions .....	38
3.15 Statistical methods .....	38
<b>4. Results .....</b>	<b>39</b>
4.1 Categorization of NGS-reported HPV integration sites .....	39
4.1.1 Qualified HPV integrations.....	40
4.2 Making template sequences .....	40
4.3 Forward and Reverse primer design .....	41
4.4 Semi-quantitative and a qualitative validation of the Polymerase Chain Reaction (PCR) products.....	41
4.4.1 T-PCR .....	42
4.4.2 TD-PCR.....	42
4.5 DNA elution from gel-bands .....	44
4.6 Sanger sequencing data analysis .....	45
4.6.1 Confirmed HPV integrations.....	45
4.6.2 Non-confirmed HPV integrations .....	47
4.7 Microhomology regions .....	48
<b>5. Discussion .....</b>	<b>52</b>
5.1 Clinical aspects .....	52
5.1.1 Higher integration rate in HPV45 positive samples with a CIN3 diagnostic category.....	52
5.1.2 HPV45 positive samples with more than one reported HPV integration.....	53
5.1.3 Localization of HPV breakpoints and integrations .....	54
5.1.4 Why chromosomal integrations as biomarkers? .....	56
5.1.5 The corona pandemic and increasing HPV research? .....	56
5.2 Methodological consideration .....	57
5.2.1. Sample material.....	57
5.2.2 NGS reported data .....	57
5.2.3 Validation of HPV integrations.....	57
5.2.4 Template design .....	58
5.2.5 Primer design .....	59
5.2.6 Agarose gel electrophoresis and visualization of the PCR products.....	60
5.2.7 DNA elution from gel-bands.....	60
5.2.8 Analyzing Sanger sequencing data .....	60
5.3 The strengths and limitations of the study .....	62
5.3.1 Strengths.....	62
5.3.2 Limits .....	62
<b>6. Conclusion and further research.....</b>	<b>63</b>
<b>7. Literature list .....</b>	<b>64</b>
Appendix .....	
Appendix 1. REK approval .....	
Appendix 2. Data Protection Office at Ahus approval .....	
Appendix 3. Template sequences and primer pairs .....	
Appendix 4. Agarose gel runs .....	
Appendix 5. Sanger sequences from confirmed HPV integrations .....	
Appendix 2. Data Protection Office at Ahus approval .....	

## 1. Introduction

Human papillomavirus (HPV) is one of the most prevalent causes of sexually transmitted infections in both men and women (2). The virus is associated with 4.5% of all human cancers worldwide (3), including cervical cancer. Cancer refers to a group of diseases characterized by uncontrolled cell growth and cell division (4). Several risk factors for the cancer progression have been identified, including chromosomal integration, a process where a part of the viral genome attaches to the linear host genome and becoming its part (5-8). The integration process may promote cancer development by disrupting specific genes in the virus and/or host genome. Interrupted host genes may have important functions in regulating the host cell cycle (2, 6, 9, 10), increasing the risk of cancer development. Currently, there is no ideal biological diagnostic marker (biomarker) for predicting the cancer progression that may lead to unnecessary follow-up and treatment of women with minimal risk of developing high-grade lesions or cancer. Viral integration has been reported as an early event and a potential biomarker for predicting progression from lesions to cervical cancer (11). Current HPV integration research employs Next-Generation Sequencing (NGS) approaches (5, 12). NGS technology is revolutionary, but it also produces a lot of artefacts that may lead to false-positive results. Therefore, several studies have used Sanger sequencing as a gold standard to confirm NGS data (12, 13). Although NGS technology can reveal genomic information about the HPV integrations, the integration studies have mostly been focused on HPV16 and 18. However, as HPV31, 33, and 45 are also considered high-risk HPV (HR-HPV) types, these types need to be encompassed by the HPV integration studies.

### 1.1 HPV and cervical cancer

Cervical cancer is the fourth most common cancer type in women (14, 15), and in 99% of the cervical cancer cases, the disease has been linked to an HR-HPV (14). The correlation between genital HPV infection and cervical cancer was first described in the 1980s by the German virologist, Harald zur Hausen (2); a discovery that awarded him the Nobel prize for medicine in 2008 (16). HPV infection is also associated to penile, vulvar, vaginal, anal, and head and neck cancer but not so strongly as in cervical cancer cases (17). In 2018, the World Health Organization (WHO) estimated the number of women diagnosed with cervical cancer to 570, 000 with more than half of those dying from the disease (14). In Norway, in 2019, 368 were diagnosed with cervical cancer with an estimated death rate of 85 women (18). This is a

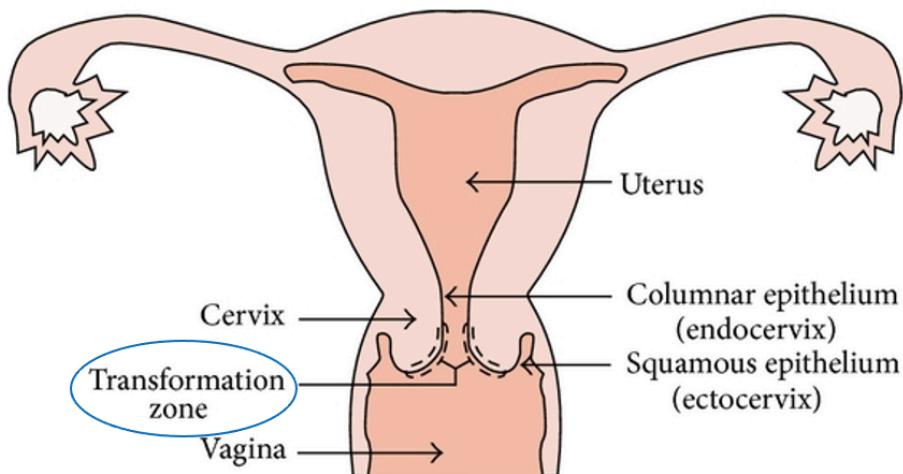
reduction from earlier decades because of organized preventative actions such as extensive screening (19). Developing countries are of major concern as they lack these organized programs. This may be the reason for the majority of cervical cancers and deaths occurring in less developed countries, accounting for more than 85% of all cervical cancer deaths (15, 20). The concern becomes even greater with the fact that HPV infections are most prevalent among young, sexually active women. However, because HPV-caused cervical cancer develops slowly in women with a normal immune system, it will mainly arise in women at their reproductive age (21, 22). This may cause complications during pregnancy and childbirth highlighting the importance of global preventative strategies. Therefore, early detection of HPV infection is crucial to limit viral pathogenesis and a potential cancer progression.

## 1.2 HPV Infection, pathology and cancer progression

Some of the risk factors for HPV infection are the number of sex partners and previous exposure to sexually transmitted diseases such as Chlamydia trachomatis, Herpes Simplex Virus type 2 (HSV-2), and The Human Immunodeficiency Virus (HIV) (21-23). Especially immunosuppressed women as those infected with HIV are at higher risk for HPV infection, persistence, and progression from HPV lesions (abnormal tissue changes) to cervical cancer (24). A correlation between countries with high HIV prevalence and cervical cancer deaths has been reported (25). Various studies have different definitions of persistent infection, however the most of them describe it as two positive HPV samples in 6-12 months (26).

HPV transmits through direct contact with infected regions of the skin or mucous membrane (27). The virus colonizes the lower portion of the uterus, part of the female reproductive tract also known as the cervix (28) (Figure 1). The virus infects basal epithelial cells at the squamocolumnar junction, a line separating ectocervical squamous and endocervical columnar epithelium. The columnar epithelium is replaced by squamous epithelium over time depending on biological changes in women (age and hormonal status). The squamocolumnar junction is localized towards the ectocervix during puberty and moves towards the endocervix years later. A new- squamocolumnar junction is found between the newly formed squamous epithelium and the columnar epithelium, and the metaplastic epithelium is referred to as the transformation zone (29), marked with the blue circle in Figure 1. Metaplasia is referred to as a change or replacement from one epithelium type to another (29). The transformation region is especially susceptible to carcinogenesis (28, 30, 31). As the transition zone includes two types of epithelial

cells, glandular and squamous, two types of cancers, adenocarcinoma (ADC) and squamous cell carcinoma (SCC) can arise in the cervix (32). SCC occurs in the squamous cells located on the outer level of the cervical canal, while ADC develops in the glandular cells located on the inner level of the cervical canal. The prognosis of SCC is better than ADC (33). HPV18 and 45 are more prevalent in ADC (34), and HPV16, 31, and 33 in SCC (35). One study reported more cases of viral integration in SCC than in ADC (12).

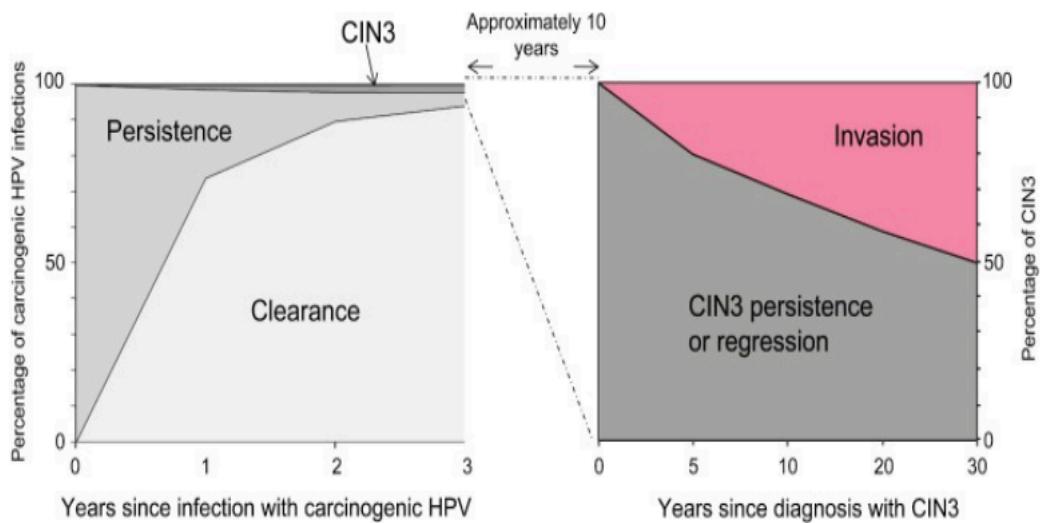


**Figure 1: The reproductive tract of women.** The lower portion of the uterus named cervix consists of two different epithelia, glandular/columnar and squamous. The transformation zone (circled in blue) is a region especially susceptible to carcinogenesis. Figure obtained and reconstructed with permission from Bengtsson et al. (36).

Most sexually active women will undergo an HPV infection in their lifetime (27). Large proportions of the individuals infected are without noticeable symptoms. Therefore, the virus is most frequently transmitted unknowingly (37). Not knowing when an infection started or for how long it has persisted leads to late discoveries of cytological abnormalities.

The natural history of cervical cancer is the gradual progress from low cervical intraepithelial neoplasia (CIN) Grade 1 (CIN1) (abnormal tissue growth) to CIN Grade 2 (CIN2) (moderate) and finally to CIN Grade 3 (CIN3) (severe neoplasia and micro-invasive lesions (abnormal tissue change) (2). In 90% of the cases, the infection disappears within several months while in 10% of the cases the infection persists and may progress to invasive cancer (8). The progression towards cancer usually takes 10 to 20 years, however, some lesions become cancerous more rapidly, in a year or two (2). Infections with multiple HPV types may increase the risk of a persistent HPV infection (38). This might be caused by the host immune system being under stress from fighting multiple HPV types, requiring more time to combat each type.

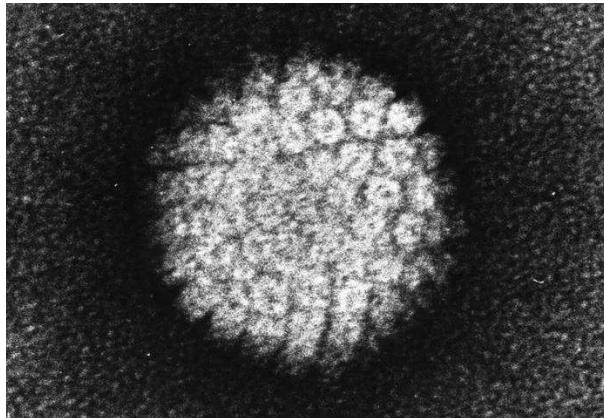
Figure 2 illustrates the percentage of carcinogenic HPV infections (y-axis) and years (x-axis) of HPV persistence and clearance. The figure illustrates the low percentage of cases that progresses to CIN3 (left graph, Figure 2), the high percentage of persistence or regression after 10 years with CIN3, and the small percentage that progresses to invasion if no treatment has been offered (right graph, Figure 2) (39). As only a small percentage progresses to invasion it is important to identify the CIN cases that could potentially progress to invasive cancer. According to the WHO reports, two of the symptoms of invasive cervical cancer are pelvic pain and irregular bleeding (27). Viral integration has been reported as an important factor in the progress from precancerous lesions (CIN2/CIN3) to invasive carcinoma (40). HPV can occur in different genomic structures with genes encoding crucial events in the virus's life cycle.



**Figure 2: Human papillomavirus (HPV) infection and carcinogenesis.** The left graph illustrates the percentage of HPV infections that regresses and the small percentage that persists and may progress to cervical intraepithelial grade 3 (CIN3). The right graph illustrates the high percentage of CIN3 persistence or regression after 10 years with CIN3 and the small percentage that progresses to invasive cancer if no treatment has been performed. Figure obtained with permission from Schiffman et al. (39).

### 1.3 HPV genome structure

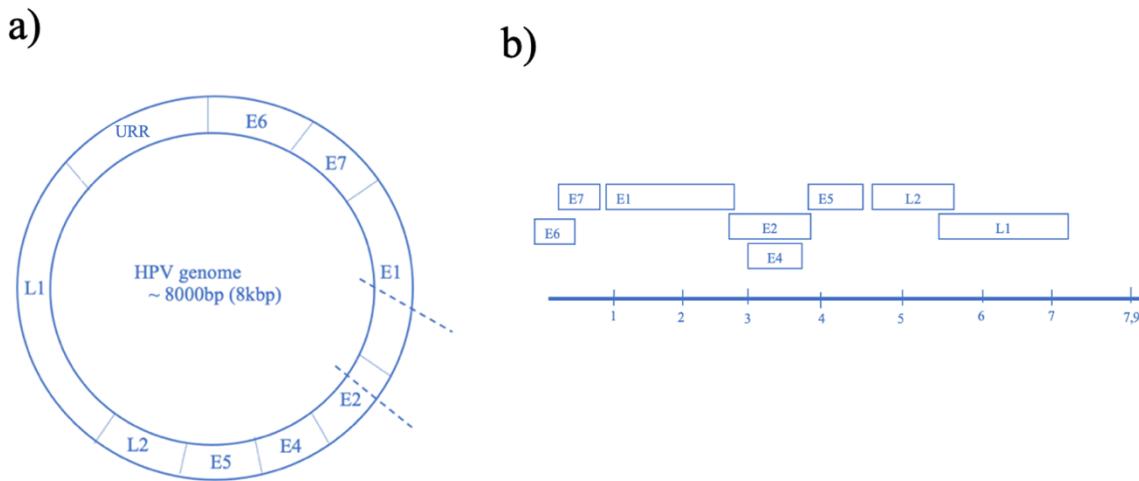
When viewed in an electron microscope the HPV virion may resemble a golf ball because of its circular form and the small patches on the surface, as seen in Figure 3 (2).



*Figure 3: Human papillomavirus (HPV) viewed in an electron microscope. Figure obtained from (41).*

In an infection, the virus can both be presented in an episomal or linear form, or a combination (42). When in an episomal form, HPV is circular and may produce viral particles (43), while the linear form is how the virus genome is seen in when integrated. Figure 4 exhibits two HPV structures, circular (Figure 4a) and linear (Figure 4b) with the gene positions approximately adjusted to HPV31, 33, and 45 by use of The Papilloma Episteme Database (PaVE) (44, 45). The specific gene regions for each of the HPV types are presented in Table 1.

HPV is a double-stranded deoxyribonucleic acid (dsDNA) virus, and the genome consists of approximately eight thousand base pairs (bp) (8kbp) (2, 42). The genome contains eight genes classified into early (E) and late (L) genes depending on the timely expression at different stages during the viral life cycle. HPV genome is composed of six early genes, E1, E2 E4, E5, E6, and E7, two late genes, L1 and L2, and a noncoding upstream regulatory region (URR) (2, 42). The early genes have essential roles in replication, cellular transformation, and viral transcription and are also involved in oncogenesis (2, 46). A break in the E1 and E2 genes as a potential outcome of an HPV integration can lead to a disruption or deletion that hinders an optimal function of HPV genes, causing a malignant transformation (2, 9, 10). An example of dysregulatory breaks is marked with striped lines in Figure 4a. The late gene products are important for the structural capsid proteins and the virion assembly (46). L1 encoding for the major virion particle is the most conserved gene used to classify HPVs into distinct types and as the target epitope in vaccine production (2). The URR region is responsible for the replication and transcription of viral DNA (46).



**Figure 4: Human papillomavirus (HPV) genome episomal (4a) and integrated (4b).** The genome consists of ~8000 base pairs constituting 8 genes divided into early (E)(E1-E7), late(L) genes (L1, L2), and an Upstream regulatory region (URR). The striped lines seen in the episomal structure to the left, represent breaks in the E1 and E2 gene often associated with integration events causing malignant transformation. The gene positions have been approximately adjusted to HPV31, 33, and 45 by use of Papillomavirus Episteme (PaVE) database (44, 45). The figures have been reconstructed with permission from Raybould et. al (47)

**Table 1: Human papillomavirus (HPV) genes and specific sequences.**

Gene	Function	HPV31-Sequence (bp)	HPV33-Sequence (bp)	HPV45- Sequence (bp)
L1	Major capsid protein	5532-7066	5594-7093	5608-7147
L2	Minor capsid protein	4171-5571	4210-5613	4236-5627
E1	DNA replication	862-2751	879-2813	914-2845
E2	Negative regulator of transcription	2693-3811	2749-3810	2769-3875
E4	Maturation, virion release	E1^E4: 862-877, 3295-3578	E1^E4: 879-894, 3351-3577	E1^E4: 914-929, 3392-3648
E5	Maintenance of and proliferation	3816-4070	3854-4081	3909-4130
E6	Viral oncoprotein	108-557	109-558	102-578
E7	Viral oncoprotein	560-856	573-866	587-987

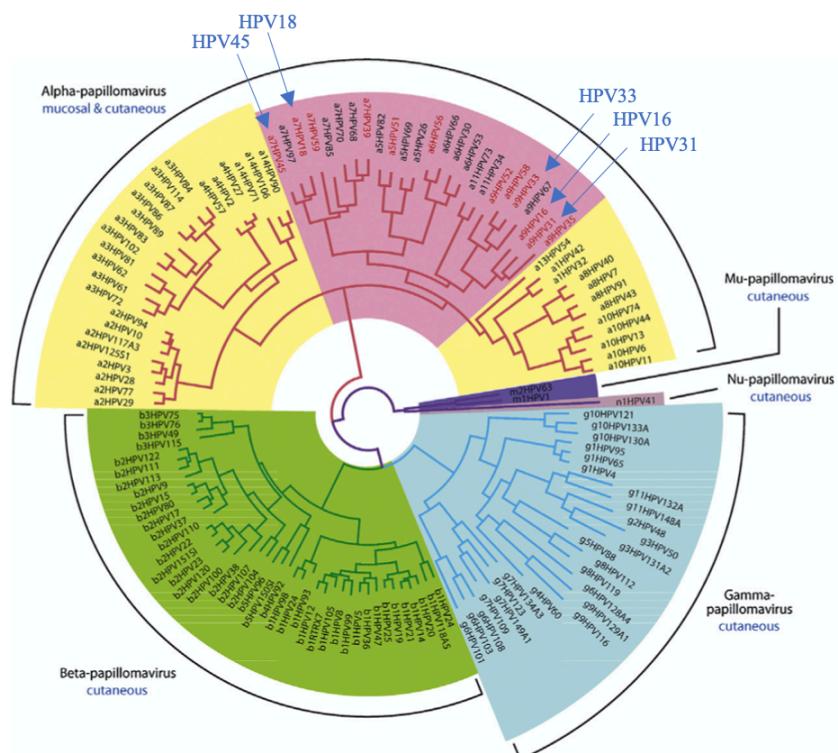
<sup>1</sup> Shows an overview of the 8 genes in HPV 31, 33, and 45, the function and localization of each gene in the genome provided in base pairs (bp). The gene positions are specific adjusted for HPV31, 33, and 45 by use of the Papillomavirus Episteme (PaVE) database (44, 45). Abbreviations: DNA= Deoxyribonucleic acid

### 1.3.1 Classifications of Human papillomaviruses (HPVs)

HPV belongs to the family *Papillomaviridae* and is divided into genera, among these *alphapapillomavirus*, *betapapillomavirus*, and *gammapapillomavirus* (48). Members of the *Alphapapillomavirus* are strongly linked to cervical cancer development (28, 49, 50). Currently, more than 200 HPV types have been identified (2, 9), with over 40 of them capable of infecting the human anogenital tract (21). The HPV types are categorized into HR and low-risk (LR-

HPV) based on their carcinogenic potential. 14 HR-HPV have been noted, besides HPV16 and 18, HR-HPVs also encompasses HPV31, 33, and 45 (49). HPV16 (57%) is the most frequently detected high-risk type, followed by type 18 (16%), 31 (4%), 33 (5%), and 45 (5%) (51). HPV16 and 18 are responsible for 70% of the HPV infections that are found in invasive cancer (49). 12 HPVs are LR, of which the most frequent types are 6 and 11 (49), usually causing benign genital warts (52).

L1 is the most conserved gene and is usually used to classify new Papillomaviruses (53). HPV types shares at least 90% sequence similarity in the L1 open reading frame (ORF) (10). A phylogenetic tree presented in Figure 5 exhibits the evolutionary relationship between different HPV types based on the alignments of E1, E2, L1, and L2 gene sequences (52, 54). The evolutionary distance between HPV18 and 45 is small, as well as the distance between HPV16 and 31. Although HPV33 is further apart from HPV16 and 31, this type is still considered related to HPV16 and 31. HPV types close on the phylogenetic tree may share similar biological or pathological characteristics (53).



**Figure 5: Phylogenetical tree of Human papillomaviruses (HPVs).** Representing sequence similarities in the E1, E2 (early), L1, and L2 (late) genes between HPVs. HPV45, HPV18, HPV33, HPV16, and HPV31 are pointed out with blue arrows to mark the small distances. HPV18 is close to HPV45, similar to HPV16 and HPV31. While HPV33 is more distant. The figure is obtained and reconstructed with permission from Egawa et al. (52).

## 1.4 HPV life cycle

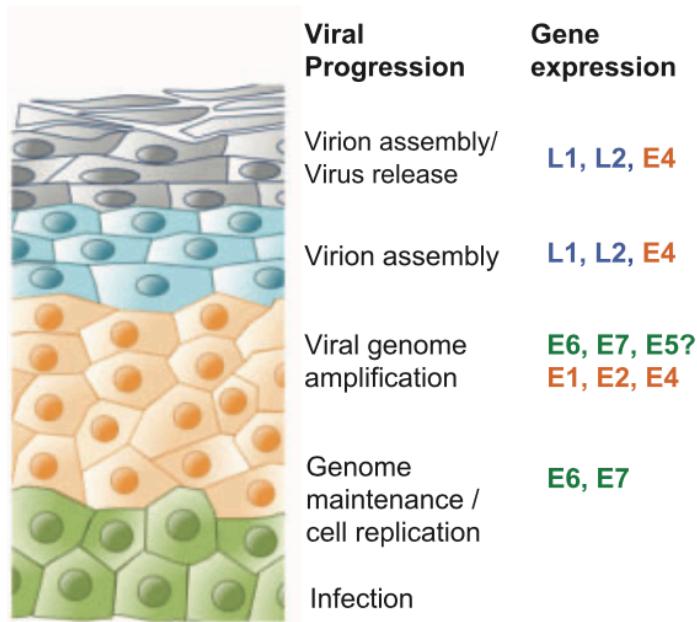
Papillomaviruses have a different life cycle than most other virus families as they need epidermal- or mucosal epithelial cells under continuous proliferation (55). HPV encodes only eight genes and uses host cell factors to regulate viral transcription and replication (2, 31). In this way, the virus can survive and transmit further.

HPV access the basal lamina through micro-wounds (55), and binds receptors named heparan sulfate proteoglycan on the basal cells. The receptors recognize L1 and L2 (minor) capsid proteins on the surface of the virus (47, 56). The early genes (E1, E2, E5, E6, and E7) are expressed early in the epithelium differentiation stage, E4 is expressed throughout the whole life cycle, and the late genes (L1 and L2) at the late stages (31). When HPV infects, the host cell factors interact with URR and initiate E6 and E7 transcription. E6 and E7 gene products functions to destabilize the cell growth-regulatory pathway to optimize the conditions for viral replication. The HPV DNA replication starts as the basal cells differentiate and move towards the epithelial surface. In the replication progress, viral DNA becomes settled in the entire epithelium while intact virions are only found in the upper layers. Conversely, in warts, the cells replicate and proliferate in all the epidermal layers except the basal layer (2).

Figure 6 demonstrates the epithelial layers and genes expressed at different differentiation stages in the life cycle. The E1 gene encodes a helicase (42), an enzyme important to unwind the complementary strands during viral DNA replication (57). The E2 gene encodes a DNA binding protein functioning as a negative regulator of E6 and E7 expression. When E2 down-regulates E6 and E7 transcription, the normal cell cycle and host differentiation process can continue (2). The E5 gene has an important role in maintaining the cell genomes and proliferation (58), whereas E4 is responsible for the maturation and release of viral particles (2). L1 and L2 proteins pack the viral genomes on the surface (30). When differentiated epithelium cells access the surface, the cells will release viral particles as part of the renewal process (2, 6). The virus will take advantage of this renewal process (6), to transmit and infect another host.

The papillomavirus lifecycle usually takes 2-3 weeks, representing the time required for a cervical cell to migrate from the basal layer to the upper epithelial layers, mature, undergo senescence, and die (59). The long differentiation process may reflect a defense mechanism of the virus to avoid a strong immune response from the host. If the virus had attacked the host

more aggressively, it would risk being discovered by the host immune cells. To complete the lifecycle, the cell must reach the terminal differentiation step, an important step for viral construction and release.



**Figure 6: Human papillomavirus (HPV) life cycle.** The figure demonstrates epithelial layers, and at which stages in the differentiation process, the genes are expressed. Early genes (E1-E7) are mainly expressed at early stages, whereas the late genes (L1 and L2) are expressed at late stages. Figure has been reconstructed with permission from Bravo et al. (60).

When the virus genome is in an episomal form it can create virions that may release to infect another host, but when the virus is being in an integrated state it may change cell functions that facilitate replication of the viral genome (42). HPV integration promotes carcinogenesis in various ways, the most important is the disruption or break in the E1 and E2 ORFs resulting in overexpression of the E6 and E7 oncogenes (2, 9, 10). Overexpression of the oncogenes may increase the negative interactions with human tumor suppressors to cause genomic instability and loss of cell-cycle control (2, 5, 51). HPV integration is not a normal part of the viral life cycle but may represent an intra-host selection advantage (61) and set a course towards malignant transformation. However, the integration is still a dead end for the virus as it can no longer form an episomal genome that can be packed and transmitted to a new host (6). Disruption of the viral oncogenes and human tumor suppressor genes are highly connected to carcinogenesis.

## 1.5 Viral oncogenes and human tumor suppressor genes

The viral oncogenes E6 and E7 complement multiple characteristics for cancer development, so-called hallmarks of cancer, a concept developed by Weinberg & Hanahan (62, 63). These characteristics are: 1) the ability of E7 to induce angiogenesis, 2) E6 and E7 evading the immune system through inhibition of interferon signaling, and 3) deregulating cellular energetics, and 4) inducing genomic instability and mutation. The latter is the most important factor achieved through viral integration proceeding towards dysregulation of cellular pathways (64).

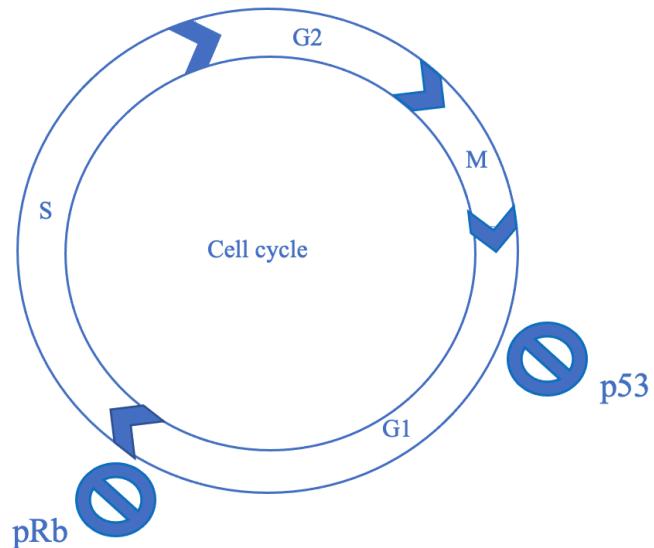
### 1.5.1 HPVE6 and tumor suppressors

HR- HPVE6 binds the tumor suppressor protein 53 (p53) and marks it for degradation through a ubiquitin ligase (2, 64). p53 is referred to as the guardian of the genome, because of its important role in preventing tumorigenesis. Its major function is to regulate the host cell-cycle by preventing cell proliferation of cells with damaged genome (65). p53 protects the genomic integrity by either triggering apoptosis or inducing cell-cycle arrest in G1 (cell growth) until the damaged genome is repaired (2, 51). This is usually done by p53 activation of p21 which further interacts with the protein retinoblastoma (pRb) (2). Mutated p53 has been observed in around half of all human cancers indicating the importance of the protein as tumor suppressor (65). However, in cervical cancer, p53 is a usually wild-type and not mutated (2). In cervical cancer cases, the overexpression and attachment of E6 to p53 inhibits the proteins' activity directly. Conversely, LR-HPVE6 does not bind p53 at detectable levels (2). Figure 7 illustrates the host cell cycle and where in the cell cycle the tumor suppressor is active. E6 can also inhibit other cell cycle regulators such as BAK, C-MYC, and Paxicillin. BAK regulates apoptosis, C-MYC drives proliferation, and Paxicillin is involved in the regulation of the actin cytoskeleton linked to tumor metastasis (66). E6 can also induce the expression of telomerase, an enzyme associated with replicative immortality (64). A potential malignant outcome is further enhanced if the pRb activity is also reduced by HPVE7.

### 1.5.2 HPVE7 and tumor suppressors

E7 mainly contributes to oncogenesis through interaction with the family pRb, which comprises RB1, RBL1, and RBL2 (51). pRb regulates the transition from G1 to S phase at the start of the cell cycle (Figure 7) (2). The S-phase is one of the major phases in the cell cycle where the replication of the whole genome takes place (67). When pRb binds and inhibits E2F

transcription factors, the damaged cell is unable to enter the S-phase (68, 69). HPVE7 binds pRb and targets it for degradation. This results in the release and activation of the E2F transcription factor allowing the cell to enter the S phase, even though the cells' genetic material is damaged (2, 51). For LR- HPVs, the E7 affinity for pRb is reduced (2). E7 can also cause upregulation of protein 16 (p16), a cyclin-dependent kinase inhibitor (CDK), acting as a tumor suppressor in the cell cycle (42).



**Figure 7: Host cell cycle and important tumor suppressors.** Illustration showing where in the host cell cycle the tumor suppressors, protein 53 (p53) and protein Retinoblastoma (pRb) are active.

## 1.6 Mechanism of HPV integration

The integration mechanism is a biological process found in several viruses including retroviruses (70). It has also been reported that the currently severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) ribonucleic acid (RNA) may be reverse- transcribed and integrated into the human genome (71). However, the HPV integration mechanism is not fully understood as the virus does not encode a protein facilitating the integration such as integrase in retroviruses (70). HPV integration can be detected in a precancerous lesion, but the integration frequency increases towards the development of invasive cancer (8, 12, 72). Although HPV integrations are frequently seen in HPV-related cancer, it is not a necessary event for cervical cancer progress (6). There is a huge research field studying alternative ways to produce neoplasia, whether there are HPV integration sites in the host, epigenetic factors of E2, external factors like viral type, and viral load (6, 42).

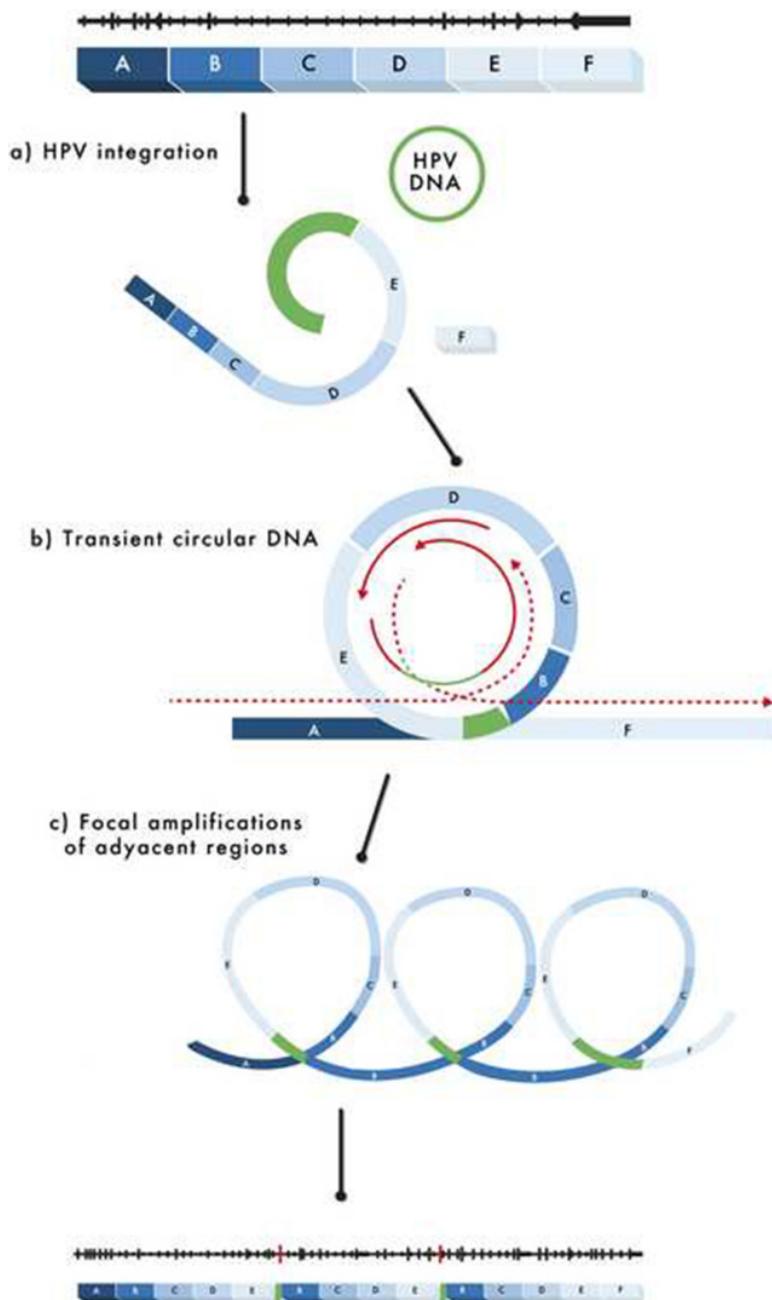
The HPV integration process mainly starts with DNA damage or double-strand break potentially caused by oxidative stress or HPV proteins (7). Oxidative stress is caused by an imbalance between the production and accumulation of reactive oxygen species (ROS) in cells and tissues (73). Increased oxidative stress is associated with inflammation, chemical stress, ultraviolet (UV) exposure, and oxidative phosphorylation. Inflammation and oxidative stress are considered cofactors for enhancing viral integration and deregulation of cellular and viral oncogenes. DNA damage induces the DNA damage response (DDR) pathway to repair the damage before continuing the cell to the cell division. p53 is required for recruiting the base excision repair machinery (BER) to repair the oxidative damage (70). The viral oncoprotein E6 and E7 disrupt cell cycle checkpoint controls by inactivating p53 and pRb. As a result, damage response fails to repair the break.

Unrepaired breakpoints are essential for an HPV integration to occur (47, 70). Breaks in HPV DNA are most likely introduced by the E1 gene product during the viral replication process, followed by a failure to repair the break. This can cause re-circularization to fail forcing the HPV DNA to remain in its linear state necessary for integration to occur. When breaks in both the human and HPV genome appear, a fusion can occur between the genomes either through homologous or nonhomologous recombination. The non-homologous end-joining (NHEJ) is considered likely involved in HPV integration, leading to incorporation of HPV DNA. The presence of short identical sequences at the integration breakpoint, named microhomology regions, may indicate a microhomology-mediated DNA repair pathway during the fusion of human and viral DNA. Although microhomology sequences are observed at the integration, it is not a necessary event for an integration to occur (70). Two events have been described that could mediate the HPV integration into the human genome, a single genome integration into the cellular DNA or integration as multiple tandem head-to-tail repeats (6), involving different recombination models.

### 1.6.1 Looping model

The looping model is the most established model demonstrating the occurrence of two breakpoints found in the HPV16 positive SiHa cell line (42). The model states that HPV integration is mediated by DNA replication and recombination that may lead to DNA concatemers (several copies of the same DNA sequence) (74). As illustrated in Figure 8, the process starts with HPV integration between region E and F (Figure 8a), followed by forming a short-term circular DNA that includes the viral sequences (Figure 8b). Meanwhile, the DNA

polymerase initiates replication and forms concatemers while focal amplification and rearrangements are made close to the viral integration (Figure 8c) (42). This process can lead to disruption of genes participating in tumorigenesis, oncogene amplification, inter or intrachromosomal rearrangements, and/or genetic instability (74).



**Figure 8: Looping model found in the SiHa cell line.** The model describes an integration event that can cause formation of deoxyribonucleic acid (DNA) concatemers. The green region demonstrates the Human papillomavirus (HPV) genome. Figure obtained with permission from Oyervides-Muñoz et al.(42).

## 1.6.2 Microhomologies

Another integration model based on microhomologies has been proposed by *Hu. et al* (12). The model is based on the discovery of microhomology-rich regions between the viral and the host genome near the integration sites (12). The illustration of microhomology-sequences in the viral and host genome is presented in Figure 9. *Hu. et al.* (12) have focused on two integration mechanisms: fork stalling and template switching (FoSTeS), and microhomology-mediated break-induced replication (MMBIR). FoSTeS is based on viral genome integration during a pause in the replication fork. The HPV hijacks the pathway and exchanges the host genome template to be able to integrate its own. On the other hand, in MMBIR, the replication break is mediated by microhomologies where HPV integrates its genome into the host DNA during the replication. Both FoSTeS and MMBIR are activated during HPV infection, especially in the presence of a break in repetitive genomic elements such as satellite DNA, Alu elements and Short Interspersed nuclear elements (SINEs). The repetitive genomic elements form microhomologies flanking the breakpoint. These formations enable HPV to hijack the DNA repair pathways to fuse its genome with the host's damaged chromosome (Chr) (42). In some cases, the position of chromosomal HPV integrations might not be randomly distributed.

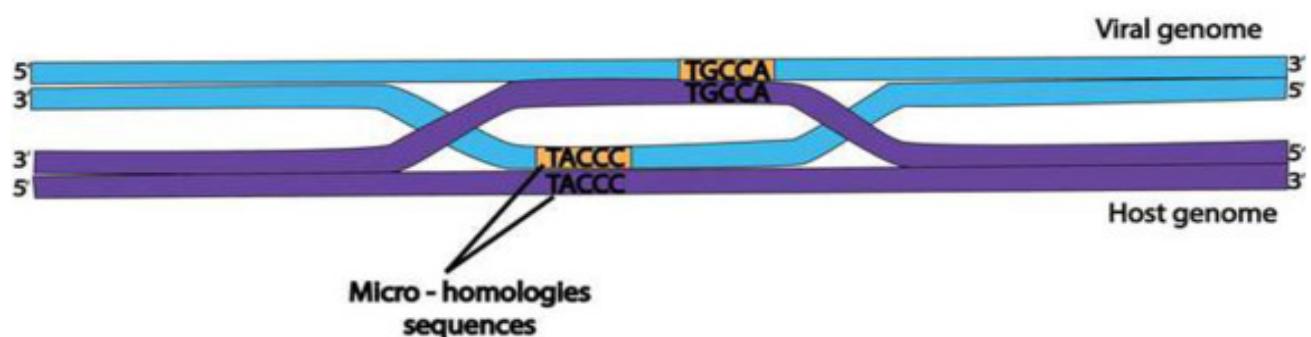


Figure 9: Microhomology regions in the viral and host genome. Figure obtained with permission from Oyervides-Muñoz et al. (42)

## 1.7 Position of HPV integrations

### 1.7.1 Hot-spot regions

HPV integrations are dispersed across the human genome (75). However, integrations in some chromosomal loci such as 1p, 3q, 6q, 11, 13q, and 20q have been reported more frequently than others (9). Among the specific regions reported are 3q28, 4q13.3, 8q24.21, 17q21 and 13q22.2 (9, 40, 70, 76). The hot-spot regions are usually associated with common fragile sites (40, 77),

defined as an unstable region susceptible to break (47). Integration near or in fragile sites has frequently been reported (9, 75, 78-80).

### 1.7.2 Transcriptionally active regions

As reported previously, HPV frequently integrates into genes that are constantly expressed during DNA transcription and repair (9, 75, 81). This may be caused by histones that do not tightly pack the DNA strand in gene regions containing frequently expressed genes. In this way, HPV can easier gain access to the host DNA strand. When integration occurs in or close to constantly expressed genes it may affect cell growth and proliferation (9). *Das et al.* showed that most of the detected integrations, were located within or nearby specific genes, such as proto-oncogene MYC and tumor protein 63 (TP63) (9). Integrations located in MYC have also been reported by *Hu et al.* (12). Interestingly, some HR-HPV types exhibited a higher integration frequency than others.

### 1.7.3 Viral integration in HPV 16, 18, 31, 33 and 45 positive samples

A study conducted by *Vinokurova et al.* noted that the HPV integration frequency may be type-dependent (8). The study showed that HPV16, 18, and 45 were more frequently able to generate the integration in comparison with type 31 and 33. Furthermore, in CIN3 lesions, 60% of HPV45-positive cases showed integrated HPV DNA, 19% of the HPV16 cases, and 10% of the HPV31 cases, whereas, in HPV18 and 33, no integration was detected. In cervical cancer samples, HPV18 stood for the highest integration frequency with 92%, followed by HPV45 (83%), HPV16 (55%), HPV33 (37%), and HPV31 (14%) (8). This distribution of frequencies may be associated with the close phylogenetic relationship between HPV18 and 45, as well as between HPV16, 31, and 33. A comprehensive study organized by The Cancer Genome Atlas (TCGA) also showed a high integration frequency of HPV18 in cervical cancer (82). TCGA is a public project responsible for studying alterations in cancer genome profiles to uncover possible prognostic or therapeutic markers. The study involved a population of 178 samples with the integration rate of 100% in HPV18 positive tumors and up to 80% in HPV16 positive tumors (82). Another cervical cancer study conducted in India including a population of 270 samples showed opposite results, with a higher integration frequency in HPV16 than in HPV18 (9). The number of studies reporting integration sites has increased in the last years due to the access to novel technologies enabling breakpoint detection in the viral genome and integration sites in the host genome (42).

## 1.8 Methods for detecting HPV integrations

HPV integration studies have been performed on clinical samples and various carcinoma cell lines such as CaSki, HeLa, and SiHa (47). CaSki and SiHa are HPV16 positive cell lines (83), while HeLa is an HPV18 positive cell line (84). CaSki is a human cervical carcinoma cell line from a 40 year old female Caucasian patient (85). Throughout the years, a broad selection of molecular methods has been employed to detect the viral state, such as immunohistochemistry (IHC), Fluorescence in situ hybridization (FISH), and Southern blot until more sensitive methods as Polymerase chain reaction (PCR) were introduced, followed by revolutionary NGS technologies (42, 47).

### 1.8.1 Polymerase Chain Reaction (PCR)

PCR is a commonly used technique because of the high sensitivity (HS) and specificity (11). PCR is beneficial to detect and amplify DNA to an optimal concentration for downstream analyses (86), for instance, visualization of the PCR products on agarose gel electrophoresis and Sanger sequencing. Several PCR variants for the detection of the viral state (episomal/integrated), including Amplification of Papillomavirus Oncogene Transcripts (APOT), Restriction Site Polymerase Chain Reaction (RS-PCR), Detection of Integrated Papillomavirus Sequences (DIPS), and Real-time PCR have been practiced. The methods are mainly based on the amplification of multiple early genes; E2, E6, and E7, and the measurement of the ratio of these genes (E2/E6, E2/E7) (42, 47). The PCR methods are simple and easy to use, and determines if integrations are present or not, but are unable to determine the integration site. The PCR methods can also provide false-negative results in samples with a low copy number of viral genomes (11, 87). Nonetheless, PCR is a useful method to obtain an exponential amount of a specific genome region that is of interest for further investigations. The APOT method provides information about the integration sites while also identifying the HPV transcript structure. However, the method relies on high RNA quality (88).

### 1.8.2 First generation sequencing

In 1977, Frederick Sanger and his colleagues developed the first-generation DNA sequencing method based on chain termination using dideoxynucleotides (ddNTPs), named Sanger sequencing (89, 90). The method reveals the nucleic acid order. As Fred stated, “*a knowledge of sequences could contribute much to our understanding of living matter*” (90). A similar

understanding can be obtained by sequencing genome regions containing HPV integrations. NGS data may generate false-positive results, therefore, Sanger sequencing has been referred to as the gold standard for validating NGS data (91). However, validation by Sanger sequencing may be relatively costly, time-consuming, and laborious (91, 92). A viral integration site can be uncovered when aligning the obtained sequence to the human and HPV genome (42).

### 1.8.3 Next-generation sequencing (NGS)

When NGS was first developed, several studies used the technology to determine the HPV integration breakpoints through either whole genome or exome sequencing (42). Whole-genome sequencing encompass sequencing of both non-coding (introns) and coding (exons) regions, while only exons are sequenced in exome-sequencing. Illumina is a second-generation sequencing technology that has made a successful contribution to the sequencing field (90, 93).

NGS technology was beneficial to the HPV integration research enabling several studies to determine integration sites. A study conducted by *Liu et.al* used HPV capture technology combined with NGS to investigate HPV integration sites in 166 women (5). The study reported several HPV integration sites primarily located in the E1 and E2 regions in samples from normal cervical epithelium and different CIN stages. The study also reported a higher integration percentage in CIN3 tissue samples in comparison to the normal epithelium, CIN1, and CIN2 (5). The disadvantage with the NGS technology is the massive amount of sequencing data generated, such as the number of reads in each sequencing run (42), which requires bioinformatics expertise to process (94). The methods may also be time-consuming, error-prone, and expensive making them unsuitable for a routine laboratory. This is especially problematic in developing countries lacking common biomedical detection instruments and methods.

## 1.9 Cervical cancer prevention

### 1.9.1 HPV vaccination

Vaccination is essential as a primary action to prevent HPV infection and transmission. A prophylactic vaccine against HPV is frequently offered in many developed countries and less in developing countries as a consequence of financial barriers. The vaccine is usually offered to females and males in the age group 9-14. This has been recommended by WHO as an action prior to sexual debut (27, 95). In Norway, the HPV vaccine has been offered to girls in 7<sup>th</sup> grade since 2009 as part of the Childhood Vaccination Programme (19, 49). Males were enrolled in

the programme in 2018 (19). The Immunization program is beneficial as primary prevention for individuals not infected with HPV and partly for people already infected with HPV (96).

The vaccine is made of virus-like particles (VLPs) from capsid proteins L1 or L2 (2). Cervarix (GlaxoSmithKline) offers a bivalent HPV vaccine (2vHPV) against HPV16 and 18; Gardasil (Merck, Kenilworth, New Jersey, USA) offers a quadrivalent (4vHPV) against type 6, 11, 16, and 18, and which was later expanded to protect against five additional HPV types; 31, 33, 45, 52 and 58 (97). This nonavalent (9vHPV) vaccine showed an additional increase in the prevention of infection and disease against nine HPV types used in the vaccine (98). Cross protection for non-vaccine HPV types in the 2vHPV and 4vHPV vaccine has been reported (27, 99, 100). A study conducted by *Malagón et. al* demonstrated that 4vHPV and 2vHPV were efficient against HPV, 31, 33, and 45 (100). This is a result of similarities between the epitopes (L1) of the vaccine-targeted and non-targeted HPV types (100, 101). However, Cervarix exhibited better cross-protection efficacy than Gardasil (100, 102, 103). Although cross-protection in the 2vHPV and 4vHPV have been reported against other phylogenetically close HPVs, the antibodies might not be sufficiently specific to prevent a possible infection by other non-targeted HR-HPVs. Because of this, the prevalence of HR-HPVs other than 16 and 18 may arise in the future (47). Therefore, research on other HR-HPVs and their molecular mechanisms may also contribute to uncovering genomic events important for vaccination research.

### 1.9.2 Cervical cancer screening

Screening is a secondary preventative action. A screening algorithm that secures broad coverage and follow-up of women with cellular abnormalities is important for reducing the cervical cancer incidence (31). Diagnostic screening is like vaccination, mainly available in developed countries. This is because of a lack of public health policy, education, media attention, clinical settings, financial support, and poor capacity for identification and follow-up treatment in developing countries (99). On a national level, Norway has several components involved in screening, follow-up and guidelines for vaccination and HPV surveillance. The Cancer Registry is an institution responsible for cancer statistics, screening and research (104). The Cancer Registry has an important role in preventing cervical cancer in all age groups. The National Institute of Public Health in Norway (FHI: Folkehelseinstituttet) is responsible for the distribution and follow-up of the vaccination program (105). Communicable Disease Notification System (MSIS: Meldingssystem for smittsomme sykdommer) is responsible for

surveillance of vaccine effectiveness in close collaboration with the National HPV Reference Laboratory at Akershus University Hospital (Ahus) (106).

Cervical cancer screening involves cytology and HPV testing of cell samples taken from the cervix. These are either offered as primary detection individually or in combination. The cytological test is based on cell collection with a brush or spatula by scraping material from the squamocolumnar junction (36). Afterwards, the material is smeared onto a glass slide, fixated, and stained with the Papanicolaou (Pap) procedure (2, 36, 107). Another way to prepare the sample is through liquid-based cytology (LBC) where the cells are collected in a suspension and applied onto a glass. Liquid-based cytology improves the quality relative to the Pap-smear because the cells are protected in a liquid suspension (36, 108).

The Norwegian Ministry of Health and Care Services has recommended HPV primary screening for women between 34-69 years of age and primary cytology screening from 25 years of age (19). Cytology screening involves visual inspection of cells by experts to identify abnormalities and due to the subjective interpretation training is continuous (99). HPV testing has higher sensitivity than cytology in identifying people at risk for developing cervical cancer. However, HPV is a common infection and hence HPV testing has lower clinical specificity (31, 51, 109), being an important argument for not introducing primary HPV screening in lower age groups (25-33) where the virus is more prevalent. If the HPV test is negative, a woman is recommended to have a new screening test in five years. In this way, the screening is more cost-effective and less burdensome for the individuals being screened (110).

Although HPV infection is considered to be common, only in small percentage of infected women the infection might progress to precancerous lesions and cancer (39). Therefore, a combination of HPV-testing and cytology is optimal in a screening process. HPV primary screening was introduced in Norway in 2015 through a pilot project in four counties (19, 111, 112). HPV primary screening will be offered to all women at the age of 34-69 from 2022. Younger women, 25-33 years of age, will still be offered cytology as the primary screening tool. The screening algorithm is outlined in a flowchart that shows an overview of how and when the follow-up is organized for early detection of cell abnormalities. For example, if a woman is HPV-negative, she will be reminded to take a new HPV test after five years (19).

Biomarkers could be helpful to guide HPV primary screening in future cervical cancer screening. Among the potential biomarkers are HPV genotyping, methylation, and the detection of HPV integration (109). Viral integration may be an early event that can occur before the morphological changes and hence potentially a biomarker for predicting the development of lesions to cervical cancer (11). The anomalies that may occur are differently classified.

#### *1.9.2.1 Classification of dysplasia and neoplasia*

Different classification systems have been developed for a common international terminology to describe cytological and histological abnormalities. Cytology reporting applies the Bethesda system while histology uses the CIN system. The Bethesda system classifies squamous cell abnormalities into four categories, Atypical squamous cells (ASC), low-grade squamous intraepithelial lesions (LSIL), high-grade squamous intraepithelial lesions (HSIL), and squamous cell carcinoma. ASC was introduced as a result of uncertainties associated with cytological evaluation. The category contains the two subcategories atypical squamous cells of undetermined significance (ASC-US) and atypical squamous cells, which cannot exclude HSIL (ASC-H) (2). The CIN system is based on tissue architecture classified into mild dysplasia (CIN1), moderate dysplasia (CIN2), severe dysplasia (CIN3), and carcinoma (113), adenocarcinoma (AC), and adenocarcinoma in situ (ACIS) (114). Proper identification of the abnormalities helps the clinicians to decide treatment options.

### 1.10 Treatment of cervical lesions

Most HPV-induced cell modifications disappear spontaneously with the help of the host immune system (2). If serious abnormalities have occurred, clinicians will consider the patients' health, stage of invasion, and eventually how the tumor has processed when deciding the appropriate treatment options (32). In Norway, women diagnosed with CIN2 or more severe lesions are recommended to have the abnormal cells removed (115). Treatment focuses on the removal of precancerous cells while minimizing harm to the cervix (2). Conization or cold knife cone is a surgical procedure frequently used. Conization can be performed with a scalpel, laser, or an electrosurgical instrument, called Loop Electrosurgical Excision Procedure (LEEP) (116). LEEP is a common procedure because it can be performed under local anesthesia and produces a tissue specimen suited for clinical evaluation (39). Another treatment option is cryotherapy, a local destruction method of cervical tissue through freezing (117). More advanced tumors may need chemoradiotherapy.

## 2. Aims of study

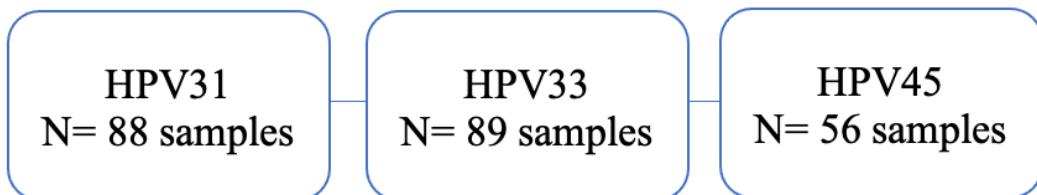
The major challenge of the HPV screening is the high number of women infected with the virus compared to the low number developing cervical cancer. Currently, there is no ideal biomarker for predicting cervical cancer progression leading to unnecessary follow-up and treatment of women with minimal risk for developing high-grade lesions or cancer. HPV integrations have been reported as a potential biomarker for predicting the development of lesions to cervical cancer (11). Much is still unknown about the integration process and most studies have been focused on HPV16 and 18 because of their high prevalence in cervical cancer cases. Less attention has been given to other HR-HPVs such as 31, 33, and 45.

This master project aimed to validate and characterize NGS-reported integration sites in women with HPV31, 33 or 45 infections, and a diagnostic category of LSIL/ASCUS, CIN2, CIN3, or cervical cancer. 88 HPV31, 89 HPV33, and 56 HPV45 samples have been sequenced previously by the HPVseq research group with Illumina NGS sequencing technology. Although the NGS may reveal genomic information about the HPV integrations, the method also may produce false-positive results. NGS data needed to be manually assessed to determine whether the reported HPV integrations were likely artefactual or could be confirmed by Sanger sequencing. Finally, the study also aimed to uncover potential integrations in hot-spot regions and to identify microhomology sequences at the integration breakpoint.

### 3. Materials and methods

#### 3.1 Study population and specimen collection

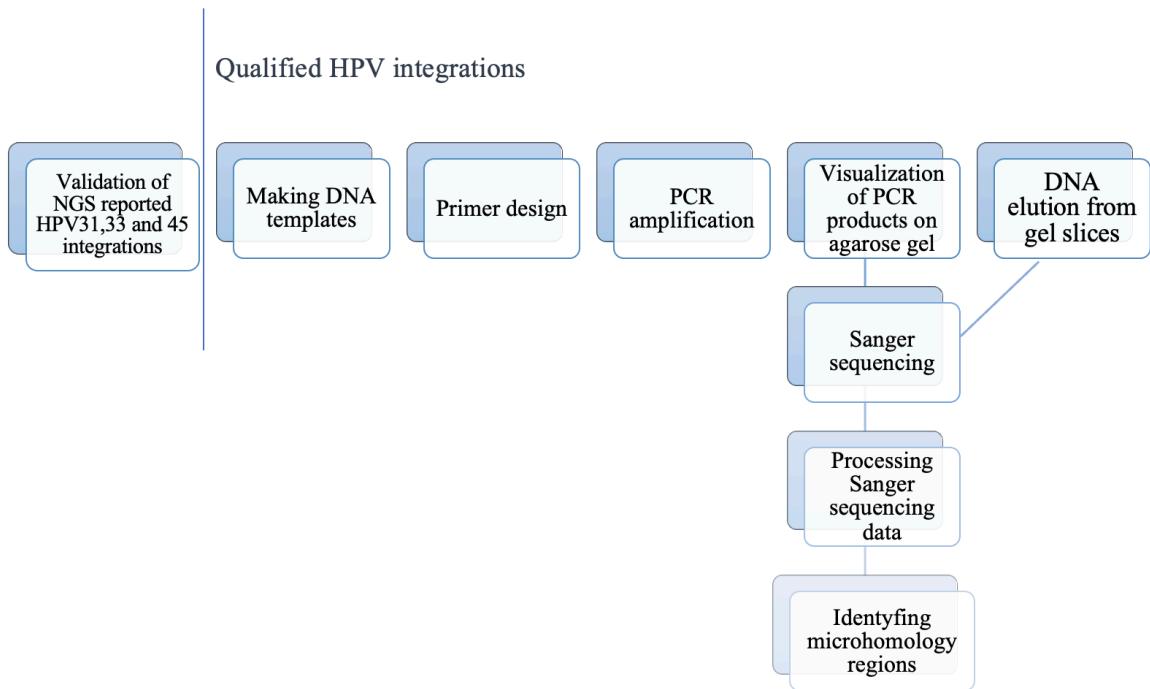
LBC samples used in the study were obtained from women who had tested positive for HPV31, 33, or 45. These were obtained from a biobank at the National HPV Reference Laboratory at Ahus. The biobank contains samples from patients attending the cervical cancer screening program between 2005-2009. In total, 88 HPV31 positive samples, 89 HPV33 positive samples, and 56 HPV45 samples were included as illustrated in Figure 10. The women who attended the screening program had a clinical history of normal cytology, LSIL/ASCUS, CIN2, CIN3, or cancer. The diagnostic category of the patients was unknown during the validation to hinder quick conclusions and sample prioritization based on the knowledge of HPV integrations. All women have given their written consent and the sample material has been pseudonymized during the work. The research has been approved by the regional committee for medical and health research ethics, Oslo, Norway (REK) [REK-reference 2017/447] (Appendix 1) and the Data Protection Office at Ahus (Appendix 2). All the experiments were performed in accordance with the committee's guidelines and regulations.



**Figure 10: Number of samples used in the study.** In total, 88 Human papillomavirus (HPV) 31-positive samples, 89 HPV33, and 56 HPV45- positive samples were obtained from women with a clinical history of normal cytology, LSIL/ASCUS, CIN2, CIN3, or cancer.

#### 3.2 Validation of HPV integration sites

HPV31, 33, and 45 positive LBC samples were previously whole genome sequenced using the Tagmentation-assisted multiplex PCR enrichment sequencing (TaME-seq) protocol (10), which employs Illumina NGS (125 bp paired-end sequencing). The validation of the NGS reported HPV integration sites was performed according to the research groups' earlier work and recommendations. Firstly, the NGS-reported integrations were validated, and the potentially true integrations qualified for further confirming analyses. An overview of the study workflow is illustrated in Figure 11.



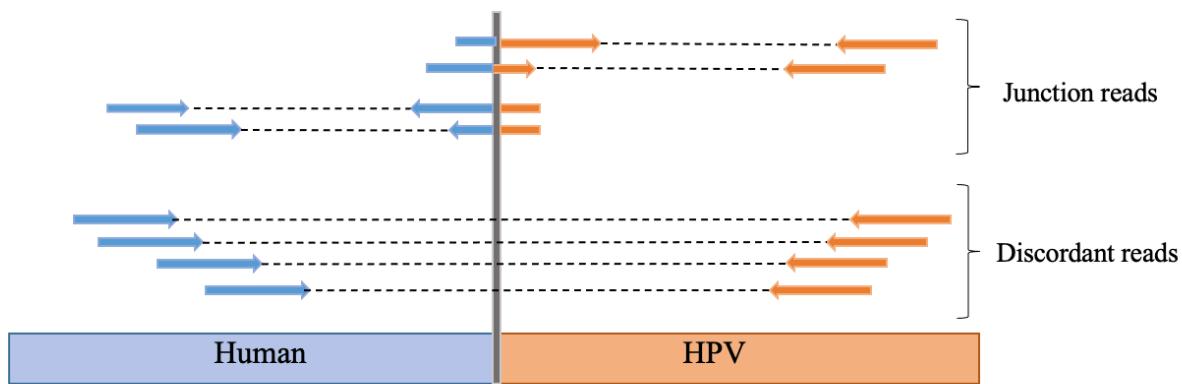
**Figure 11: Various biomolecular methods performed in the study.** Reported Human papillomavirus (HPV) integrations from next generation sequencing (NGS) Illumina sequencing were manually investigated to determine whether the read alignments were likely artefacts or could potentially be true. The potentially true HPV integrations were qualified for further analyses, involving template and primer design, polymerase chain reaction (PCR) amplification, visualization on agarose gel, Sanger sequencing, and identifying hot-spot and microhomology regions at the integration breakpoint.

### 3.2.1 Read alignment and visualization on Integrative Genomics Viewer (IGV) software

In the process of validating HPV integrations, the Illumina reads were mapped to a reference file consisting of the human reference genome (GRCh8/hg38) and 183 HPV types including types 31, 33, and 45 obtained from the Papillomavirus Episteme (PaVE) database (44, 45). Hierarchical Indexing for Spliced Alignment of Transcripts (HISAT2) and Local Alignment Search Tool (LAST) were used for read mapping. Reads from all samples were first mapped with HISAT2, while unmapped reads were further remapped with LAST algorithm to determine the exact position of HPV-human integration breakpoints (vertical black line separating human/HPV genomes). Remapping with LAST algorithm was performed since HISAT2 cannot map junction reads.

Pair-end reads were identified as discordant when one read of the read pair mapped to the HPV genome and the other to the human chromosome (HISAT2 alignment), thereby indicating a potentially true HPV integration in the region. When one individual read mapped to both the human and HPV genome and the other read either to the human or HPV genome it was

identified as a junction read (LAST alignment). The two alignment algorithms are presented in Figure 12. Binary Alignment (BAM) format was used to store the information of reads mapping to the human reference genome and genomes of 183 HPV types. The BAM files were loaded into IGV v2.8.9 software and the location of the reported integration breakpoint was used to validate the possible HPV integration. Reported integration breakpoints having  $\geq 2$  discordant read pairs or  $\geq 3$  junction reads were manually inspected to investigate whether they were the result of potentially true integration or technical artefacts.



**Figure 12: Junction and discordant reads.** An overview of junction and discordant reads aligning to the human and Human papillomavirus (HPV) genomes. Junction reads represent individual reads mapping to both the human and HPV genome (arrows colored with blue and orange) and the other read mapping to either the human (blue arrows) or HPV genome (orange arrows), whereas discordant reads represent reads of a read pair where one read maps to the human genome (blue arrows) and the other read to the HPV genome (orange arrows). Illustration obtained and reconstructed with permission from Wang et al. (1).

The reported integration breakpoints may represent more than one HPV integration per sample, however, located at different regions in genomes. If  $>1$  integration breakpoint was reported in the same human and HPV genome with the reads aligning with a gap of a few bp, only one of the integrations was used for further analyses.

### 3.2.2 Categorizing potential HPV integrations or potential artefacts?

When validating the NGS reads for the reported integration breakpoints some factors were considered as potential errors, 1) reads sharing the same start- and stop coordinates, 2) reads of same fragment/duplicates, and/or 3) reads mapping to more than one region in the genome. The main factor when considering integrations as not potentially true was the first factor 1) sharing the same start and stop coordinates.

Several examples have been collected to demonstrate the criteria included in the validation process (Figure 13-17). The reported HPV-human integrations were categorized into three

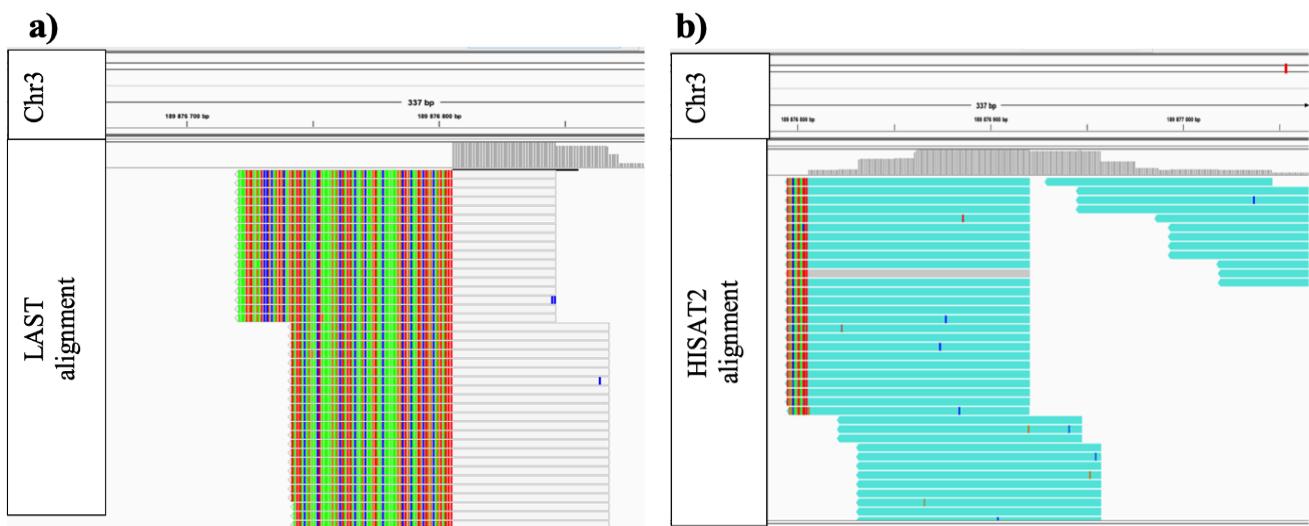
groups: “Yes” (potential integration), “No” (not a potential integration), and “Maybe” (possibly a potential integration). Further analyses were performed on the reported integrations classified into “Yes”, and “Maybe” categories.

### 3.2.2.1 Criteria for categorizing the reported integrations “Yes” (potential integration)

#### 3.2.2.1.1 Different start- and stop coordinates

Different start- and stop coordinates of the reads was a criterion for categorizing the reported integration “Yes” as they indicated amplification reactions from different fragments. Figure 13 shows junction and discordant reads visualized in IGV. Figure 13a shows mapped junction reads where the white thick line of the read maps the human chromosome 3 and the multi-colored part, mismatched bases. When mismatches sequences were separately analyzed by Nucleotide Basic Local Alignment Search Tool (BLASTn) (BLAST) the mismatched sequence was homologous to HPV, mainly HPV31, 33, and 45. The BLAST system reports the query coverage, percent identity, and e-value (118, 119)

Figure 13b shows discordant reads where the thick blue line represents one read pair mapping to human chromosome 3 (read-pair mapping to HPV not shown). Figure 13 shows examples of reads having different start and stop- coordinates associated with reads originating from different fragments, consequently, considered as a potentially true HPV integration categorized as “Yes”.

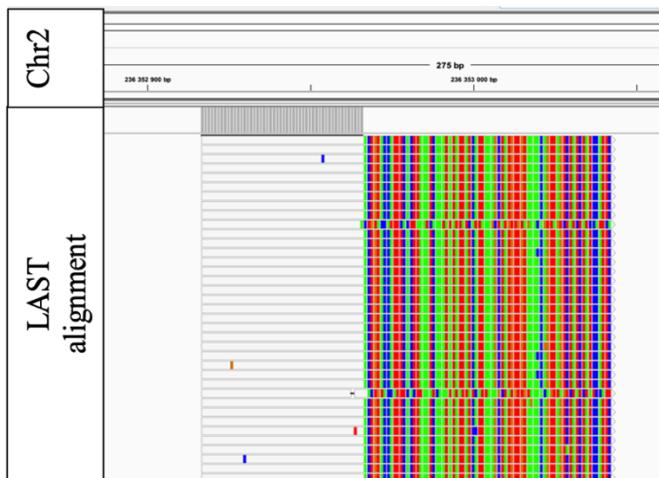


**Figure 13: Junction and discordant reads.** Screenshot from Integrative Genomics Viewer (IGV) v 2.8.9. Junction reads are shown in Figure 13a where the right-hand side of the read is mapping to the human chromosome 3 and the multi-colored left side mismatched bases. The discordant reads are shown in Figure 13b that represent mapping to the human chromosome 3 (read pair mapping Human papillomavirus (HPV) not shown). The read pairs had different start- and stop coordinates. Consequently, considered a potentially true integration and categorized as “Yes”.

### 3.2.2.3 Criteria for categorizing the reported integration “No” (not a potential integration)

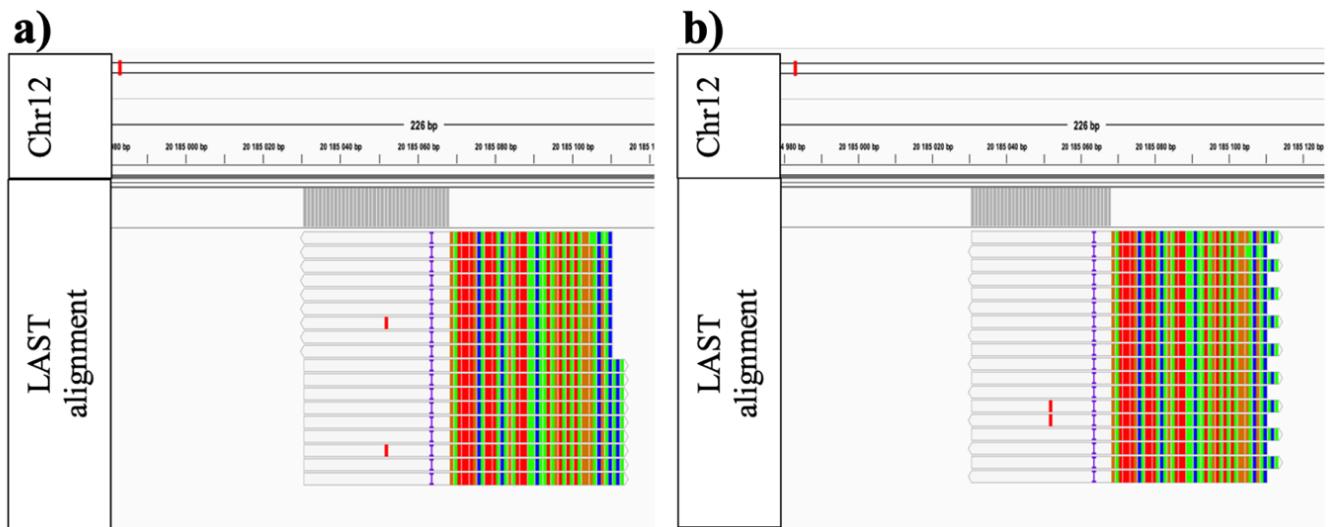
#### 3.2.2.3.1 Same start-and stop coordinates

The same start- and stop coordinates of the reads was a criterion for categorizing the reported integrations “No”. Figure 14 shows junction reads mapping to the human chromosome 2. The reads had identical start- and stop coordinates mainly associated with PCR duplicates from the same target on the template. This can occur during the library preparation step when DNA is fragmented randomly, and PCR amplified to expand the number of copies to increase the library quality for optimal sequencing. This is a common technical issue in NGS technology and may in some cases lead to false-positive results (120). When reads had the same start- and stop coordinates they were not considered a potentially true HPV integration categorized as “No”.



**Figure 14: Junction reads.** Screenshot from Integrative Genomics Viewer (IGV 2.8.9) shows junction reads sharing the same start and stop coordinates associated with polymerase chain reaction (PCR) duplicates thereby considered not a potential integration and categorized as “No”.

If the reads exhibited as the same start coordinates but different stop coordinates (Figure 15a), the reads were sorted in IGV (Figure 15b). This was performed to discover whether the forward (F) and reverse (R)-reaction belonged to the same fragment (XXXXXX/1 and XXXXXX/2), consequently, considered as a technical artefact and categorized as “No”.



**Figure 15: Junction reads.** Screenshots from Integrative Genomics Viewer (IGV) v2.8.9 shows junction reads (15a and 15b). It exhibited as the reads had same start coordinates and different stop coordinates (15a), but when sorting the reads in IGV the forward and reverse reactions belonged to the same fragment (XXXXXX/1 and XXXXXX/2) (15b) thereby considered not a potentially true integration and categorized as “No”.

### 3.2.2.3.3 Reads mapping to more than one region

Reads mapping to more than one region was a criterion for categorizing the reported integrations “No”. Figure 16 represents discordant reads containing the homopolymer of nucleotide Thymine (T) mapping to the human chromosome 12 (read-pair mapping HPV is not shown). The reads were also mapping to several other regions presented as thin blue lines on both sides of the read. Homopolymeric nucleotides, such as long polyT sequence was considered a result of insufficient priming in the PCR reaction. If the read mapped to other regions in the genome than the region used for designing it, it was not considered a potentially true integration and categorized as “No”.



**Figure 16: Discordant reads.** Screenshot from Integrative Genomics Viewer (IGV) v2.8.9 shows discordant reads. The reads consist of the nucleotide Thymine (T) in homopolymeric tracts and consequently mapping more than one region in the genome seen as thin lines on both sides of the read. This is considered not a potential integration and categorized as “No”.

### 3.2.2.4 Criteria for categorizing the reported integration “Maybe” (possibly a potential integration)

#### 3.2.2.4.1 Relatively same start- and stop coordinates

Reads that had the same start but not the same stop coordinates and did not belong to the same sequencing reaction F or R (XXXXXX/1 and XXXXXX/2) were categorized as “Maybe”. An example of junction reads sharing the same start but not the same stop coordinates are presented in Figure 17. The different stop coordinates were a result of the trimming process generating <125 bp length reads. Raw reads were trimmed to exclude reported low-quality nucleotides. These were difficult to classify into the “No” category because they could have different start and stop coordinates if they were of 125 bp length, representing amplification reactions from different fragments.



**Figure 17: Junction reads.** Screenshot from Integrative Genomics Viewer (IGV) 2.8.9 shows junction reads with the same start but different stop coordinates. This is a potentially true integration thereby categorized as “Maybe”.

### 3.3. *In silico* DNA template for primer design

An *in silico* DNA template was generated for reported integrations categorized as “Yes” and “Maybe” to enable primer design on the F and R strands essential to amplify a region containing a potentially true HPV integration. The generated templates were composed of both human and HPV-specific sequences in the orientation identified at the integration breakpoint. Templates generated from junction reads could be identified with specific product sizes, whereas the product size of discordant reads was unknown as the distance between the read pairs was not

identified. Junction and discordant reads required different procedures for designing the template as the reads look different at the integration breakpoint.

### 3.3.1 Discordant reads

DNA templates from discordant reads were generated based on the read information from the Sequence Alignment Mapping (SAM) flags. The SAM flags reports various number codes determining the read orientation at the integration breakpoint (Table 2) (121). It is important to identify the read orientation to know which strand to use for primer design. If the human or HPV read aligned to the minus strand determined from the SAM flags, the reads were reverse complemented. Subsequently, the term [NNN] was added between the human and HPV sequences to separate the genomes, human sequence placed on the 5' side and HPV on the 3' side of the term. This is a standard format in Primer3 and Primer3plus for primer suggestions on both sides of the bracket, giving a human-specific Forward (F) primer, and HPV-specific Reverse (R) primer (122).

**Table 2: Sequence Alignment Mapping (SAM)**

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

<sup>1</sup>Shows an overview of the read and genome orientation at the integration breakpoint. The SAM flags were used to determine the read orientation of discordant read pairs through number codes that represent different events. The table is inspired by earlier HPVseq group member, Sonja Lagström. Abbreviations: DNA=Deoxyribonucleic acid, HPV=Human papillomavirus

### 3.3.2 Junction reads

DNA templates from junction reads were generated based on the read alignment at the integration breakpoint. The human and HPV orientation at the integration breakpoint was determined by BLASTn. If the human alignment was on the right-hand side of the breakpoint, the whole sequence was reverse complemented. This was performed to organize the genomes in such way to ensure that the orientation of the human sequence was on the left side of the term [NNN] and the HPV sequence on the right side. Subsequently, 100-200 bp was added to increase the template size. This was done to avoid primer suggestions close to the integration breakpoint [NNN].

### 3.4 Primer design

F- and R primers were designed for every integration to flank the integration breakpoint. This is important for proper amplification if a potentially true HPV integration is present. The primer pairs were made in the Primer3 and Primer3plus websites that suggested primer pairs according to the term [NNN], separating the genomes. The F-primer was human-specific and the R-primer HPV-specific. In Primer3, the following settings were chosen for designing optimal primers to ensure a proper PCR reaction: 1) primer length ranging 18-24 bp, 2) balanced distribution of the nucleotides Adenine (A), T, Cytosine (C), and Guanine (G), 3) primers not containing homopolymeric nucleotides (for example AAAA) or heteropolymeric regions (for example CACACA), 4) F and R primer not complementary to each other to avoid primer dimers, 5) primers not being complementary to itself to avoid secondary structures and 6) primer pairs having melting temperature around 60 °C (123). Primer pairs close to the integration breakpoint, approximately 10-20 bp, were also avoided to hinder complications when performing Sanger sequencing. When Primer3 did not suggest optimal primer pairs, the Primer3plus was used (124).

F and R primers were investigated as a control step for specificity to prevent off-targeted hybridization to other genomic regions. This was done by use of BLASTn or Blast-like alignment tool (BLAT). Subsequently, the primer pairs were validated for “PCR products” (Figure 18a) and performed for a “Primer map”(Figure 18b) at Sequence Manipulation Suite, *Bioinformatics* website (125-127). “PCR products” was done to control the product length to ensure it matched the template sequence, while the “Primer map” was performed to visualize where and on which strand the primers bind. The marked pink sequence represents the F primer, while the orange represents the R primer.

a)

PCR Products results

>152 bp product from linear template sample sequence A, base 57 to base 208 (T7 - T3).  
ACAATTGATTCACAGCAGCTGTAAAGTCTATCTCGATAAAAGCTATGATCATGAAG  
CTAAACCGCAGNNNGCTTGTAAATAGCTTTAGTTGTAAATACTGCACTGCGGATCTAC  
ATTTCTGCATTGTCACTACTATCCCCACAC

b)

Primer Map results

Results for linear 228 residue sequence "sample sequence" starting "AAGCAAATCA"

>>>forward>>> 57 to 78  
1 A N Q L K S H F F V I \* S L L V S I T I \* F T A A L \* S L S S I K P M  
1 S K S T K I T F L R N I K P T C I Y N N L I H S S S V K S I F D K A Y  
1 K Q I N \* N H I S S \* Y K A Y L Y L \* Q F D S Q Q L C K V Y L R \* S L  
1 AAGCAAATCAACTAAATCACATTCTCGTAAATATAAAAGCTACTGTATCTATAACAAATTGATTCAACAGCAGCTCTGTAAAGCTATCTCGATAAAGCTA  
1 10 20 30 40 50 60 70 80 90 100  
1 TTGTTTGTGATTGATTTGTAAAGAAGCATTATATTCGGATGAACTAGATATTGTTAAACTAAGTGTGTCGAGACATTTCAGATAGAGACATTTCGGAT  
<<<reverse<<< 188 to 208  
36 I M K V N A X X L \* \* L F \* F C N T A L R I Y I F C I V T T I P T T T  
36 D H E G K R X X L V I A L L V L \* Y C T A D L H F L H C H Y Y P H H Y  
36 \* S \* R \* T R X A C N S S F S S V I L H C G S T F S A L S L L S P P L  
106 TGATCATGAAAGTAAACCGCAGNNNGCTTGTAAATAGCTCTTTAGTTCTGTAAATCTGACTGCGGATCTACATTTCGCAATTGTCACACTPATCCCCACACTA  
106 110 120 130 140 150 160 170 180 190 200  
106 ACTAGTACTCCATTGCGCTNNNGAACATTATCGAGAAAATCAAGACATTATGACGCTGACGCCAGATGATGAAAAGACGTACAGTGATGATAGGGTGGTGAT  
71 L C T M Y  
71 F V Y Y V  
71 L C V L C T  
211 CTTTGTGTACTATGTACA  
211 220  
211 GAAACACATGATAATGAT

Primer:	Sequence:
forward	5'-ACAATTGATTCACAGCAGCTC-3'
reverse	5'-GTGGTGGGGATAGTAGTGACA-3'

**Figure 18: In silico control step.** Screenshot from Sequence Manipulation Suite, Bioinformatics used to avoid off-target hybridization. Here seen the determination of the product size (18a) and visualization of the primer pairs (18b). Screenshots obtained from Bioinformatics, Stothard et al. (125-127).

### 3.5 Sample preparation and DNA extraction

The sample DNA was extracted using the automated magnetic beads technique, Nuclisens® easyMAG® (Biomérieux, USA) (128). In each specific well, 100 µL patient samples and 1000 µL lysis buffer were mixed into a homogeneous solution as recommended by the manufacturer (Biomérieux™, USA). The lysis buffer contains guanidine thiocyanate that disrupts viral particles or cells to release DNA/RNA in the sample while simultaneously inactivates RNases and DNases (129). Subsequently, 50 µL of a solution containing magnetic beads that binds sample DNA by presence of chaotropic substances was added to each well (130, 131). The eluate was stored at -80 °C to retain the DNA quality and integrity (132).

### 3.6 Measurement of DNA concentration

The DNA concentration was measured on Qubit® 3.0 Fluorometer (Life Technologies, USA) prior to the PCR reaction to ensure optimal DNA quantity in every sample (133). Qubit dsDNA

HS assay Kit (Invitrogen, Burlington, Ontario) was used to prepare the samples as recommended by the manufacturer (134). The instrument detected the fluorescent intensity in the samples converting it to DNA concentration (ng/µL) (133). When necessary, samples with the higher DNA concentration were diluted to ensure the optimal concentration range for PCR reactions (5-20 ng/µL).

### 3.7 Amplification by Polymerase Chain Reaction (PCR)

PCR was performed to amplify the DNA sequences (templates) consisting of potential human and HPV-specific regions. Amplification is essential to detect the presence of DNA products and for later analyses such as visualization on agarose gel and Sanger sequencing. The PCR reaction contained sample DNA (5-20 ng/µL), designed F-primer (human-specific) and R-primer (HPV specific) and 2 × Phusion™ Master Mix (Thermo Scientific, USA) (135), containing a Phusion polymerase (Table 3). The Phusion polymerase can generate long templates with high accuracy and speed (135), favorable for discordant reads where the specific product size was unknown. Additionally, the Phusion has a lower error rate, 1.32 % compared to other polymerases (136). Phusion polymerase is also tolerant to various inhibitors, allowing a robust amplification of the PCR products that require minimal optimizations (137).

DNA extracted from the CaSki cells was used as positive control. The positive CaSki control also had a human-specific F primer and HPV-specific R-primer with a known integration reported in the human chromosome X (138).

*Table 3: Volumes of each reagent used in the Polymerase chain reaction (PCR) setup.*

Reagent	Volume for each tube (µL)
2 × Phusion Mastermix (MM)	10
F-primer	1
R-primer	1
DNA (5-20 ng/µL)	(2)
H <sub>2</sub> O	(7)
Total volume	21

Abbreviations: DNA= Deoxyribonucleic acid, F= Forward, R=Reverse

After adding the reagents, the samples were ready for the reaction on Gene Amp® PCR system 2700 (Applied Biosystems, USA) with the following program shown in Table 4.

Table 4: Polymerase chain reaction (PCR) program.

Cycles	Temperature	Reaction step	Time
30	98 °C	Denaturation	30s
	98 °C	Primer annealing	10s
	60 °C		30s
	72 °C		15s
	72 °C	Elongation	10min
	10 °C		∞

Abbreviations: min = minutes, s = seconds

Samples not giving a successful directly Sanger sequencing result with smears and/or unspecific bindings present when visualizing PCR products on agarose gel (methods section 3.8.1 *Agarose gel electrophoresis*) were used to perform a Touch Down (TD-PCR). TD- PCR is a PCR variant where the annealing temperature is increased to avoid off-target priming (139). The annealing temperature was settled at 66 °C on Eppendorf® Mastercycler® (Eppendorf, USA) and 6 extra cycles were performed where the temperature was decreased by 1 °C for every cycle.

When performing PCR prior to the gel extraction method (methods section 3.9 *Purifying DNA fragments from Gel*) multiple PCR reaction parallels per sample were used to increase the input DNA for Sanger sequencing.

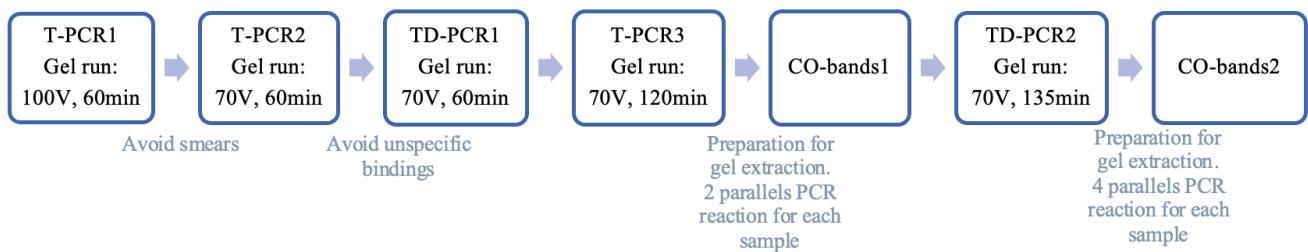
### 3.8 Analysis of PCR product

#### 3.8.1 Agarose gel electrophoresis

Agarose gel electrophoresis was performed for a qualitative and semi-quantitative visualization and analysis of the PCR products (140, 141). Agarose gel (2%, total volume 300 mL) was prepared by dissolving Ultrapure Agarose (Invitrogen, USA) in 1 × Tris-acetate-EDTA (Ethylenediaminetetraacetic acid) (TAE)- buffer. After the gel had cooled, 24 µL of a dsDNA-binding fluorescent dye, Gel Green™ (Biotium, USA) was added (142). A 25-766 bp (~800 bp) molecular weight standard (ML), Quick- Load® (New England, BioLabs, USA) with known fragment lengths was included in the setup allowing proper identification of the fragment sizes (140). Prior to the “Cut out Bands first time” (CO-bands1) and “Cut out Bands second time” (CO-bands2) gel extractions, a higher amount of the PCR products was loaded into each gel-well to increase the DNA amount. Subsequently, the gel was run on Power PAC basis

electrophoresis (BioRad, USA) at 100 Volt (V) for 60 minutes (min) (T-PCR1). When applying an electrical field, the negatively charged DNA will move towards the positively charged anode. The PCR products were visualized under UV-light by use of a Molecular Imager Gel Doc™ XR+ Imaging system (BioRad, USA) and results interpret at Image Lab 6.1 (Biorad, USA) software.

Samples with unsuccessful Sanger sequencing results or un-specific primer binding appearing as smears on agarose gels were used for further method adjustments. The adjustments were performed in the PCR reactions and the voltage- and running time conditions in the gel run. In addition to performing gel extractions (Figure 19).



**Figure 19: Various Polymerase chain reaction (PCR) and agarose gel adjustments.** Shows the various PCR reactions and, voltage- and time conditions used for visualization on 2% agarose gel, and the gel extraction processes.

T-PCR1= Traditional PCR first run, followed by a 2% agarose gel run at 100 V for 60 minutes

T-PCR2= Traditional PCR second run, followed by a 2% agarose gel run at 70 V for 60 minutes

T-PCR3= Traditional PCR third run, followed by a 2% agarose gel run at 70 V for 120 minutes

TD-PCR1 = Touch Down PCR first run, followed by a 2% gel run at 70 V for 60 minutes

TD-PCR2 = Touch Down PCR second run, followed by a 2% agarose gel run at 70 V for 135 minutes

CO-bands1= Cut out bands first time.

CO-bands2= Cut out bands second time

Abbreviations: Min= Minutes, V= Voltage

### 3.9 Purifying DNA fragments from Gel

Samples which failed when directly sequenced but had visible PCR products when visualized on agarose gels were used for gel extraction. This was done to determine whether one of the gel-bands could potentially contain an HPV integration. Gel-bands containing >30 bp long DNA amplicons were cut from the gel, dissolved, DNA isolated, and purified with Wizard® SV Gel and PCR Clean-Up System Kit (Promega, USA) as recommended by the manufacturer (143). The bands were visualized and cut by UV-light on Gel Doc™ XR+ Imaging system (BioRad, USA) or by Blue light on Safe Imager™ 2.0 Blue Light Transilluminator (Invitrogen, USA). The gel extractions were performed twice; 1) Initially by eluting DNA in 50µL room

temperature nuclease-free water (CO-bands1) and 2) eluting DNA in 15 µL 37 °C nuclease-free water in a prolonged incubation step (CO-bands2). The process was repeated to increase the input DNA for Sanger sequencing the second time. The DNA eluates were stored at 4 °C or -20 °C (143).

### 3.10 Preparing sequencing PCR

A separate sequencing PCR reaction was performed to amplify enough product to improve the Sanger sequencing quality while simultaneously incorporating fluorochrome-labeled ddNTPs. The reagents were added to a final volume of 10 µL (Table 5).

*Table 5: Volume of each reagent used in the sequencing Polymerase Chain Reaction (PCR) setup.*

Reagent	Volume for each tube (µl)
dH <sub>2</sub> O	5.5
5 × seq-buffer	1.5
Big dye terminator v. 1.1	1.0
F primer/R primer	1.0
Template (diluted 1:10/undiluted)	1.0
Total	10.0

*Abbreviations: dH<sub>2</sub>O = distilled water (H<sub>2</sub>O), F = Forward, R = Reverse*

The sequencing PCR reaction was performed on Gene Amp® PCR system 2700 (Applied Biosystems, USA) with the program shown in Table 6.

*Table 6: Sequencing polymerase chain reaction (PCR) program.*

Cycles	Temperature	Reaction step	Time
25	96 °C	Denaturation	10s
	50 °C	Primer annealing	5s
	60 °C	Elongation	1min
	4 °C		∞

*Abbreviations: m = minutes, s = seconds*

### 3.11 Precipitation of PCR sequencing products

The PCR sequencing precipitation was performed to remove enzymes, nucleotides, primers, and buffer that may interfere with the Sanger sequencing (144). The PCR sequencing products were mixed with 90 µL 69% isopropanol and incubated for 15-30 min. PCR sequencing products mixed with isopropanol solution were initially centrifuged with Eppendorf® 5810R

(Eppendorf, USA) by 3000 × g at 21 °C for 30 min followed by upside down centrifugation at 700 × g and 21 °C for 1 minute to further eliminate isopropanol debris. This step prevented alcohol debris residuals from interfering with the Sanger sequencing. Finally, 12 µL HiDi™ Formamide was added to each of the sequencing PCR products and transferred to a MicroAmp® Optical 96-Well Reaction Plate (Applied Biosystems, USA), ready for Sanger sequencing. Highly Deionized Formamide (Hi-Di) is used to resuspend samples and keep the DNA denatured prior to capillary electrophoresis (145).

### 3.12 Sanger sequencing

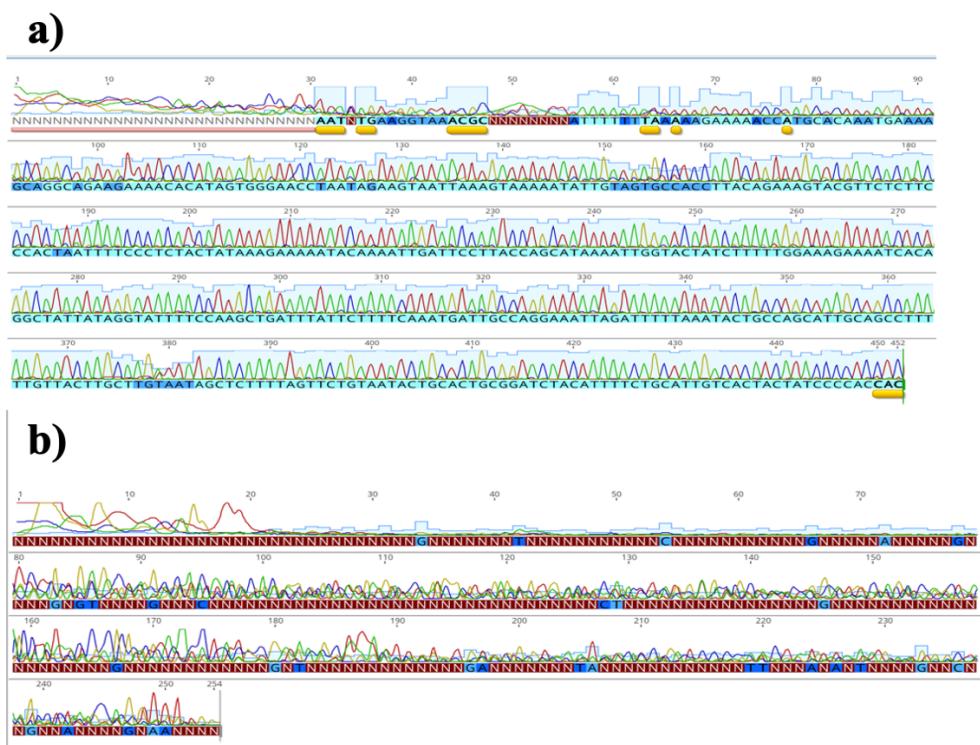
HPV integrations that were considered potentially true (“Yes” and “Maybe”) and qualified for further analyses were Sanger sequenced by dye terminator technique. This was used to uncover the nucleotide order to determine whether the sequence mapped to both the human and HPV genomes as a sign of a true integration. The qualified HPV integrations were Sanger sequenced with POP-7™ Polymer (Applied Biosystems, USA) by 3130 XL 16-capillary Genetic Analyzer (Applied Biosystems, USA). The DNA strand is terminated in each direction of the F and R reactions by incorporating fluorescently labeled ddNTPs (ddATP, ddTTP, ddGTP, and ddCTP). This results in DNA fragments of various lengths (146-149). Various-sized fragments migrated through the capillary electrophoresis. This is a device with a thin polymeric capillary through which fragments move depending on their size, short fragments move quicker through the capillary relative to large fragments (147). The capillary is not isolated at one point in the capillary, enabling the laser beam to excite the ddNTPs in the fragments. The fluorescent-labeled ddNTP emits and excites light at a specific wavelength corresponding to ddATP, ddGTP, ddCTP, or ddTTP determined by the detector. The results show a chromatogram with the nucleotides on the x-axis and light intensity on the y-axis. The ddNTPs are represented by the following colors; red ddATP, green ddTTP, blue ddCTP and yellow ddGTP. The peak far to the left in the chromatogram corresponds to the shortest fragment. The Sanger sequencing data was further processed.

### 3.13 Processing sequencing data

Sequencing data from both the F and R-sequencing reactions were loaded onto Geneious Prime v2020.2.2 and the chromatograms were investigated for peak appearance. Clear separated sequencing peaks with high quality (Figure 20a) provided readable and identifiable sequences, while sequences with no- or low-quality Sanger sequences were either not identifiable (Figure

20b) or only partly identifiable. If a sequence was of high quality but had some regions where the instrument had challenges to separate the nucleotides because of technical artefacts (150), the sequences were either trimmed or edited. Examples of edited regions are seen as yellow lines underneath the sequence in Figure 20a. If the F- and R-sequences were of high-quality the sequences were assembled to generate a continuous sequence (151). If a continuous sequence is defined as a set of overlapping sequences to make a connected sequence (151). If a continuous sequence was present the pairwise identity between the F and R- sequences were determined, green color above the chromatogram showed 100 % identity, green-brown color  $\geq 30\%$  identity but  $<100\%$ , and red color  $<30\%$ .

The homology of the Sanger sequences was identified by BLASTn or BLAT. If the continuous sequence or one of the F/R -sequence was homologous to the same human chromosome and HPV type as originally reported from the NGS data, the HPV integration was considered confirmed. Whenever the sequence mapped to both human and HPV genome each of the specific sequences were divided by the term [NNN]. The human sequence represented in blue and was placed on the left side, and the HPV sequence represented in orange and was placed on the right side of the term. If the human part of the sequence mapped to a known human gene, the percent identity from BLASTn was reported.



**Figure 20: Sanger sequencing chromatograms of low and high-quality.** Screenshot a) a high-quality sequence enabling identification of the sequence, whereas screenshot b) demonstrates low-quality sequence giving unreadable sequences. The screenshots were obtained from Geneious Prime v2020.2.2.

### 3.14 Determining microhomology regions

BLASTn and/or BLAT were used to identify short homologous sequences known as microhomology regions at the integration breakpoint in the confirmed HPV integrations. The search for microhomology regions was made either by a search with the continuous sequence or one of the (F/R) sequences. The position of the sequence homologous to each of the human and HPV genomes was identified by BLASTn and/or BLAT. If overlapping sequences were present between the human and HPV genome it was designated as a microhomology sequence. For instance, if a 150 bp sequence was used for search in BLASTn and BLAT and the human genome was homologous to 1-113 bp of the sequence and the HPV genome to the 110-150 bp it was a sign of 4bp microhomology sequence. The specific bases were identified on Geneious Primer v2020.2.2.

### 3.15 Statistical methods

For an overview of the DNA concentration distributions in the 21 samples, descriptive statistics were used to determine whether the data were normally distributed. Shapiro-Wilk test showing a p-value  $<0.05$  represented non-normally distributed data and the min, max, and median were reported. Parametric Chi-square of independence was used to determine whether there was a significantly higher number of integrations in one HPV type. Chi-square test showing a p-value  $<0.05$  represented a significantly higher number. All statical analyses were performed in Statistical Package for Social Sciences (SPSS) v27.

## 4. Results

### 4.1 Categorization of NGS-reported HPV integration sites

In total 88 HPV31, 89 HPV33, and 56 HPV45 samples have been sequenced with NGS technology. BAM files containing Illumina paired-end sequenced reads aligned to human reference genome hg38 and 183 HPV genomes were loaded onto IGV v2.8.9 for a visual inspection and evaluation of whether reported integrations were likely artefact or true. The reported integration breakpoints from both alignments were categorized as “Yes”, “No” or “Maybe” potential integrations. Figure 21 represents a summary of; 1) the validation results, 2) the qualified group for further investigations, and 3) the confirmed and non-confirmed HPV-integrations.

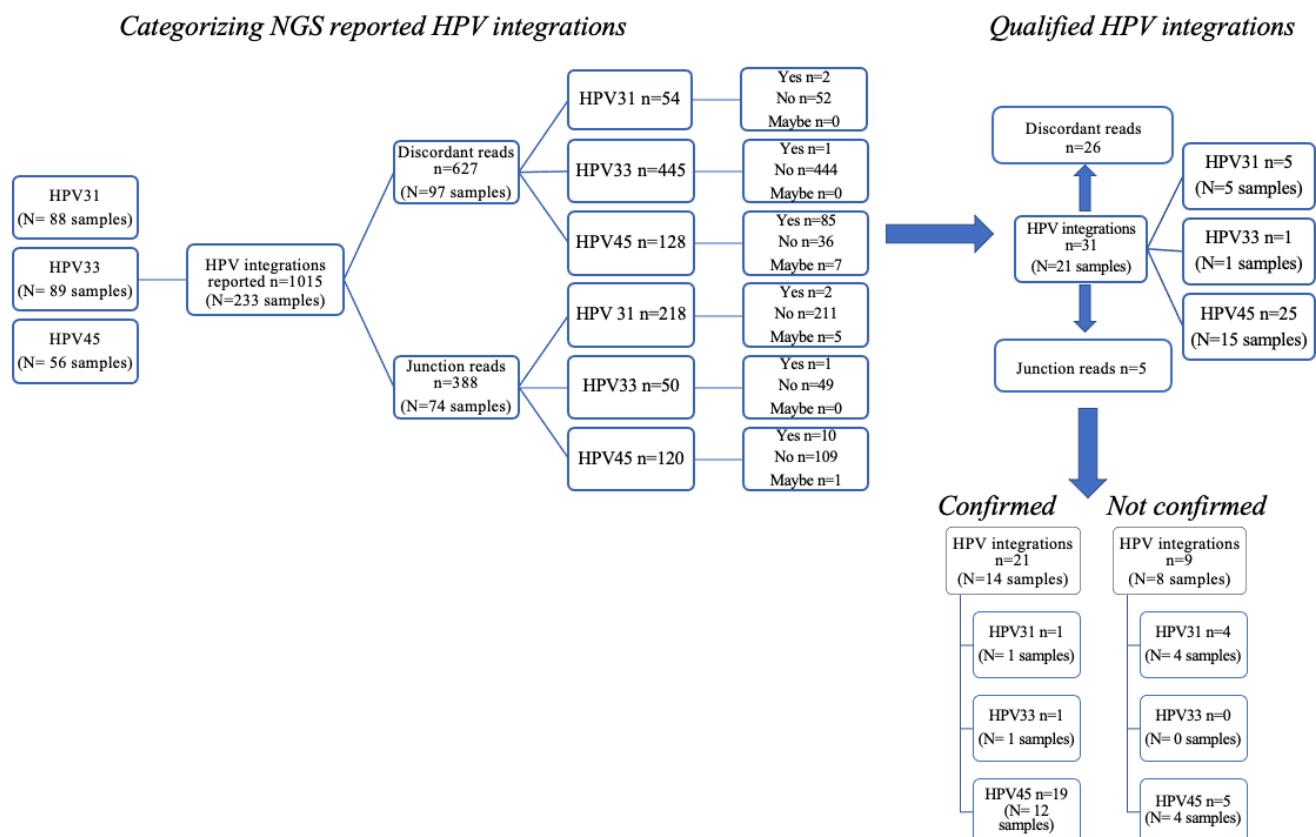
Across the three HPV types, a total of 1015 possible HPV integrations covering both the discordant and junction reads were reported. Discordant reads accounted for 62% (627/1015) of the total, while junction reads accounted for 38% (388/1015). HPV31 positive samples accounted for 8.6% (54/627) of the integrations reported from discordant reads, of which 3.7% (2/54) were categorized as “Yes”. Further, 71% (445/627) of the integrations reported from discordant reads were HPV33 positive samples, of which nearly 100% were categorized as “No” and only one integration categorized as “Yes”. HPV45 positive samples accounted for 20% (128/627) of the integrations reported from discordant reads, of which 72% (92/128) were considered potential integrations (85/128 “Yes” and 7/128 “Maybe”).

When categorizing the 388 reported integrations from junction reads, only a small proportion, 0.9% (2/218) of the HPV31 positive integrations were considered potentially true, categorized as “Yes”. Similarly, only one (1/50) reported integration in the HPV33 positive samples was categorized as “Yes”. In the HPV45-positive samples 9.2% (11/120) reported HPV integrations were considered potential (10 “Yes” and 1 “Maybe”). Both algorithms produced a high number of calls considered “No” based on the known technical error which is detectable when the same start- and stop coordinates are found in the F/R reactions from the same fragment. Breakpoints reported at around 955 bp, 3440 bp, and 3940 bp in HPV31- positive samples, 7130 bp and 7230 bp in HPV-33 and 1394 bp and 5461 bp in HPV45- positive samples were typically classified “No”, usually observed with a polyT in the region.

#### 4.1.1 Qualified HPV integrations

A total of 31 HPV integrations from 21 patient samples were qualified for further investigations. For the 31 qualified HPV integrations, certain integrations had both junction and discordant reads reported at the integration breakpoint, even though template sequences were designed for in total 26 discordant reads and 5 junction reads (Table 8). 9/31 integrations were categorized as “Maybe” mainly in junction reads called by LAST.

The qualified HPV integrations included 16% (5/31) HPV31 positive samples, 3.2% (1/31) HPV33, and 81% (25/31) HPV45 positive samples. The qualified samples had between one and four HPV integrations reported. Sample 13a was coinfecte with HPV45 and HPV31. The clinical diagnostic category CIN3 dominated the sample population, constituting (29/31) of the qualified samples analyzed, while only one sample was in each of the categories, CIN2, and cancer. A breakpoint in the HPV E1 and E2 genes were reported in 17 integrations.



**Figure 21: Validation process, qualified group, and confirmed and non-confirmed HPV integrations.**  
An overview of the process of validating next generation sequencing (NGS) reported Human papillomavirus (HPV) integrations, followed by further investigations of the qualified group and the proportion of confirmed and non-confirmed HPV integrations with Sanger sequencing.

#### 4.2 Making template sequences

Templates were extracted from IGV v.2.8.9 for each qualified integration. An overview of the *in silico* template sequences for the 31 qualified integrations are shown in Supplementary Table 3A (Appendix 3). The table reports each template relative to sample ID, HPV breakpoint, and human chromosome integration site. Furthermore, the number of integrations per sample is interpretable in the same table. For example, sample 1 has two integrations reported (a and b), while sample 21 has four integrations reported (a, b, c and d)

Some template sequences were extended to enable design of appropriate and specific primers in the Primer3 and Primer3plus. The programs did not find suitable primer pairs for some of the template sequences that may represent a challenging region. The average length of template fragments was 215 bp.

#### 4.3 Forward and Reverse primer design

The designed primers ensured optimal PCR amplification across the reported HPV integration. The F primers were human-specific, while the R primers were HPV-specific. Primers used for each template are presented in Supplementary Table 3A (Appendix 3). Some of the primers contained regions of somewhat lower specificity.

#### 4.4 Semi-quantitative and a qualitative validation of the Polymerase Chain Reaction (PCR) products

The DNA concentrations in the samples were non-normally distributed ranging from 0.446-39.0 ng/µL (min- max) with a median value of 8.40 ng/µL, representing a wide distribution of the concentrations.

For a visual semi-quantitate and qualitative validation of the PCR reactions and products and an indirect determination of the primer specificity, an agarose gel electrophoresis was performed. When gel smears and un-satisfactory Sanger sequencing results (no or low-quality) were observed, new optimized reactions were set using different voltage- and running time conditions. The Supplementary Table 4A (Appendix 4) shows results of the various gel runs of samples 1a-21d linked to gel picture A-α. The screenshot from the various gel runs in each sample is placed approximately next to each other in Supplementary Table 4a (Appendix 4) as some of the gels were run under different conditions and some contained smears, making

accurate relative placement challenging. The fragment sizes of the ML were also placed in relative approximate positions.

#### 4.4.1 T-PCR

Initially, a traditional PCR (T-PCR1) was performed and run on a 2% agarose gel at 100 V for 60 min. Samples 1a, 3a, 6b, 8a, 9a, 11a, 11b, 12a, 12b, 14a, and 15a used in T-PCR1 provided high-quality Sanger sequencing results when sequenced directly, while the remaining samples provided no- or low-quality Sanger sequences that did not provide identifiable sequences (Figure 23). T-PCR1 of samples 18b, 19a, 20a, and the CaSki control was unsuccessful (Figure 22) (Supplementary Table 4A), probably the result of a pipetting error. Therefore, control gels were run with CaSki showing a distinct band at ~500 bp as expected. Sample 21 was not included in the T-PCR1 setup nor directly Sanger sequenced as the sample could not be found in the biobank when the other samples were prepared for T-PCR1 and Sanger sequencing.

Sample ID: 18b	Sample ID: 19a	Sample ID: 20a	Pos CaSki ctr
Y1	Z1	$\beta$ 1	Ca1

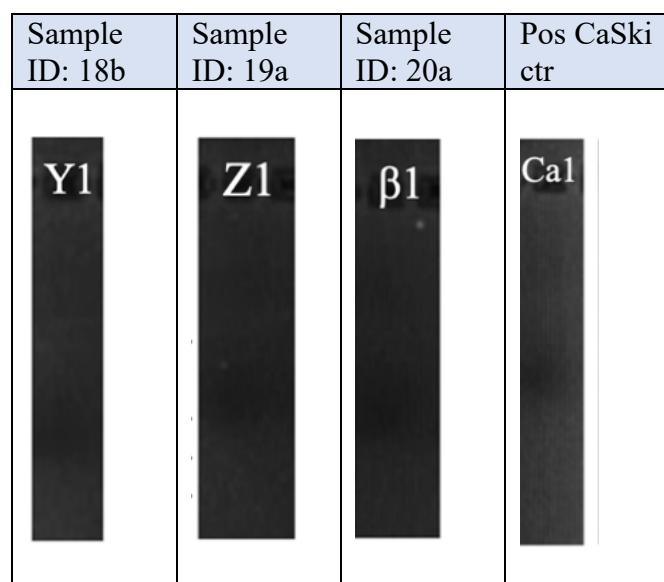


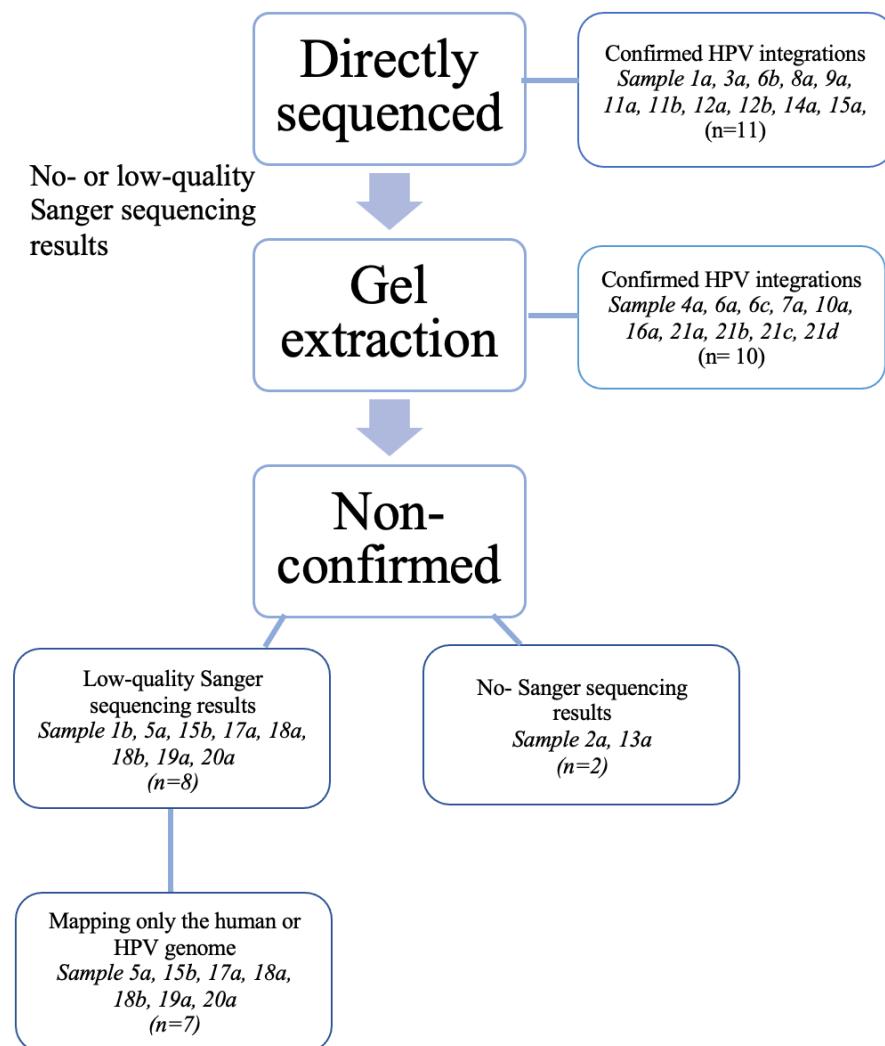
Figure 22: Samples and positive (pos) CaSki control (ctr) not showing Polymerase Chain Reaction (PCR) products when performing Traditional PCR, followed by 2% agarose gel run at 100 V for 60 minutes (T-PCR1)

#### 4.4.2 TD-PCR

TD-PCR was performed in an attempt to eliminate un-specific bands. The specificity was in several samples increased showing a better band separation when running a gel at 70 V for a prolonged time, 120-135 min. Multiple gel bands were still present in 84% (26/31) of the reported HPV integration breakpoints, some with the combination of both weak bands and

distinct bands with a visually higher amount of PCR products. Fragments <30 bp were considered to be primer dimers.

Prior to the PCR reactions, the fragment size of discordant read pairs were unknown. When visualizing the gel and eliminating the gel-bands that represented short DNA amplicons, the PCR products were of 100-800 bp in length. Templates made from junction reads had predicted product sizes between 174-297 bp, however, only 1/5 samples displayed a gel-bands that matched the predicted product size.



**Figure 23: Confirmed and non-confirmed number of Human papillomavirus (HPV) integrations.** Presents the qualified HPV integrations either confirmed by directly Sanger sequencing or by Sanger sequencing of the gel-eluates. In addition to the non-confirmed HPV integrations being a result of no- or low-quality sequence despite the adjustments.

Sample 2a contained smears with several bands in the T-PCR1, however, not reproduced in consecutive gel runs. Similarly, when performing T-PCR1 on sample 13a, several bands were identified, but only a few or no bands were identified in subsequent PCR reactions (Figure 24).

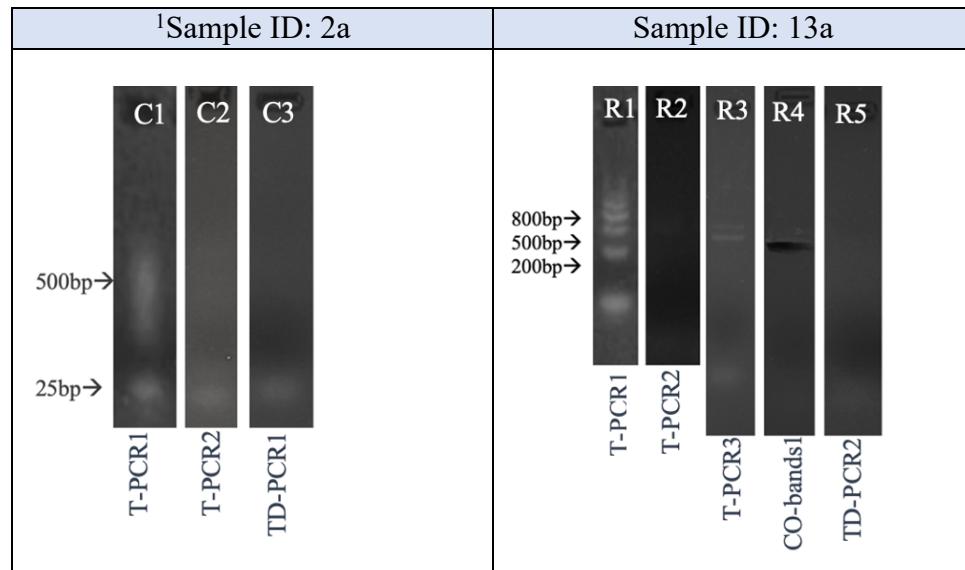


Figure 24: Samples with not reproducible agarose gel results.

#### 4.5 DNA elution from gel-bands

Gel extractions were used for samples with visible gel-bands but showed no- or low-quality Sanger sequencing when sequenced directly (Figure 23). Each of the agarose >30 bp gel- bands were cut out and placed in a microcolumn. An overview of the bands used for further investigations is presented in Supplementary Table 4A (Appendix 4).

Sanger sequencing of CO-bands1 was unsuccessful. Measuring the DNA concentrations in 10 randomly undiluted samples that were unsuccessfully sequenced, including the positive CaSki-control showed low values (0.0212-2.520 ng/µL and the positive CaSki control 0.5 ng/µL). Conversely, performing CO-bands2 increased the DNA concentrations of samples to 3.11-164 ng/µL and the positive CaSki control to 194 ng/µL. The filter was also reeluted in 25 µL nuclease-free water to control the DNA concentration. The eluate in 25 µL showed consistently 3-4 times lower DNA concentration.

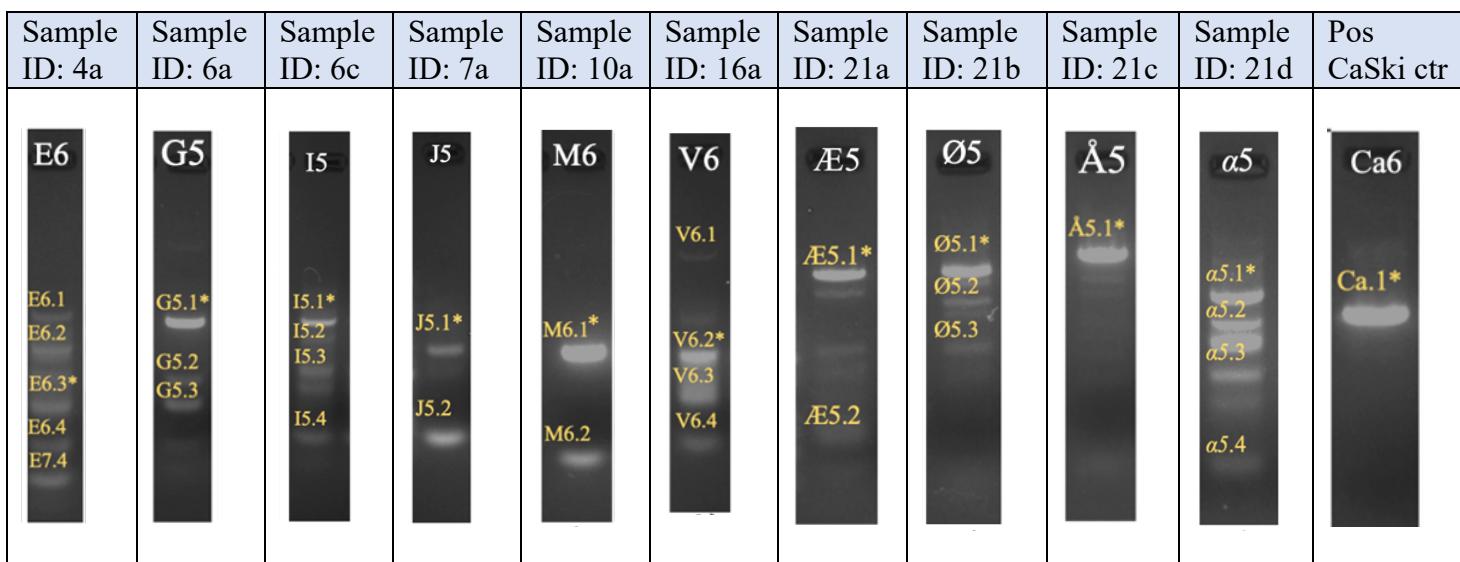
## 4.6 Sanger sequencing data analysis

Sanger sequencing was performed to detect hybrid sequences harboring both human and HPV sequences. The proportion of the confirmed and non-confirmed HPV integrations are shown in Figure 21 and 23.

### 4.6.1 Confirmed HPV integrations

68% (21/31) of the qualified HPV integrations in a total of 14 samples were confirmed by Sanger sequencing (Table 8), including one HPV31, one HPV33, and 19 HPV45- samples. The CIN3 diagnostic category accounted for 95% (20/21) of the confirmed HPV integrations. The remaining sample was in the cervical cancer diagnostic group. The Sanger sequencing chromatograms from the confirmed integrations are presented in Supplementary Table 5A-5Y (Appendix 5). Three out of eight (3/8) HPV integrations classified into the “Maybe” category was confirmed.

Ten out of 21 (10/21) integrations were confirmed with the gel extraction method and the rest were confirmed when sequenced directly. The gel-bands that resulted in confirmed HPV integrations were E6.3\*, G5.1\*, I5.1\*, J5.1\*, M6.1\*, V6.2\*,  $\Delta$ E5.1\*,  $\emptyset$ 5.1\*,  $\text{A}5.1^*$  and  $\alpha$ 5.1\* presented in Figure 25 (Supplementary Table 4A, Appendix 4). The positive CaSki-control was both confirmed when sequenced directly and by the gel extraction method, seen as gel-band Ca.1\* in Figure 25 (Supplementary Table 4a, Appendix 4). The other gel-bands from the samples either resulted in no- or low-quality sequences, in certain cases partly identifiable, mapping to either the human or HPV genome. It was typical the distinct gel-bands in CO-bands2 with a DNA concentration  $>10.0$  ng/ $\mu$ L that provided high-quality Sanger sequencing results.



*Figure 25: Human papillomavirus (HPV) integrations confirmed by the gel extraction method. TD-PCR followed by 2% agarose gel run at 70V for 135 minutes (TD-PCR2)*

A continuous sequence with approximate >90% pairwise identity between the F- and R- sequences was found in 62% (13/21) of the confirmed HPV integrations. This includes the following samples: 1a, 4a, 6a, 6c, 7a, 9a, 10a, 11a, 12b, 14a, 21a, 21b and 21c, in addition to the positive CaSki-control.

The F-sequence in sample 12a mapped to more than one human chromosome (chr14, 12, 1, 17, 16) and HPV45. However, the R-sequence mapped only to human chromosome 3 and HPV45 as originally reported, thereby the discovered integration was considered confirmed (Supplementary table 5M and 5N Appendix 5). The F-sequence from sample 15a also mapped to more than one human chromosome (chr5, 4, and 11), HPV45, and partly to HPV97 and 18 (Supplementary Figure 5Q and 5R, Appendix 5). The R- sequence, however, mapped to more than one chromosome, among these chr8 and HPV45 thereby confirming the original NGS result.

#### 4.6.1.1 Identity to known human genes

Four out of 21 (4/21) confirmed HPV integrations displayed >99% identity to known human genes identified by BLASTn. Sample 1a showed identity to a gene encoding the tumor suppressor protein p63 (Supplementary Figure 5A, Appendix 5), sample 6a showed identity to the SH3 domain and tetratricopeptide repeats 2 (SH3CT2) (Supplementary Figure 5D,

Appendix 5), sample 7a to the NHS like 1 transcript and sample 14a to Wilms proteins (Supplementary Figure 5G, Appendix 5).

#### 4.6.2 Non-confirmed HPV integrations

Ten out of 31 (10/31) qualified HPV integrations were not confirmed by directly Sanger sequencing nor by the gel extractions (Figure 21, Figure 23, and Table 8). When performing CO-bands2, it was typically the gel bands eluted in 15 µL and contained a DNA concentration of <5 ng/µL that provided no or low-quality Sanger-sequencing results.

Sample 1b gave low-quality Sanger-sequencing result when sequenced directly and by the gel-extraction method, not providing an identifiable sequence. Sample 2a and 13a did not provide Sanger sequencing results when sequenced directly and did not provide reproducible results in later gel runs to enable CO-bands2 (Figure 24) (Supplementary Table 4A, Appendix 4).

Only low-quality sequences were obtained when sample 15a was Sanger sequenced directly. Moreover, the gel extraction method prior to sequencing also failed to increase the quality of the obtained sequences. However, the F and R sequences from gel-band F6.3 mapped to more than one human chromosome, one of them being chromosome 17 as originally reported but did not map to HPV.

A similar result was obtained from the sample 15b. However, the F-sequence contained partly identifiable sequence mapping to chromosome 8. This was previously reported when NGS was applied in an attempt to detect HPV integrations, but the obtained sequence did not map to any HPV type.

Sanger sequencing of sample 17a resulted in low-quality sequences. However, sequencing of gel-band W5.1 and W5.2 was more successful. Unfortunately, F-sequence of W5.1 mapped to more than one chromosome and not to any HPV type, while R-sequence of W5.2 mapped to HPV45 but not to the human genome.

Sample 18a also had low-quality Sanger sequencing results when sequenced directly. Sequencing of gel-bands X5.1 and X5.2 resulted in somewhat low-quality Sanger sequencing results. The identifiable part of the sequences mapped to more than one human chromosome but not to the chromosome as originally reported and not to HPV.

Comparable results were obtained from sample 18b. Partly identifiable sequences were obtained from gel-bands Y6.1 and Y6.2 that mapped to more than one human chromosome, not including the human chromosome originally reported and not mapping to any HPV type.

Direct Sanger sequencing of sample 19a once again resulted in a low-quality sequence. Sequencing of gel-bands Z6.1, Z6.2, Z6.3, and Z6.4, resulted in a combination of low-quality and partly identifiable sequences that mapped both the chromosome 6 as originally reported and to several other human chromosomes and not to HPV.

Sequencing of the extracted gel-bands  $\beta$ 5.1 and  $\beta$ 5.2 from sample 20a was more successful than sequencing directly. However, partly identifiable sequences mapped to the human chromosome 15 and not to chromosome X as originally reported. In addition, the sequence did not map to any HPV type.

#### 4.7 Microhomology regions

Overlapping sequences between the human and HPV genomes at the integration breakpoint were identified as microhomology regions. Overlapping regions were identified with BLAT and BLASTn, while the specific overlapping sequences were identified with Geneious Prime v2020.2.2. A microhomology region at the integration breakpoint was identified in 24% (5/21) of the confirmed HPV integrations, ranging in length from 3 bp to 12 bp (Table 7).

*Table 7: Microhomology regions*

<sup>1</sup> Sample ID (n=5 samples)	Microhomology sequences (bp)
1a	ATT (3)
6a	GATAAT (6)
8a	ACTGTT (6)
14a	AAAGGAA (7)
15a	CAGATAGAAAGG (12)

<sup>1</sup> Representing the 24% (5/21) of the confirmed HPV integrations identified with a microhomology region at the integration breakpoint. The table demonstrates the sample ID and the microhomology sequence.

Abbreviations: bp = base pairs, HPV = Human papillomavirus

Table 8 shows the 31 qualified NGS reported HPV integrations, the location of the human and HPV breakpoints, the number of discordant and junction reads at each integration, whether the reported integration was classified as “Maybe” or “Yes” categories during the validation

process, whether the integrations had been confirmed or not confirmed with Sanger sequencing and whether a microhomology sequence was identified.

Table 8: Qualified 31 Human papillomavirus (HPV) integrations from 21 samples.

Sample ID (n=31 integrations) (n=21 samples)	Diagnostic category	HPV			Human GRCh38/hg38)			Junction reads	Discordant reads pairs	Potential integration	Confirmed with Sanger sequencing <sup>1</sup>	Microhomology (bp) <sup>2</sup>
		Type	Position	Gene	Chr	Position	Locus					
1a	CIN3	45	1393	E1	3	189876913	3q28	0	22	Yes	Yes (A)	Yes (3)
1b	CIN3	45	4358	L2	3	189955746	3q28	0	2	Yes	No (B)	No
2a	CIN3	31	1235	E1	7	26640555	7p15.2	16	0	Maybe	No (C)	No
3a	CIN3	31	6169	L1	13	73210124	13q22.1	0	7	Yes	Yes (D)	No
4a	CIN3	45	2624	E1	1	8859323	1p36.23	0	2	Yes	Yes (E)	No
5a	CIN3	45	5168	L2	17	27329570	17q11.1	12	0	Maybe	No (F)	No
6a	CIN3	45	2624	E1	5	149029187	5q32	0	43	Yes	Yes (G)	Yes (6)
6b	CIN3	45	2888	E2	5	148950210	5q32	0	26	Yes	Yes (H)	No
6c	CIN3	45	3390	E2	3	116874819	3q13.31	0	2	Maybe	Yes (I)	No
7a	CIN3	45	892	E7	6	138481138	6q24.1	0	3	Yes	Yes (J)	No
8a	CIN3	45	3669	E2	X	114942562	Xq23	72	0	Yes	Yes (K)	Yes (6)
9a	CIN3	45	4865	L2	15	58280064	15q21.3	0	66	Yes	Yes (L)	No
10a	Cancer	45	6852	L1	11	102911491	11q22.2	0	17	Yes	Yes (M)	No
11a	CIN3	45	2127	E1	13	48516810	13q14.2	0	9	Yes	Yes (N)	No
11b	CIN3	45	3893	E2	13	48491541	13q14.2	0	23	Yes	Yes (O)	No
12a	CIN3	45	1646	E1	3	160748989	3q25.33	0	3	Yes	Yes (P)	No
12b	CIN3	45	6852	L1	3	160749198	3q25.33	0	2	Yes	Yes (Q)	No
13a	CIN3	45	1646	E1	11	102867368	11q22.2	0	32	Yes	No (R)	No
14a	CIN3	45	3390	E2	11	102831633	11q22.2	0	77	Yes	Yes (S)	Yes (7)
15a	CIN3	45	2127	E1	8	86002065	8q21.3	0	30	Yes	Yes (T)	Yes (12)
15b	CIN3	45	5114	L2	8	85947193	8q21.3	0	37	Yes	No (U)	No
16a	CIN3	33	4389	L2	5	53351927	5q11.2	0	2	Yes	Yes (V)	No
17a	CIN3	31	3689	E2	2	23668560	2p24.1	11	0	Maybe	No (W)	No
18a	CIN3	45	2875	E2	1	209430141	1q32.2	0	13	Yes	No (X)	No
18b	CIN3	45	5115	L2	1	209409245	1q32.2	0	2	Yes	No (Y)	No
19a	CIN2	31	4893	L2	6	80251636	6q14.1	15	0	Maybe	No (Z)	No
20a	CIN3	31	214	E6	X	43763326	Xp11.1	4	2	Maybe	No ( $\beta$ )	No
21a	CIN3	45	1646	E1	8	109485377	8q23.1	0	4	Maybe	Yes ( $\mathcal{A}$ E)	No

21b	CIN3	45	2875	E2	9	125230778	9q33.3	0	8	Yes	Yes ( $\emptyset$ )	No
21c	CIN3	45	5394	L2	8	109595503	8q23.1	0	7	Maybe	Yes ( $\text{\AA}$ )	No
21d	CIN3	45	4358	L2	9	125230411	9q33.3	0	3	Yes	Yes ( $\alpha$ )	No

<sup>1</sup>Represents HPV integrations confirmed with Sanger sequencing. The number in the brackets indicates agarose gel position in Supplementary Table 4A (Appendix 4)

<sup>2</sup>The microhomology sequence length identified at the integration breakpoint.

Abbreviations: CIN3= Cervical intraepithelial neoplasia grade 3, Chr= Chromosome, GRCh8/hg38= Human reference genome 38

## 5. Discussion

The aim of this study was to validate NGS reported HPV integrations in HPV31, 33, and 45 positive samples and to characterize hot-spot and microhomology regions at the integration breakpoints. Integrations are potential biomarkers for predicting cancer progression.

### 5.1 Clinical aspects

#### 5.1.1 Higher integration rate in HPV45 positive samples with a CIN3 diagnostic category

In this study, 68% of the HPV31, 33, and 45 positive samples qualified for validation were confirmed. In HPV45 positive samples, the number of qualified and confirmed integrations (20) was higher than in HPV31 (0) and HPV33 (1) positive samples. Since HPV31 had zero in proportion, comparing HPV33 and HPV45 positive samples revealed that HPV45-positive samples had a significantly higher proportion of integrations ( $p < 0,00001$  Chi-square test).

A previous study reported that the frequency of HPV integrations was higher in HPV 16, 18, and 45-positive samples than in 31 and 33-positive samples (8). Moreover, in the same study, APOT was used as a detection method for HPV integrations at the mRNA level. Consequently, results may be biased due to integrations not being transcribed rather than not being present. The number of integration studies performed at the DNA level in HPV31, 33, and 45 positive precancerous lesions and cancer is limited, making a proper comparison with other studies difficult.

In this study HPV45-positive samples with a CIN3 diagnostic category dominated the population, both in the qualified and confirmed group. The APOT study (8), also showed that in CIN3 cases, HPV45-positive samples showed the highest proportion of integrated HPV, followed by HPV16 and HPV31, whereas no integrated DNA was detected in HPV18 and 33 (8). Unlike the previous study, this study confirmed integration in one HPV33-positive CIN3 sample.

A study conducted by *Liu et al.* on a DNA level based on NGS technology and validation with Sanger sequencing observed a higher rate of HPV integrations in CIN3 than in normal epithelium, CIN1, or CIN2 (5). The highest number of confirmed samples was also categorized as CIN3 in this study. One of the confirmed HPV integrations was a HPV33 positive CIN3 sample, similar to the findings in this study. However, the study was limited by the low number

of HPV31, 33, and 45 samples in the study population, in addition to the small proportion of HPV integrations confirmed by Sanger sequencing (5). According to previous findings reflecting a higher integration rate in CIN3 samples, from biological aspect, it may not be surprising that the most prevalent HPV integrations qualified for further analyses were HPV45-positive CIN3 samples.

Only one HPV45-positive sample was qualified and confirmed in the diagnostic group of cervical cancer. The APOT study (8), showed that the integration rate in cervical cancer cases was higher in HPV18, followed by HPV45, 16, 33, and 31. The reason for the higher integration rate in HPV18 and HPV45 positive cervical cancer cases might be caused by the higher prevalence of HPV18 and 45 in ADC. As ADC develops in glandular cells localized in the inner cervical canal (34), a proper sampling including cells from the inner cervical canal might be challenging to obtain. Therefore, precancerous lesions become more difficult to detect which increases the probability of persistence and progression to cervical cancer. The difficulty of obtaining adequate samples may increase even more when/if self-sampling gets introduced in screening.

The large CIN3 group and the lack of a balanced distribution of other diagnostic groups was a result of the categories being unknown during the validation process to prevent biased sample prioritization and quick conclusions based on knowledge about HPV integrations. Due to this fact, it was challenging to compare the rate of HPV integrations across multiple diagnostic groups and to uncover a potentially higher integration rate in late CIN stages and cancer. To confirm that viral integration is an early event as previously reported (152, 153), samples with confirmed HPV integrations in earlier stages than CIN3 need to be obtained. Alternatively, obtaining a follow-up sample from the women with confirmed HPV integrations having CIN3 would also be informative, however, this would not be in line with human ethics. Data from the women's clinical histories can also be useful in identifying possible risk factors such as previous exposure to sexually transmitted diseases, periods of immunosuppression, which raise the risk of HPV infection, persistence, integrations, and cancer progress.

### 5.1.2 HPV45 positive samples with more than one reported HPV integration

Mainly HPV45 positive samples were reported with more than one HPV integration. These samples may cause a higher grade of instability, more likely leading to a cancer progression. The detection of HPV in an integrated form may also depend on the rate of episomes in the

sample as a low number of integration sites could be undetected by the presence of a high background of episomal HPV (11, 87).

### 5.1.3 Localization of HPV breakpoints and integrations

#### 5.1.3.1 *E1 and E2*

A proportion of the qualified HPV integrations had breakpoints reported in HPV E1 or E2 genes. The study by *Liu et al.* also reported HPV integration breakpoints mainly located in E1 and E2 genes (5). The E2 protein is a known negative regulator of E6 and E7 expression. When the E2 gene is disrupted by HPV integration, the E2 gene expression is hindered resulting in the E6 and E7 overexpression. Since E2 and E1 share the same ORF, a break in E1 can also result in an E2 break. Disruption of E1 and E2 may lead to a higher oncogenic potential (154, 155). This may not be surprising as overexpressed HPVE6 and HPVE7 inhibits the activity of important cell cycle regulators, p53 and pRb (2). Consequently, the cell with damaged DNA would be allowed to continue its cycle with a potential malignant tendency.

#### 5.1.3.2 *Detected HPV integrations in human genes*

##### 5.1.3.2.1 Previously reported cancer-related genes

Two of the confirmed HPV integrations mapped to previously reported cancer-related genes, TP63 and Wilms protein. HPV integration into TP63 was similar to earlier findings (9, 79). A study performed on HPV-induced cervical neoplasia demonstrated a correlation between increased expression of the p63 gene and aggressive cancer progression (156). HPV integration into the gene encoding p63 has also been shown to have a critical outcome in head and neck, and penile cancer (157-159).

Another confirmed sample mapped to the tumor-associated protein, Wilms protein. This protein has been identified as highly responsible for carcinogenic development in various cancer types, including gynecological tumors such as ovarian cancer. Wilms protein in gynecological cancer studies was also found associated with poor prognosis (160, 161). These findings reflect the importance of the proteins also in other cancer types, and the potential outcome by disrupting the encoding gene.

##### 5.1.3.2.2 Previously not reported cancer-related genes

Two of the confirmed HPV integrations mapped to genes encoding SH3 domain and tetratricopeptide repeats 2 (SH3CT2), and NHS like 1 transcript. SH3CT2 gene has not been

previously associated with cancer development (162, 163). Another sample mapped to NHS like 1 transcript that is partly expressed in endometrium tissue, however, not associated with cervical cancer (164). Although these genes may not be specifically linked to cancer, they could play a role in the formation of an irregular cell population with a distinct morphology that could have been identified during screening.

#### *5.1.3.3 Non-random distribution of integration sites?*

14/31 qualified HPV integration had human breakpoints identified at 1p, 3q, 6q, 11q and 13q chromosome loci, similar to a previous HPV 16 and 18 study (9). Two HPV integrations were reported with a breakpoint in 3q28, two in 11q22.2, and one in 13q22.2 specific regions. Two of the confirmed samples had identical integration sites in the human chromosomal locus 11q22.2. Integrations in 3q28, 11q22.2, and 13q21-22 have also been reported previously (9, 40, 70, 76), indicating a non-random distribution of the integration sites. This might be because DNA is less densely packed and less coiled in regions with expressed genes, allowing for HPV integration (75).

When identifying an HPV integration pattern for instance hot-spot regions, it may be easier to develop a new method and implement HPV integration as a potential biomarker. This may be done by design of specific primer pairs to the hot-spot regions and perform PCRs. PCR is cheaper than NGS and does not require analysis of the whole genomes to detect HPV integrations. The implementation of such method could be beneficial in developing countries not having the access to expensive instruments and equipment.

#### *5.1.3.4 Microhomology regions identified at the integration breakpoint*

Microhomologies at the breakpoint was identified in 24% (5/21) of the confirmed HPV integrations. Microhomology regions at the integration breakpoint have also been identified in other studies (12, 165). The discovered microhomology indicated that the fusion between viral and human DNA may have occurred during the microhomology-mediated DNA repair pathways (12). However, the detected 3bp microhomology region was short and more likely randomly distributed than the 12bp microhomology region. Breakpoints in HPV E1 and E2, integrations in hot-spot regions, and human tumor suppressors, and the presence of microhomologies can be useful in uncovering important events of the integration.

#### 5.1.4 Why chromosomal integrations as biomarkers?

As illustrated in Figure 2 in the introduction section (*1.2 HPV infection, pathology, and cancer progression*), although a woman may have had CIN3 for 10 years, there is still a high rate of persistence and regression and only a small percentage that progresses to invasive cancer without treatment (39). The challenges arise when CIN3 lesions are discovered in screening and the decision to treat or not to treat has to be made. As dysplasia can regress, there is a high risk of overtreatment. Overtreatment may damage the cervix, resulting in later pregnancy and birth complications. Conversely, not treating and waiting for a potential regression may lead to a progression to cancer that would require even larger and more complex interventions. However, current treatment for precancerous lesions is relatively efficient in terms of preventing further complications. Introducing detection of HPV integrations as part of the screening program may guide and customize prevention and treatment options in line with a personalized medical focus. Harald zur Hausen made a revolutionary discovery in the 1980s correlating the HPV virus to cervical cancer (2). Hence it is important to continue the research to uncover other molecular mechanisms of the virus.

#### 5.1.5 The corona pandemic and increasing HPV research?

The SARS-CoV19 pandemic may have made cervical cancer screening more challenging (166). Several women may have declined or delayed taking a cell sample to avoid putting extra strain on the health care system and out of fear of infection. In this context, the importance of self-sampling tests has become even more attractive (167). It is also conceivable that the importance and impact of vaccination has become clearer because of the corona pandemic.

Developing countries may have lacked biomolecular instruments and detection facilities in the past, but this may have improved during the corona pandemic. More laboratories have opened for detecting the coronavirus using technologies adaptable for a range of viral diagnostics. This may be useful for future clinical diagnostics and research that contributes to international cooperation by sharing knowledge worldwide. This may be especially evident through the work of practicing preventative strategies and developing new molecular methods and biomarkers, including HPV integrations. A higher contribution to the HPV research field is linked to an increase in clinical discussions and methodological considerations. Establishing a new biomarker could also be necessary, especially during the ongoing corona pandemic to prevent HPV-positive women who are not at risk for cervical cancer development from attending the screening.

## 5.2 Methodological consideration

### 5.2.1. Sample material

A low viral load in the samples could be undetected by the presence of high background episomal HPV, causing no or low- sequencing yield (11, 87). Additionally, all molecular methods depend on a relatively high concentration of input DNA material reflecting once again the importance of high viral load in the sample.

### 5.2.2 NGS-reported data

NGS technology is time-consuming in processing sequencing data and requires a high storage capacity and bioinformatics expertise (94). This can be challenging when implementing the method in the routine, especially in developing countries where preventive actions and infrastructure are still lacking. Another challenge is that patients may find the ethical consequences of sequencing genetic material strange and frightening. Therefore, protocols and legislations for the use of genetic material in diagnostics and research must be well established. Patients must also be well informed about the biomarkers and methods through good communication with a health provider that uses understandable language. It is also important to obtain informed consent from all women participating in HPV research programs while also clearly stating which results are passed to them. The research laws and guidelines are both regulated on an international (European) level, and national level to encourage good and ethical medical and health research. On a national level, it is regulated by the medical and health research act, controlled by REK (168).

### 5.2.3 Validation of HPV integrations

All HPV integrations reported were manually processed and categorized. Some samples had both discordant and junction reads reported in the same location, whereas others had either discordant, or junction reads. Either junction or discordant reads was used to create a DNA template. Therefore, when 25 discordant and 6 junction reads were used for template design it might have appeared as a small number compared to the number originally reported from the NGS. The polyT regions might have interfered with the mapping process when creating the BAM files causing the high number of false-positive integrations, especially evident in HPV33-positive samples. The settings could have been adjusted in such manner to eliminate integrations in HPV regions frequently erroneously reported. However, adjustment of IGV settings might also exclude the potentially true integrations.

For a sample to be qualified for further analyses, several conditions needed to be satisfied. However, the criteria might have been too strict as 3/8 HPV integrations classified as “Maybe” were confirmed. These confirmed HPV integrations were almost the only ones in the “Maybe” category with >2 discordant reads reported, in addition to junction reads. When aligning the junction reads, the reads typically aligned with same start and stop coordinates but with 1-2 reads shifting to the right, giving origin to different start and stop coordinates.

Conversely, the non-confirmed HPV integrations in the “Maybe” category had reads aligning with the same start- and stop coordinates with usually a few shorter reads as a result of trimming. Consequently, it was challenging to exclude short reads as they could align with the same or different start- and stop coordinates if the reads were of 125 bp.

Most reads with identical start and stop coordinates were categorized as “No” during the validation process, indicating a PCR artefact. The IGV software settings could also have been changed to filter away reads with the same start and stop coordinates to avoid false-positive integrations. However, this would result in missing a potentially true HPV integration if it was covered with only a few reads. IGV also has a weakness of not displaying all reads. Therefore, it is possible that an integration categorized as “No” could potentially be true.

#### 5.2.4 Template design

*In silico* templates covering both human and HPV genome were manually designed. Although the templates were in certain cases expanded to obtain proper primer pairs, the regions in the middle might still be challenging to sequence. In certain cases, this part of the sequence could have provided no- or low-quality Sanger sequencing result, necessary to identify homology to either the human or HPV genome.

The lack of recommended primer pairs was primarily a result of unbalanced distribution of the bases. A template with a high G/C ratio required a higher temperature in the denaturation step, which could result in secondary structures or primer dimers. Moreover, secondary structures or primer-dimers might have affected the activity of the DNA polymerase in the PCR (169) reducing its efficiency and leading to the amplification of several non-targeted regions. Non-specific amplification was observed in several samples, Supplementary Table 4A (Appendix 4). Multiple gel-bands observed in these samples were mostly related to designed primer pairs with known but unavoidable low specificity. However, cases where the primer pairs should

have been specific also generated several products observed as multiple gel-bands. The latter was mainly a pattern of several weak bands, but one distinct gel-band containing much PCR product.

As templates generated from discordant reads almost exclusively led to confirmed HPV integrations, the HISAT2 algorithm might have been more reliable than LAST algorithm. The LAST alignment can determine the correct position of the human-integration breakpoint (10). However, HISAT2 algorithm is more sensitive, specific, and more stringent. As a result, LAST alignments reports higher number of false-positive integrations than HISAT2. However, the integration calls from NGS were disproportionately represented by discordant and junction reads. The relative high number of calls from discordant reads may have increase the possibility for confirming and integration from these.

Another possibility for the high number of confirmed HPV integrations with templates generated from discordant reads may be that the templates from junction reads were not optimally designed. Consequently, this might have led to incorrect primer pairs, suboptimal PCR reactions, and thereby Sanger sequencing. Template expansion may introduce error as additional design steps are required. However, one of the confirmed HPV integrations was based on an extended template made from junction reads, while other non-confirmed integrations with templates made from junction reads did not require expansion. This reflects the possibility that the junction reads templates were not made incorrectly although the template was extended. Chimeric HPV-DNA sequences might have also occurred during the PCR reactions causing unspecific products. These sequences occur when a single DNA strand is amplified from more than one template. During a simultaneous amplification of homologous sequences, a generation of chimeric DNA molecules is a common artefact (170-172).

### 5.2.5 Primer design

During the investigation of primer pairs specificity, several F and R-primers exhibited a cross-binding to different human chromosomes and several HPV types, respectively. The latter was mainly linked to phylogenetically close HPVs. Primer pairs homologous to other HPV types or human chromosomes are not unusual as the primer sequence is relatively short, ~ 25 bp, making it difficult to completely exclude cross-binding to other genome regions. Cross-binding was prevented by expanding the template sequences, however, in certain cases cross-binding was unavoidable. Other primer pairs may be used to increase the primer specificity.

### 5.2.6 Agarose gel electrophoresis and visualization of the PCR products

TD-PCR might have improved the precision and sensitivity (173). However, the DNA products amplified by TD-PCR were loaded on a gel that was run at a lower voltage condition compared to the T-PCR1. The lower voltage might have improved band-separation and decreased the number of smears.

### 5.2.7 DNA elution from gel-bands

It was typically the gel-elutes containing high DNA concentrations that resulted in high-quality Sanger sequences. During the first time of CO-bands1 extraction, visualization of the weak band under the UV-light was difficult. As a result, the weak bands were orientated according to the ML. This process required a longer time and thereby prolonged UV-exposure. This may have contributed to DNA degradation (174), lowering its DNA concentration ultimately leading to no or Sanger sequencing results. In addition, during CO-bands1 the gel-bands were eluted in higher nuclease-free water volume, reducing the DNA concentration even more. When performing CO-bands2 extraction the gel eluates were not diluted prior to the sequencing PCR step as the components potentially interfering with the PCR were most likely were eliminated during the washing steps in the gel extraction procedure. As DNA concentration is crucial an ethanol precipitation of the samples or adding several parallels in the PCR can be performed.

Samples confirmed by the gel extraction method usually had multiple gel-bands with one distinct band containing highly concentrated PCR product. This band usually provided high-quality Sanger sequencing results, while the rest of the gel bands typically provided no or low-quality sequences. The low-quality sequence was either not identifiable or partly identifiable mapping to either human or HPV genome (Figure 23).

### 5.2.8 Analyzing Sanger sequencing data

Sanger sequencing is still a common method in validating NGS results and has been used previously to validate reported HPV integrations (5).

Several samples were observed with continuous sequences. However, cases of the non-continuous sequences were mainly a result of one of the F/R sequences having low-quality, either partly or entirely, and the other F/R sequence having a high-quality mapping to both the

genomes. Although the sequences were corrected for technical artefacts, these corrections were not performed to such an extent that the whole sequence was changed. Some of the sequences also mapped multiple human chromosomes without this being an issue during the primer design. This was especially evident in low-quality sequences. HPV integrations confirmed by the gel extraction method had the highest proportion of continuous sequences, reflecting sequencing unspecific PCR reactions. Multiple washing steps during the gel extractions may have contributed to obtaining high-quality sequences by removal of possible inhibitor such as nucleotides or other components from the PCR.

The non-confirmed HPV integrations were mainly a result of 1) no Sanger sequencing results despite adjustments, 2) low-quality sequencing results not giving an identifiable sequence or 3) low-quality sequencing with partly identifiable sequence mapping to either the human or HPV genome (Figure 23). Therefore, the non-confirmed HPV integrations may still contain an HPV integration. Especially the low-quality sequences with a partly identifiable sequence that only mapped to either the human or HPV genome. It is possible that the unidentified portion could have mapped both the human genomes given a high-quality sequence would have been present, which would have complemented the requirements of a confirmed HPV integration.

Besides, low DNA concentrations in the samples, no or low-quality sequence may also be a result of poor washing steps prior to the Sanger sequencing. An improper washing process can cause nucleotides and other cell components in the sequencing PCR to interfere with the Sanger sequencing. The non-confirmed proportion may also be sensitive to the sequencing protocol as this is the setup for bacterial detection in the hospital. Nonetheless, if this had a large impact, none of the qualified HPV integrations would have been confirmed by Sanger sequencing.

In some cases, low-quality sequencing results in non-confirmed integrations were linked to challenging primer design seen as multiple gel bands. However, some of the HPV integrations confirmed by Sanger sequencing also had known challenges in the primer design and contained several gel bands. HPV integrations having several gel bands that were still confirmed by sequencing directly might be either, 1) several weak gel bands but one distinct band with highly concentrated PCR products. This band may have caused the strongest signal during the sequencing, or 2) only weak gel bands where the band containing HPV integration was randomly sequenced.

## 5.3 The strengths and limitations of the study

### 5.3.1 Strengths

All the integration sites reported were manually processed and validated without previous knowledge of the diagnostic categories. In addition, templates and primer pairs for each integration breakpoint were designed following our own approach, depending on how the reads aligned to the region (discordant or junction reads). The primer pairs were also designed to be target-specific; F-primer to the human genome and the R-primer to the HPV genome. During the performance of almost every molecular method, such as the PCR reactions, visualization of the PCR products on agarose gel electrophoresis, DNA elution from gel-bands, and Sanger sequencing, a positive CaSki-control was included. If the samples were not confirmed by directly Sanger sequencing as a result of no- or low Sanger sequencing results, adjustments of the PCR reactions and gel runs were performed. Sanger sequencing was used to confirm in total 68% (21/31) of the NGS reported HPV integrations, and hot-spot regions and microhomology region were identified at the integration breakpoint.

### 5.3.2 Limits

The positive CaSki- control was not included in the DNA extraction set up as control of a proper extraction. Nevertheless, if there had been poor performance in the extraction, no PCR products would have been present. The CaSki-control was not included in T-PCR1 because of a known pipetting error.

Further, when performing the molecular methods, no negative control was included. However, 31 HPV integrations were qualified with unique breakpoints reported followed by design of unique templates and primer pairs for each integration breakpoint. Involving a negative control for each reported HPV integration would have been relatively time-consuming and not important as each integration was unique. If potential contamination was present, other integrations reported would have been Sanger sequenced. However, potential contaminants in reagents might have caused some of the weak bands in the samples with multiple gel-bands, reflecting the importance of including a negative control. Unspecific bindings seen as multi gel-bands were a concern in several samples. In addition, not all of the qualified HPV integrations were confirmed. Another limitation of the study is that HPV integrations were only confirmed if they mapped the same human chromosome and HPV type as reported from the NGS data, not whether it was at an identical chromosomal locus and HPV gene.

## 6. Conclusion and further research

The aim of this study was to validate NGS reported HPV integrations in HPV31, 33, and 45 positive samples. Additional aims were to uncover potential integrations in hot-spot regions and to identify microhomology sequences at the integration breakpoint. Although NGS technology can reveal genomic information about the HPV integrations, the validation process of reported HPV integrations was relatively time-consuming. Integrations in HPV31, 33, and 45 positive samples were validated, with mostly integrations in HPV45 positive CIN3 samples being confirmed. Some of the HPV integrations were observed in hot-spot regions and with microhomology regions at the integration breakpoint, alluding to the mechanisms responsible for integrations. Samples collected at an earlier stage or follow-up samples from the women with confirmed HPV integrations were not part of this study design but could reveal whether chromosomal integration can be used as a biomarker for predicting cancer development. The result of this study confirms previous findings and reflects the importance of determining viral integrations. The results also shows that the TaME-seq protocol can identify HPV integrations in human chromosomes and HPV breakpoint. Still, more studies are needed for other HR-HPVs besides type 16 and 18, with a larger study population and a balanced distribution of diagnostic groups. Also, longitudinal studies to the extent possible will be an important contribution to the validation of whether HPV integration may be used as a biomarker.

Future studies should also aim to exclude off-target cross-binding of primers, closer characterization of chromosomal integration sites, to uncover cancer-related genes, including genomic distance and potential impact on their function, followed by confirmation through functional studies. Further investigations of the non-confirmed HPV integrations are also important for adjusting the methods by e.g. primer design, different PCR and gel conditions, and efforts to increase the DNA concentrations prior to a potential gel extraction method. It is also important to study more closely the events of the DDR which allows the HPV virus to integrate into the human genome.

## 7. Literature list

1. Wang J, Mullighan CG, Easton J, Roberts S, Ma J, Rusch MC, et al. CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nature Methods*. 2011;8(8):652-4.
2. Burd EM. Human Papillomavirus and Cervical Cancer American Society for Microbiology Journals. 2003;16(1):1-17.
3. Martel Cd, Plummer M, Vignat J, Franceschi S. Worldwide burden of cancer attributable to HPV by site, country and HPV type. *International Journal of Cancer*. 2017;141(4):664-70.
4. National Institutes of Health (NIH) NIH Curriculum Supplement Series. Understanding Cancer [Bethesda, Md]. NIH; 2007.
5. Liu Y, Zhang C, Gao W, Wang L, Pan Y, Gao Y, et al. Genome-wide profiling of the human papillomavirus DNA integration in cervical intraepithelial neoplasia and normal cervical epithelium by HPV capture technology. *Scientific Reports*. 2016;6:35427.
6. McBride AA, Warburton A. The role of integration in oncogenic progression of HPV-associated cancers. *PLoS Pathogens*. 2017;13(4):e1006211.
7. Williams VM, Filippova M, Soto U, Duerksen-Hughes PJ. HPV-DNA integration and carcinogenesis: putative roles for inflammation and oxidative stress. *Future Virology*. 2011;6(1):45-57.
8. Vinokurova S, Wentzensen N, Kraus I, Klaes R, Driesch C, Melsheimer P, et al. Type-Dependent Integration Frequency of Human Papillomavirus Genomes in Cervical Lesions. *Cancer Research*. 2008;68(1).
9. Das P, Thomas A, Mahantshetty U, Shrivastava SK, Deodhar K, Mulherkar R. HPV Genotyping and Site of Viral Integration in Cervical Cancers in Indian Women. *PLoS One*. 2012;7(7):e41012.
10. Lagström S, Umu SU, Lepistö M, Ellonen P, Meisal R, Christiansen IK, et al. TaME-seq: An efficient sequencing approach for characterisation of HPV genomic variability and chromosomal integration. *Scientific Reports*. 2019;9(524).
11. Abreu ALP, Souza RP, Gimenes F, Consolaro MEL. A review of methods for detect human Papillomavirus infection. *Virology Journal*. 2012;9:262.
12. Hu Z, Zhu D, Wang W, Li W, Jia W, Zeng X, et al. Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism. *Nature Genetics*. 2015;47(2):158-63.
13. Gao G, Wang J, Kasperbauer JL, Tombers NM, Teng F, Gou H, et al. Whole genome sequencing reveals complexity in both HPV sequences present and HPV integrations in HPV-positive oropharyngeal squamous cell carcinomas. *BMC Cancer*. 2019;19(352).
14. World Health Organization. Cervical cancer: WHO; [cited 2020 10.09.2020]. Available from: [https://www.who.int/health-topics/cervical-cancer#tab=tab\\_1](https://www.who.int/health-topics/cervical-cancer#tab=tab_1).
15. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer*. 2015;136(5):e359-86.
16. Stanley M. Pathology and epidemiology of HPV infection in females. *Gynecologic Oncology*. 2010;117(2):S5-S10.
17. Forman D, Martel Cd, Lacey CJ, Soerjomataram I, Lortet-Tieulent J, Bruni L, et al. Global burden of human papillomavirus and related diseases. *Vaccine*. 2012;30:F12-23.
18. Kreftregisteret. Livmorhalskreft Kreftregisteret [updated 22.02.2021; cited 2021 22.03]. Available from: <https://www.kreftregisteret.no/Temasider/kreftformer/Livmorhalskreft/>.

19. Jonassen C, Christiansen IK, Tropé A. Livmorhalskreft og Human papillomavirus: Kreftregisteret; [updated 2019; cited 2020 11.08]. Available from: <https://www.kreftregisteret.no/screening/livmorhalsprogrammet/Helsepersonell/Faglig-Radgivningsgruppe/kvalitetsmanual2/2.-livmorhalskreft-og-human-papillomavirus>.
20. Cutts F, Franceschi S, Goldie S, Castellsague X, Sanjose Sd, Garnett G, et al. Human papillomavirus and HPV vaccines: a review: World Health Organization (WHO); [cited 2021 28.01]. Available from: <https://www.who.int/bulletin/volumes/85/9/06-038414/en/>.
21. Baseman JG, Koutsky LA. The epidemiology of human papillomavirus infections. *Journal of Clinical Virology*. 2005;32:16-24.
22. Arbyn M, Castellsagué X, Sanjosé Sd, Bruni L, Saraiya M, Bray F, et al. Worldwide burden of cervical cancer in 2008. *Annals of Oncology*. 2011;22(12):2675-86.
23. Smith JS, Herrero R, Bosett C, Muñoz N, Bosch FX, Eluf-Neto J, et al. Herpes Simplex Virus-2 as a Human Papillomavirus Cofactor in the Etiology of Invasive Cervical Cancer. *Journal of the National Cancer Institute*. 2002;94(21):1604-13.
24. Bosch FX, Qiao Y-L, Castellsagué X. The epidemiology of human papillomavirus infection and its association with cervical cancer. *International Journal of Gynecology and Obstetrics*. 2006;94:S8-S21.
25. Stelzle D, Tanaka LF, Lee KK, Khalil AI, Baussano I, Shah ASV, et al. Estimates of the global burden of cervical cancer associated with HIV. *The Lancet Global Health*. 2021;9(2):e161-e9.
26. Koshiol J, Lindsay L, Pimenta JM, Poole C, Jenkins D, Smith JS. Persistent Human Papillomavirus Infection and Cervical Neoplasia: A Systematic Review and Meta-Analysis. *American Journal of Epidemiology*. 2008;168(2):123-37.
27. World Health Organization (WHO). Human papillomavirus (HPV) and cervical cancer: WHO; 2020 [cited 2020 25.08]. Available from: [https://www.who.int/news-room/fact-sheets/detail/human-papillomavirus-\(hpv\)-and-cervical-cancer](https://www.who.int/news-room/fact-sheets/detail/human-papillomavirus-(hpv)-and-cervical-cancer).
28. Schiffman M, Wentzensen N. Human Papillomavirus Infection and the Multistage Carcinogenesis of Cervical Cancer. *Cancer Epidemiology, Biomarkers & Prevention*. 2013;22(4):553-60.
29. Sellors JW, Sankaranarayanan R. Colposcopy and Treatment of Cervical Intraepithelial Neoplasia: A Beginners' Manual International Agency for Research on Cancer, ; 2003.
30. Doorbar J, Quint W, Banks L, Bravo IG, Stoler M, Broker TR, et al. The Biology and Life-Cycle of Human Papillomaviruses. *Vaccine*. 2012;30:F55-F70.
31. Schiffman M, Castle PE, Jeronimo J, Rodriguez AC, Wacholder S. Human papillomavirus and cervical cancer. *The Lancet* 2007;370(9590):890-907.
32. Wang X, Huang X, Zhang Y. Involvement of Human Papillomaviruses in Cervical Cancer. *Frontiers in Microbiology*. 2018;9:2896.
33. Jung EJ, Byun JM, Kim YN, Lee KB, Sung MS, Kim KT, et al. Cervical Adenocarcinoma Has a Poorer Prognosis and a Higher Propensity for Distant Recurrence Than Squamous Cell Carcinoma. *Gynecological Cancer*. 2017;27(6):1228-36.
34. Tjalma WAA, Depuydt CE. Don't Forget HPV-45 in Cervical Cancer Screening. *American Journal of Clinical Pathology*. 2012;137(1):161-3.
35. Bulk S, Berkhof, Bulkman NWJ, Zielinski GD, Rozendaal L, Kemenade FJv, et al. Preferential risk of HPV16 for squamous cell carcinoma and of HPV18 for adenocarcinoma of the cervix compared to women with normal cytology in The Netherlands. *British Journal of Cancer*. 2006;94(1):171-5.
36. Bengtsson E, Malm P. Screening for Cervical Cancer Using Automated Analysis of PAP-Smears. *Hindawi*. 2014;2014:842037.

37. Institute for Quality and Efficiency in Health Care (IQWIG). Cervical cancer: Human papillomaviruses (HPV). Cologne, Germany: (IQWiG); 2006.
38. Ho GYF, Bierman R, Beardsley L, Chang CJ, Burk RD. Natural History of Cervicovaginal Papillomavirus Infection in Young Women. *The New England Journal of Medicine*. 1998;338:423-8.
39. Schiffman M, Wentzensen N, Wacholder S, Kinney W, Gage JC, Castle PE. Human Papillomavirus Testing in the Prevention of Cervical Cancer *Journal of the National Cancer Institute*. 2011;103(5):368-83.
40. Schmitz M, Driesch C, Jansen L, Runnebaum IB, Dürst M. Non-Random Integration of the HPV Genome in Cervical Cancer. *PLoS One*. 2012;7(6):e39632.
41. Medical Xpress. 92% of HPV-caused cancers could be prevented by vaccine: health authority: Medical Press; 2019 [cited 2020 17.12]. Available from: <https://medicalxpress.com/news/2019-08-hpv-caused-cancers-vaccine-health-authority.html>.
42. Oyervides-Muñoz MA, Pérez-Maya AA, Rodríguez-Gutiérrez HF, Gómez-Macias GS, Fajardo-Ramírez OR, Treviño V, et al. Understanding the HPV integration and its progression to cervical cancer. *Infection, Genetics and Evolution*. 2018;61:134-44.
43. Stanley MA. Epithelial Cell Responses to Infection with Human Papillomavirus. *Clinical Microbiology Reviews*. 2012;25(2):215-22.
44. Doorslaer KV, Qina Tan SX, Bandaru S, Gopalan V, Mohamoud Y, Huyen Y, et al. The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. *Nucleic Acids Research*. 2013;41(D1):D571-8.
45. Doorslaer KV, Li Z, Xirasagar S, Maes P, Kaminsky D, Liou D, et al. The Papillomavirus Episteme: a major update to the papillomavirus sequence database. *Nucleic Acids Research*. 2017;45(D1):D499-D506.
46. Ajila V, Shetty H, Babu S, Shetty V, Hegde S. Human Papilloma Virus Associated Squamous Cell Carcinoma of the Head and Neck. *Journal of Sexually Transmitted Diseases*. 2015;2015:791024.
47. Raybould R, Fiander A, Hibbitts S. Human Papillomavirus Integration and its Role in Cervical Malignant Progression *The Open Clinical Cancer Journal*. 2011;5:1-7.
48. Bernard H-U, Burk RD, Chen Z, Doorslaer Kv, Hausen Hz, Villiers E-Md. Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology*. 2010;401(1):70-9.
49. Meisal R, Rounge TB, Christiansen IK, Eieland AK, Worren MM, Molden TF, et al. HPV Genotyping of Modified General Primer-Amplicons Is More Analytically Sensitive and Specific by Sequencing than by Hybridization. *PLoS ONE*. 2016;12(1):e0169074.
50. Doorslaer KV, Burk RD. Evolution of Human Papillomavirus Carcinogenicity. *Advances in Virus Research*. 2010;77:41-62.
51. Crosbie EJ, Einstein MH, Franceschi S, Kitchener HC. Human papillomavirus and cervical cancer. *The Lancet*. 2013;382(9895):889-99.
52. Egawa N, Doorbar J. The low-risk papillomaviruses. *Virus Research*. 2017;231:119-27.
53. Villiers E-M, Fauquet C, Broker TR, Bernard H-U, Hausen H. Classification of papillomaviruses. *Virology*. 2004;324(1):17-27.
54. Baum D. Reading a Phylogenetic Tree: The Meaning of Monophyletic Groups. *Nature Education* 2008;1(1):190.
55. Harald zur Hausen. Papillomaviruses and cancer: from basic studies to clinical application. *Nature Reviews Cancer*. 2002;2:342-50.
56. Strati K. Changing Stem Cell Dynamics during Papillomavirus Infection: Potential Roles for Cellular Plasticity in the Viral Lifecycle and Disease. *Viruses*. 2017;9(8):221.

57. Huttner D, Hickson ID. Brenner's Encyclopedia of Genetics 2013 [cited 2020 08.11]. Available from: <https://www.sciencedirect.com/science/article/pii/B9780123749840006872>.
58. Venuti A, Paolini F, Nasir L, Corteggio A, Roperto S, Campo MS, et al. Papillomavirus E5: the smallest oncoprotein with many functions. *Molecular Cancer*. 2011;10:140.
59. LaCour DE. Human papillomavirus in infants: transmission, prevalence, and persistence. *Journal of Pediatric and Adolescent Gynecology*. 2012;25(2):93-7.
60. Bravo IG, Félez-Sánchez M. Papillomaviruses. *Evolution, Medicine & Public Health*. 2015;2015(1):32-51.
61. Pett M, Coleman N. Integration of high-risk human papillomavirus: a key event in cervical carcinogenesis? *The Journal of Pathology*. 2007;212(4):356-67.
62. Hanahan D, Weinberg RA. The Hallmarks of Cancer. *Cell*. 2000;100(1):57-70.
63. Hanahan D, A. Weinberg R. Hallmarks of Cancer: The Next Generation. *Cell*. 2011;144(5):646-74.
64. Mesri EA, Feitelson M, Munger K. Human viral oncogenesis: A cancer hallmarks analysis. *Cell Host & Microbe* 2014;15(3):266-82.
65. Toufektchan E, Toledo F. The Guardian of the Genome Revisited: p53 Downregulates Genes Required for Telomere Maintenance, DNA Repair, and Centromere Structure. *Cancers*. 2018;10(5):135.
66. Mantovani F, Banks L. The Human Papillomavirus E6 protein and its contribution to malignant progression. *Oncogene*. 2001;20(54):7874-87.
67. Takeda DY, Dutta A. DNA replication and progression through S phase. *Oncogene*. 2005;24(17):2827-43.
68. Talluri S, Dick FA. Regulation of transcription and chromatin structure by pRB. *Cell Cycle*. 2012;11(17):3189-98.
69. Giacinti C, Giordano A. RB and cell cycle progression. *Oncogene*. 2006;25(38):5220-7.
70. Senapati R, Senapati NN, Dwibedi B. Molecular mechanisms of HPV mediated neoplastic progression. *Infectious Agents and Cancer* 2016;11:59.
71. Zhang L, Richards A, Khalil A, Wogram E, Ma H, Young RA, et al. SARS-CoV-2 RNA reverse-transcribed and integrated into the human genome. *bioRxiv*. 2020;2020.12.12.422516.
72. Huang J, Qian Z, Gong Y, Wang Y, Guan Y, Han Y, et al. Comprehensive genomic variation profiling of cervical intraepithelial neoplasia and cervical cancer identifies potential targets for cervical cancer early warning. *Journal of Medical Genetics*. 2019;56(3):186-94.
73. Pizzino G, Irrera N, Cucinotta M, Pallio G, Mannino F, Arcoraci V, et al. Oxidative Stress: Harms and Benefits for Human Health. *Oxidative Medicine and Cellular Longevity*. 2017;2017:8416763.
74. Akagi K, Li J, Broutian TR, Padilla-Nash H, Xiao W, Jiang B, et al. Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. *Genome Research*. 2014;24(2):185-99.
75. Christiansen IK, Sandve GK, Schmitz M, Dürst M, Hovig E. Transcriptionally Active Regions Are the Preferred Targets for Chromosomal HPV Integration in Cervical Carcinogenesis. *PLoS One*. 2015;10(3):e0119566.
76. 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(5):5852-64.
77. Thorland EC, Myers SL, Gostout BS, Smith DI. Common fragile sites are preferential targets for HPV16 integrations in cervical tumors. *Oncogene*. 2003;22(8):1225-37.

78. Kraus I, Driesch C, Vinokurova S, Hovig E, Schneider A, Doeberitz MvK, et al. The Majority of Viral-Cellular Fusion Transcripts in Cervical Carcinomas Cotranscribe Cellular Sequences of Known or Predicted Genes. *Cancer Research*. 2008;68(7):2514-22.
79. Wentzensen N, Ridder R, Klaes R, Vinokurova S, Schaefer U, Doeberitz MvK. Characterization of viral-cellular fusion transcripts in a large series of HPV16 and 18 positive anogenital lesions. *Oncogene*. 2002;21:419-26.
80. Dall KL, Scarpini CG, Roberts I, Winder DM, Stanley MA, Muralidhar B, et al. Characterization of Naturally Occurring HPV16 Integration Sites Isolated from Cervical Keratinocytes under Noncompetitive Conditions. *Cancer Research*. 2008;68(20).
81. Bodelon C, Untereiner ME, Machiela MJ, Vinokurova S, Wentzensen N. Genomic characterization of viral integration sites in HPV-related cancers. *International Journal of Cancer*. 2016;139(9):2001-11.
82. The Cancer Genome Atlas Research Network. Integrated genomic and molecular characterization of cervical cancer. *Nature*. 2017;543(7645):378-84.
83. Sahab Z, Sudarshan SR, Liu X, Zhang Y, Kirilyuk A, Christopher M. Kamonjoh, et al. Quantitative Measurement of Human Papillomavirus Type 16 E5 Oncoprotein Levels in Epithelial Cell Lines by Mass Spectrometry. *Journal of Virology*. 2012;86(17):9465-73.
84. Xiao C-Y, Fu B-B, Li Z-Y, Mushtaq G, Kamal MA, Li J-H, et al. Observations on the expression of human papillomavirus major capsid protein in HeLa cells. *Cancer Cell International*. 2015;15:53.
85. Ontology search (OLS). CaSki [cited 2021 29.04]. Available from: [https://www.ebi.ac.uk/ols/ontologies/efo/terms?short\\_form=EFO\\_0006549](https://www.ebi.ac.uk/ols/ontologies/efo/terms?short_form=EFO_0006549).
86. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*. 2012;13(1):134.
87. Snijders PJF, Hogewoning CJA, Hesselink AT, Berkhof J, Voorhorst FJ, Bleeker MCG, et al. 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. *International Journal of Cancer*. 2006;119(5):1102-7.
88. Vojtechova Z, Sabol I, Salakova M, Turek L, Grega M, Smahelova J, et al. Analysis of the integration of human papillomaviruses in head and neck tumours in relation to patients' prognosis. *International Journal of Cancer*. 2016;138(2):386-95.
89. Totomoch-Serra A, Marquez MF, Cervantes-Barragán DE. Sanger sequencing as a first-line approach for molecular diagnosis of Andersen-Tawil syndrome. *F1000 Research*. 2017;6:1016.
90. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. *Genomics*. 2016;107(1):1-8.
91. Cario RD, Kura A, Suraci S, Magi A, Volta A, Marcucci R, et al. Sanger Validation of High-Throughput Sequencing in Genetic Diagnosis: Still the Best Practice? *Frontiers in Genetics*. 2020;11:592588.
92. Anderson MW, Schrijver I. Next generation DNA sequencing and the future of genomic medicine. *Genes (Basel)*. 2010;1(1):38-69.
93. Slatko BE, Gardner AF, Ausubel FM. Overview of Next-Generation Sequencing Technologies. *Current Protocols in Molecular Biology*. 2018;122(1):e59.
94. Kulkarni P, Frommolt P. Challenges in the Setup of Large-scale Next-Generation Sequencing Analysis Workflows. *Computational and Structural Biotechnology Journal*. 2017;15:471-7.
95. Petrosky E, Bocchini JA, Hariri S, Chesson H, Curtis CR, Saraiya M, et al. Use of 9-Valent Human Papillomavirus (HPV) Vaccine: Updated HPV Vaccination Recommendations

- of the Advisory Committee on Immunization Practices. Morbidity and Mortality Weekly Report. 2015;64(11):300-4.
96. Beachler DC, Kreimer AR, Schiffman M, Herrero R, Wacholder S, Rodriguez AC, et al. Multisite HPV16/18 Vaccine Efficacy Against Cervical, Anal, and Oral HPV Infection. Journal of the National Cancer Institute. 2015;108(1).
97. Cheng L, Wang Y, Du J. Human Papillomavirus Vaccines: An Updated Review. Vaccines. 2020;8(3).
98. Joura EA, Giuliano AR, Iversen O-E, Bouchard C, Mao C, Mehlsen J, et al. A 9-Valent HPV Vaccine against Infection and Intraepithelial Neoplasia in Women. The New England Journal of Medicine. 2015;372(8):711-23.
99. Bosch FX, R.Broker T, Forman D, Moscicki A-B, L.Gillisone M, Doorbar J, et al. Comprehensive Control of Human Papillomavirus Infections and Related Diseases. Vaccine. 2013;31:H1-H31.
100. Malagón T, Drolet M, Boily M-C, Franco EL, Jit M, Brisson J, et al. Cross-protective efficacy of two human papillomavirus vaccines: a systematic review and meta-analysis. The Lancet Infectious Diseases. 2012;12(10):781-9.
101. Harari A, Chen Z, Rodríguez AC, Hildesheim A, Porras C, Herrero R, et al. Cross-protection of the Bivalent Human Papillomavirus (HPV) Vaccine Against Variants of Genetically Related High-Risk HPV Infections. The Journal of Infectious Diseases. 2016;213(6):939-47.
102. Verdenius I, Groner JA, Harper DM. Cross protection against HPV might prevent type replacement. The Lancet Infectious Diseases. 2013;13(3):195.
103. Draper E, Bissett SL, Howell-Jones R, Waight P, Soldan K, Mark Jit NA, et al. A Randomized, Observer-Blinded Immunogenicity Trial of Cervarix® and Gardasil® Human Papillomavirus Vaccines in 12-15 Year Old Girls. PLoS ONE. 2013;8(5):e61825.
104. Kreftregisteret. HPV-relatert forskning [updated 20.01.2021; cited 2021 03.05.2021]. Available from: <https://www.kreftregisteret.no/Forskning/Om-forskningen/HPV/>.
105. Folkehelseinstituttet (FHI). Vaksine mot HPV (humant papillomavirus): FHI; 2018 [cited 2020 16.12]. Available from: <https://www.fhi.no/sv/vaksine/barnevaksinasjonsprogrammet/vaksinene-i-barnevaksinasjonsprogrammet/vaksine-mot-hpv-humant-papillomavirus/>.
106. Folkehelseinstituttet (FHI). Overvåking av HPV i Meldingssystemet for smittsomme sykdommer (MSIS) : FHI; 2016 [cited 2020 16.12]. Available from: <https://www.fhi.no/hn/helseregistre-og-registre/msis/msis-biobank/hpv-i-msis/>.
107. Mokobi F. Papanicolaou Staining (Pap stain) for Pap Smear / Pap Test: Microbe Notes; 2020 [updated 09.09.2020; cited 2020 16.12]. Available from: <https://microbenotes.com/papanicolaou-staining/#basic-procedure-for-papanicolaou-staining-pap-stain>.
108. Karnon J, Peters J, Platt J, Chilcott J, McGoogan E, Brewer N. Liquid-based cytology in cervical screening: an updated rapid and systematic review and economic analysis. Health Technology Assessment programme. 2003.
109. 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.
110. Silver MI, Rositch AF, Burke AE, Chang K, Viscidi R, Gravitt PE. Patient Concerns About Human Papillomavirus Testing and 5-Year Intervals in Routine Cervical Cancer Screening. Obstetrics & Gynecology. 2015;125(2):317-29.
111. Kreftregisteret. HPV i primærscreening [updated 21.05.2019; cited 2020 23.09]. Available from:

<https://www.kreftregisteret.no/screening/livmorhalsprogrammet/Helsepersonell/screeningstrategi-og-nasjonale-retningslinjer/HPV-i-primarscreening/>.

112. Engesæter B, Hidle BvD, Hansen M, Moltu P, Staby KM, Borchgrevink-Persen S, et al. Quality assurance of human papillomavirus (HPV) testing in the implementation of HPV primary screening in Norway: an inter-laboratory reproducibility study. *BMC Infectious Diseases*. 2016;16(1):698.
113. Buckley C, Butler E, Fox H. Cervical intraepithelial neoplasia. *Journal of Clinical Pathology* 1982;35:1-13.
114. Graue R, Lönnberg S, Skare GB, Sæther SMM, Bjørge T. Atypical glandular lesions of the cervix and risk of cervical cancer. *Acta Obstetricia et Gynecologica Scandinavica*. 2019;99(5):582-90.
115. Nygård M AT, J B, B H, B H, O-E I, Juvkam K-H, et al. HPV-test i primærscreening mot livmorhalskreft. Oslo; 2013.
116. Cooper DB, Carugno J, Menefee GW. Conization Of Cervix. StatPearls [Internet]. 2020.
117. D'Alessandro P, Arduino B, Borgo M, Saccone G, Venturella R, Cello AD, et al. Loop Electrosurgical Excision Procedure versus Cryotherapy in the Treatment of Cervical Intraepithelial Neoplasia: A Systematic Review and Meta-Analysis of Randomized Controlled Trials. *Gynecology and Minimally Invasive Therapy*. 2018;7(4):145-51.
118. Kerfeld CA, Scott KM. Using BLAST to Teach “E-value-tionary” Concepts. *PLoS Biology*. 2011;9(2):e1001014.
119. Newell PD, Fricker AD, Roco CA, Chandransu P, Merkel SM. A Small-Group Activity Introducing the Use and Interpretation of BLAST. *Journal of Microbiology & Biology Education*. 2013;14(2):238-43.
120. Ebbert MTW, Wadsworth ME, Staley LA, Kaitlyn L. Hoyt, Pickett B, Miller J, et al. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics*. 2016;17:239.
121. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*. 2009;25(16):2078-9.
122. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3--new capabilities and interfaces. *Nucleic Acids Res*. 2012;40(15):e115.
123. Dieffenbach CW, Lowe TM, Dveksler GS. General concepts for PCR primer design. *Genome Research*. 1993;3:S30-7.
124. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JAM. Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Research*. 2007;35(2):W71-W4.
125. Stothard P. The Sequence Manipulation Suite: PCR-products: Bioinformatics; 2000 [cited 2020 03.01]. Available from: [http://www.bioinformatics.org/sms2/pcr\\_products.html](http://www.bioinformatics.org/sms2/pcr_products.html).
126. Stothard P. The Sequence Manipulation Suite: Primer-map: Bioinformatics; 2000 [cited 2020 03.01]. Available from: [https://www.bioinformatics.org/sms2/primer\\_map.html](https://www.bioinformatics.org/sms2/primer_map.html).
127. Stothard P. The Sequence Manipulation Suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques*. 2000;28:1102-4.
128. Biomérieux. NUCLISENS® EASYMAG®: Biomereieux; [cited 2020 10.12]. Available from: <https://www.biomerieux-usa.com/clinical/nuclisens-easymag>.
129. Biomérieux. NucliSENS® Lysis Buffer.
130. Ali N, Rampazzo RdCP, Costa ADT, Krieger MA. Current Nucleic Acid Extraction Methods and Their Implications to Point-of-Care Diagnostics. *BioMed Research International*. 2017;2017:9306564.

131. NUCLISENS® EASYMAG®: Nucisens easyMAG; [cited 2020 10.12]. Available from: <https://www.biomerieux-nordic.com/product/nuclisensr-easymagr>.
132. Anchordoquy TJ, Molina MC. Preservation of DNA. Mary Ann Liebert. 2008;5(4).
133. Life technologies. Qubit® 3.0 Fluorometer. 2014.
134. Applied Biosystems. Qubit® dsDNA HS Assay Kits. 2015.
135. Thermo Fisher Scientific. Phusion High-Fidelity PCR Master Mix. 2018.
136. Thermo Fisher Scientific. PCR Fidelity Calculator [cited 2021 12.03]. Available from: <https://www.thermofisher.com/no/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-scientific-web-tools/pcr-fidelity-calculator.html>.
137. Thermo Fisher Scientific. Phusion™ High-Fidelity DNA Polymerase [cited 2020 02.01]. Available from: <https://www.thermofisher.com/order/catalog/product/F-530XL#/F-530XL>.
138. Raybould R, Fiander A, W.G.Wilkinson G, Hibbitts S. HPV integration detection in CaSki and SiHa using detection of integrated papillomavirus sequences and restriction-site PCR. Journal of Virological Methods. 2014;206:51-4.
139. Green MR, Sambrook J. Touchdown Polymerase Chain Reaction (PCR). Cold Spring Harbor Protocols. 2018;2018(5).
140. Garibyan L, Avashia N. Research Techniques Made Simple: Polymerase Chain Reaction (PCR). Journal of Investigative Dermatology. 2014;133(3):1-4.
141. Lee PY, Costumbado J, Hsu C-Y, Kim YH. Agarose gel electrophoresis for the separation of DNA fragments. Journal of Visualized Experiments. 2012(62):3923.
142. Biotium. GelGreen® Nucleic Acid Gel Stain [cited 2021 04.05]. Available from: <https://biotium.com/product/gelgreen-nucleic-acid-gel-stain/>.
143. Promega. Wizard® SV Gel and PCR Clean-Up System. 2010.
144. Applied Biosystems. Advances in Fast PCR Contribute to a Fast Resequencing Workflow. USA2008.
145. Applied Biosystems. Hi-Di™ Formamide. 2012.
146. Pareek CS, Smoczyński R, Tretyn A. Sequencing technologies and genome sequencing. Journal of Applied Genetics. 2011;52(4):413-35.
147. Shen C-H. Diagnostic Molecular Biology [Chapter 11]: Academic Press; 2019. 277-302 p.
148. Lee LG, Connell CR, Woo SL, Cheng RD, McArdle BF, Fuller CW, et al. DNA sequencing with dye-labeled terminators and T7 DNA polymerase: effect of dyes and dNTPs on incorporation of dye-terminators and probability analysis of termination fragments. Nucleic Acids Research. 1992;20(10):2471-83.
149. Dey P. Sanger Sequencing and Next-Generation Gene Sequencing: Basic Principles and Applications in Pathology. Singapore: Springer Singapore; 2018. 227-31 p.
150. Applied Biosystems. Troubleshooting Sanger sequencing data. 2016.
151. National Human Genome Research Institute (NIH). Contig [cited 2021 20.02]. Available from: <https://www.genome.gov/genetics-glossary/Contig>.
152. Gallo G, Bibbo M, Bagella L, Zamparelli A, Sanseverino F, Giovagnoli MR, et al. Study of viral integration of HPV-16 in young patients with LSIL. Journal Clinical Pathology. 2003;56(7):532-6.
153. Ho C-M, Lee B-H, Chang S-F, Chien T-Y, Huang S-H, Yane C-C, et al. Integration of human papillomavirus correlates with high levels of viral oncogene transcripts in cervical carcinogenesis. Virus Research. 2011;161(2):124-30.
154. Anayannis NV, Schlecht NF, Ben-Dayan M, Smith RV, Belbin TJ, Ow TJ, et al. Association of an intact E2 gene with higher HPV viral load, higher viral oncogene

- expression, and improved clinical outcome in HPV16 positive head and neck squamous cell carcinoma. *PLoS ONE*. 2018;13(2):e0191581.
155. Ma M, Feng Y, Fan P, Yao X, Peng Y, Dong T, et al. Human papilloma virus E1-specific T cell immune response is associated with the prognosis of cervical cancer patients with squamous cell carcinoma. *Infectious Agents and Cancer*. 2018;13(1).
156. Vasilescu F, Ceaușu M, Tănase C, Stănculescu R, Vlădescu T, Ceaușu Z. P53, p63 and Ki-67 assessment in HPV-induced cervical neoplasia. *Romanian Journal of Morphology and Embryology*. 2009;50(3):357-61.
157. Citro S, Bellini A, Medda A, Sabatini ME, Tagliabue M, Chu F, et al. Human Papilloma Virus Increases ΔNp63α Expression in Head and Neck Squamous Cell Carcinoma. *Frontiers in Cellular and Infection Microbiology*. 2020;10:143.
158. Koneva LA, Zhang Y, Virani S, Hall PB, McHugh JB, Chepeha DB, et al. HPV Integration in HNSCC Correlates with Survival Outcomes, Immune Response Signatures, and Candidate Drivers. *Molecular Cancer Research*. 2018;16(1).
159. Bernhard MC, Zwick A, Mohr T, Gasparoni G, Khalmurzaev O, Matveev VB, et al. The HPV and p63 Status in Penile Cancer Are Linked with the Infiltration and Therapeutic Availability of Neutrophils. *Molecular Cancer Therapeutics*. 2021;20(2).
160. Coosemans A, Nik SA, Caluwaerts S, Lambin S, Verbist G, Bree RV, et al. Upregulation of Wilms' tumour gene 1 (WT1) in uterine sarcomas. *European Journal of Cancer*. 2007;43(10):1630-7.
161. Coosemans A, Moerman P, Verbista G, Maes W, Neven P, Vergote I, et al. Wilms' tumor gene 1 (WT1) in endometrial carcinoma. *Gynecologic Oncology*. 2008;111(3):502-8.
162. National Center for Biotechnology Information (NCBI). SH3TC2 SH3 domain and tetratricopeptide repeats 2 [ Homo sapiens (human) ] [updated 08.04.2021; cited 2021 08.04]. Available from: <https://www.ncbi.nlm.nih.gov/gene/79628>.
163. Azzedine H, Salih MA. SH3TC2-Related Hereditary Motor and Sensory Neuropathy 2008 [updated 11.03.2021; cited 2021 11.05]. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK1340/>.
164. National Center for Biotechnology Information (NCBI). NNSL1 NNSL like 1 [ Homo sapiens (human) ] [updated 02.03.2021; cited 2021 06.04]. Available from: <https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=DetailsSearch&Term=57224#gene-expression>.
165. Parfenov M, Pedamallu CS, Gehlenborg N, Freeman SS, Danilova L, Bristow CA, et al. Characterization of HPV and host genome interactions in primary head and neck cancers. *Proceedings of the National Academy of Sciences (Proceedings of the National Academy of Sciences of the United States of America) (PNAS)*. 2014;111(43):15544-9.
166. Poljak M, Cuschieri K, Waheed D-E-N, Baay M, Vorsters A. Impact of the COVID-19 pandemic on human papillomavirus-based testing services to support cervical cancer screening. *Acta Dermatovenerologica Alpina, Pannonica et Adriatica*. 2021;30(1):21-6.
167. Kreftregisteret. Livmorhalssjekk hjemme kan øke screening-deltakelsen 2017 [updated 28.04.2020; cited 2021 13.05]. Available from: <https://www.kreftregisteret.no/Generelt/Nyheter/hjemmetest-av-livmorhals/>.
168. The Norwegian National Research Ethics Committees. The Health Research Act 2020 [cited 2021 20.04]. Available from: <https://www.forskningssetikk.no/en/resources/the-research-ethics-library/legal-statutes-and-guidelines/the-health-research-act/>.
169. Assal N, Lin M. PCR procedures to amplify GC-rich DNA sequences of *Mycobacterium bovis*. *Microbiological Methods* 2021;181:106121.
170. National Center for Biotechnology Information (NCBI). Chimera Detection in 16S rRNA Sequences at NCBI [cited 2021 17.04]. Available from: <https://www.ncbi.nlm.nih.gov/genbank/rRNACHIMERA/>.

171. Omelina ES, Ivankin AV, Letiagina AE, Pindyurin AV. Optimized PCR conditions minimizing the formation of chimeric DNA molecules from MPRA plasmid libraries. *BMC Genomics*. 2019;20(536).
172. Kalle E, Kubista M, Rensing C. Multi-template polymerase chain reaction. *Biomolecular Detection and Quantification*. 2014;2:11-29.
173. Korbie DJ, Mattick JS. Touchdown PCR for increased specificity and sensitivity in PCR amplification. *Nature Protocols* 2008;3:1452-6.
174. Ballari RV, Martin A. Assessment of DNA degradation induced by thermal and UV radiation processing: implications for quantification of genetically modified organisms. *Food Chemistry*. 2013;141(3):2130-6.

# Appendix

Appendix 1. REK approval

Appendix 2. Data Protection Office at Ahus approval

Appendix 3. Template sequences and primer pairs

Appendix 4. Agarose gel runs

Appendix 5. Sanger sequences from confirmed HPV integrations

# Appendix 1. REK approval



Region:	Saksbehandler:	Telefon:	Vår dato:	Vår referanse:
REK sør-øst C	Anders Strand		19.08.2020	5853
Deres referanse:				

Hanne Irene Kraus Christiansen

## 5853 HPV genetisk variasjon som markør for kreftutvikling

**Forskningsansvarlig:** Akershus universitetssykehus HF

**Søker:** Hanne Irene Kraus Christiansen

### REKs vurdering

REK viser til endringsmelding mottatt 06.08.2020, for prosjekt 2017/447 «HPV genetisk variasjon som markør for kreftutvikling». Sekretariatet i REK sør-øst C har behandlet meldingen på fullmakt fra REK sør-øst C, med hjemmel i helseforskningsloven §11.

Den omsøkte endringen består i at Adina Repesa (OsloMet) inkluderes som prosjektmedarbeider. Komiteen godkjenner dette.

### Vedtak

Godkjent

Komitén har vurdert endringsmeldingen og godkjenner prosjektet slik det nå foreligger med hjemmel i helseforskningslovens § 11.

Tillatelsen er gitt under forutsetning av at prosjektet gjennomføres slik det er beskrevet i prosjektendringsmeldingen og endringsprotokoll, og de bestemmelser som følger av helseforskningsloven med forskrifter.

Vennligst oppgi vårt referansenummer i korrespondanse.

Med vennlig hilsen,  
Jacob Hølen  
Sekretariatsleder, REK sør-øst C

Anders Strand  
Rådgiver

**Klageadgang**

Du kan klage på komiteens vedtak, jf. forvaltningsloven § 28 flg. Klagen sendes til REK sør-øst C. Klagefristen er tre uker fra du mottar dette brevet. Dersom vedtaket opprettholdes av REK sør-øst C, sendes klagen videre til Den nasjonale forskningsetiske komité for medisin og helsefag (NEM) for endelig vurdering.

## Appendix 2. Data Protection Office at Ahus approval



Akershus universitetssykehus  
HF

Postadresse:  
Postboks 95  
1478 Lørenskog

Sentralbord:  
02900

Org.nr:  
NO 983 971 636 MVA  
[www.ahus.no](http://www.ahus.no)

### PERSONVERNOMBUDETS UTTALELSE

Til: Irene Kraus Christiansen  
Leder Nasjonalt referanselaboratorium for human papillomavirus (HPV)  
Mikrobiologi og smittevern  
Diagnostikk og teknologidivisjonen  
Akershus universitetssykehus HF

Kopi: Truls Leegaard  
Konstituert avdelingsleder  
Mikrobiologi og smittevern  
Diagnostikk og teknologidivisjonen  
Akershus universitetssykehus HF

Fra: Personvernombudet ved  
Akershus universitetssykehus HF

Dato: 01.02.2021

Offentlighet: Ikke unntatt offentlighet

Sak: Personvernombudets uttalelse i forbindelse med  
endringsmelding.

---

Saksnummer/  
Personvernnummer: 18/12755  
2017\_109

### Personvernombudets uttalelse i forbindelse med endringsmelding i prosjektet «HPV genetisk variasjon som markør for kreftutvikling»

Viser til innsendt endringsmelding i ovennevnte prosjekt. Det følgende er et formelt svar på  
meldingen.

Endringene består i tilførsel av Adina Repesa og Jean-Marc Costanzi som nye  
prosjektmedarbeidere.

Personvernombudet har ingen innvendinger til endringene da de ikke synes å innebære økt  
risiko for personvernet. Endringene er godkjent av REK i vedtak med referanse 5853.

Det forutsettes imidlertid at personvernet fortsatt ivaretas slik som beskrevet i opprinnelig  
uttalelse fra personvernombudet, inkludert endringsmeldinger og eventuell etterfølgende  
korrespondanse.

Med vennlig hilsen  
for personvernombudet

Hans Tangen  
Personvernrådgiver/jurist  
Akershus universitetssykehus HF

Epost: [forskning.personvern@ahus.no](mailto:forskning.personvern@ahus.no)  
Web: [www.ahus.no](http://www.ahus.no)

### Appendix 3. Template sequences and primer pairs

Supplementary table 3A: Templates and primer pairs from the 31 qualified Human papillomavirus (HPV) integrations.

Sample ID	HPV			Human		Template Human [NNN] HPV	Forward primer 5'→3'	Reverse primer 3'→5'	Product length (bp)
	Type	Break point (bp)	Gene	Break-point (bp)	Chr Locus				
1a	45	1393	E1	Chr3: 189876 913	3q28	AAGCAAATCAACTAAAATCACATTCTCGTAATATAAGCCTACTTGTATCTATAACAATTGATTCACAGCAGCTCTGAAAGTCTATCTCGATAAAGCCTATGATCATGAAGGTAACCGCGA[NNN]GCTTGTAAATAGCTCTTTAGTCTGTAAACTGCACTGCGGATCTACATTCTGCATTGTCAGTACATCCCCACCAACTCTTGTACTATGTACA	5'-ACAATTGATT CACAGCAGCT C-3'	3'-GTGGTGGG GATAGTAG TGACA-5'	152
1b	45	4358	L2	Chr3: 189955 746	3q28	CCAGAAAACAGAAATCAGATCCTACCCAGATCCTCAATCAGCAAATCTGAGTGCCAACAAGCTCTCACAGTGGGAGGACAAATAGCAATCTACCTTGTCAAAGCTCAAGAACCTTGTCAAATTGGGTTTACGTTTAAAATCAGGTACACACATAAAAACCCATTACGCCGGGACAGTAGCTCACACCTGTAACTCAGCACTTGGGAGGCCAAGGCAGTGATCACCTGAGGTAGGAGTTCGAGACTAGCCTGGCCAACATAGTGAACCCCCATCTACTAAATA[NNN]TGATAAAATTATA CAGTGGCTAGCCTTGGAAATTGGTGGCCTTGGCATTGGTACCGGCAGTGGTCTGGAGGCCGTACGGCTATGTACCCCTT	5'-AGCACTTTGG GAGGCCAA-3'	3'-TACGGCCT CCAGAACCC ACT-5'	181
2a	31	1235	E1	Chr7: 266405 55	7p15.2	CCATCCACAGCAGGGCACACGGGCTGCCAGTGGCTGCCGGCACACAGGTACATGCACATAGTGCTCTGCACGCCATAGCTTGTCAAAGAGTGTAGGAAGGCATTGTTTCACTTTAAAAGGACCCATGCTGTGACATTGAGGTCAACCCAAATAGTCAGGCCACACGAATGCTACATCTACAAGTCCATGTATGCAATGTACATTAGATATAGTTGTGTT[NNN]TTAAGTGTATTAGTAGTTGTGATTATAATTAGTCACGGTTAAAGCTATATGCATAGAAAATAACAGTAAACAGCAA	5'-AAAAGGACCC ATGCTGTGAC-3'	3'-GCATATAG CTTTAACC GTGGA-5'	174
3a	31	6169	L1	Chr13: 732101 24	13q22.1	CCTTATATCACTGAGTTCTAAGTCTTATTACTGCCAGAATTATATCAGAGAATCTGACTCTTCACTTGTGCAAACACTTCAGGGTGGACTCTTGTGAGGTAGGGAAATTACTGAGAACCTTACCATTTGATCCTCTATAATAGCACGGTGACCCCTGGAGACACCAATCCAATGATTCAACACCAACTGGCGCTTTTTCTGAACCTACAGTGCTCACAGAGATCACGTCTAATGAGAGAGAGAGGGCAGTCAAAATAAATGGGAACAATAGGGACTTGTGAGGCCAGAGGAGGAATCTGCTTCAGAGGGGATGCCACGATCTGGGAGGCTTGGTGAACAGGGCC[NNN]ACCATTAAAGATAATCTGGATATTACATATAAGAATTACAAATGTCAAAGGAACATTACTTTAGTGTCTGTAAAGCAGTAAATCCATAGCTC CAAAGCCTGTATCAACCATAACCATACCCATCTGTATAACTGAATTTTAATTCTAATGGAGGACAACTACAGGGGTAATAGCATTGTACTACAAGGACTACCTTACCCCAATGCTCTCCAATAGGTGGTTGCAACCAAGTAA	5'-CACGATCTGG GGAGGCTTT-3'	3'-TGGTTGAT ACAGGCTT TGGAG-5'	154
4a	45	2624	E1	Chr1: 885932 3	1p36.2 3	TGCACAGTCAGGGAGCTGGGGAGCCCTTGTGGACATTCTGTTATTGTTAATGTCAGGATTAAACCTGAGAAGCAGAACCAAATAAGGAGAGGTAAAGTTAGCCAAGCATGTCGGCTAGCTATGGTCCCAGCTACTCAGCAGGCTGAGGCCAGAGGATTGCTTGAGCCCCAGAGGTCAAGGCTGCCATAAGCCAAAATCGCACCAC TGCACTCCAGTCTGGGTGACACAGTGAAGAGGCCTTCTCAAAGTAATAACTGGCTCTGATTGTGGGGCTGGATAAGC[NNN]ATTCAACAGGGCTGTCATTATTCTATATAC TGGATTACCATTTTATCAAATGGAAATGCATGTGAAATGTTAGGATGTTAATAGGATTGGAGGGACATTAAATATGGCAACCAATGCTCTCCAATAGGTGGTTGCTATAACTTATAGGATTACCATCTAA	5'-AGAGGTCAAG GCTGCCATAA-3'	3'-GGGTGACG GTATTAC ATTTCCA-5'	204
5a	45	5168	L2	Chr17: 27329 570	17q11.1	TAAC TGGGTTGTGCGCCGCATCAGACATTGTTCATCTGTCTTAAGAGCCTGAGGGTGGGAGA TGTTTGATGACAAATAAGGGTTGAGGGGGTCAGGGGTGCCAGAAGGACACAGAGGCCACACGGGAAGACAGAACGCTTGGATGGGGCTGCTTCTGCCCTGGGATGAGGAACAGGTGAGGGGC GGGTGAGCAGGTGTCGAGGGAGGGGTGACTTGAGGCTGACAGTGCAC[NNN]AAATAGGGGTA	5'-CTGAGGGTGG	3'-GCAATGGG	240

						GGGTACATTTACCATGATATAAGCCCCATTGCTGCTACAGAGGAAATTGAATTGCAGCCTTA ATTAGTGCTAC	TGGAGATGTT-3'	GCTTATATC ATGGT-5'	
6a	45	2624	E1	Chr5: 149029 187	5q32	CAAACCCACCTAAGTAGAGATGAAGAACAGGTCTGGTGTAGGTAGGATGGGCCACCATGCC ATGCCTCCAGGTGACCCAGATCAGTGGCCTTCTAGGAGCCTGGCATTCAAGACCTCAGTCA TCATTAAGGAACAGGGCAGAATTATCCTTCACCCATCATCTTAAGAGCTGTAGGGGCCATACTCA TCTTGTGTTAACCCATGCACTTACAGTTGGGAACCAAGGTATACTGTGAAACTCAGAGTCCGAGAA GAAAACCTGAAACACAGGCTATGTCAAGAGGTATGGGTACTGTAGGGGGTGGGGGGTGG GGACAGTGGGGTCTGTAGAATCTGAAGATCAGAGAGATGCAGCTCCAAATCCTCTCAGCG AGTTACAGATGGAGGCCAGGACACAACCTCCAGCACAGAACACGTGGAGCACATGGTCTGCCT CCCACCTCTAAATGTATCACATCCTCCTTGCCTCCTCATTCTATTATG[NNN]TTTCAATT TTATCATTATTCATATACTGGATTACCAATTATCAAATGGAAATGCATGTGAAATGTAAAT ACCGTCACCTACTTCTAAATATGGCATTATTATCTTGTGCTGGATCAATATTGGATGTTAAT AGGATTGGAGGACATTAGCTGTAAATAATGGTTATGCTTCTGTCTATACTTATAGGATTACCA TCTAAATGCATTCTCATATAATTCAAATATGTCAAACACGTGTGTGGCATCATCCAAACAT GGCTACCTTAGTATCTGCTAACGGTTCTAACCAAAATGGCTGTTGAATTACAAAT	5'- GAGATGCAGC TCCCAAATCC- 3'	3'- CGGTATT ACATTCC ACATGCA- 5'	228
6b	45	2888	E2	Chr5: 148950 210	5q32	GTTCCTTCTTACTTCCTTGGCATCCCTCTCAGGGCAGCCAGTTCTCACTGCTCCAGAACCTCCTTC TTGGAGCTCTGGCTGTACTGGCTGGGGAGCCGTCTGAAGAACCCAGGTAATG[NNN]A CTTACAGTGTGGAAAATGCAAACTATTACAGCAAGGGAACATGGTATTACCAAACGTGAGCC ACCAGGTGGTGCCTCTAGTAACATTCAAAAG	5'- ACTTCCTTGGC ATCCCTCTC-3'	3'- CTGGTGGC TCAGTTG GTAA-5'	188
6c	45	3390	E2	Chr3: 116874 819	3q13.3 1	TTAGTTAAAAGTAAAATAACTCATTCTACAATCTTAGCTAGTTAATGGCAAATTGGAAATT CAAACATGTTTTTATTTCTTCTACAAGCCCTGGCTTGTCTACATATAACATTAGGGT CAGAAAATAGGAACCAACTCTGGTATTTCTCAGCATACTATGATTATGCTGAAATAGAACAGC TGGAAAGTCAAACAGAAGATGATGAAGCTGCCTAGAAGTAAACACAGGAAGACATAGCTACTA CTCTTGAGCTGGATGAGAGATGGTAT[NNN]AGTACCAAGTGAACACGGTATCGCTACTCAGA TTGTTAGACAGCTACAACACGCCCTCACGTGACCCCCAAACCGCATCCGGGGACCCCAA ACCCACATCCAGACGCCGGCTACTAACGACCTAGACAGTGTGACTCACAGAGCAGCACCAC GGACGTGTCACACCCACGTGACAAACCCGCTCTGTGTTCAAGTACAAGTAACAAACAAAAGAA GGAAAGTGTGTAAGTGGTAACACTACGCCCTATAATACACTTAAAGGTGACAAAAACAGTTGAA ATGTTAAGATATAGGTAC	5'- ACAGAACATG ATGAAGCTGC C-3'	3'- CGTCTGGA TGTGGGT TTG-5'	195
7a	45	892	E7	Chr6: 138481 138	6q24.1	AACTATATCGATGTTCTTCTTCCAGTATTGCCAAGCCAAACAATTATGAAGGGT AGGTAATAATAATCCCTCTATTAAATGCACTTGTGTTGATTTCATTAAGAAAATA[NNN]AT GTTTATAGCTTATGTACAAAAACAGCCATTACACCCGTTCCCTCCCGTGGTACCTCTGG ATCCGCCATTGTAGATTATTGGTTAGTTGCA	5'- TTTCCAGTATT GCCAAGCCC-3'	3'- AGGAAACG GGGTGAA TGG-5'	151
8a	45	3669	E2	ChrX: 114942 562	Xq23	ACCTGGGGAAAGGGCAGCTGGTCACAGCTTCAGCAGACTAAACATTCCGCCACCTCT GAAGAGAGCAGTGGATCTCCAGCACAGCACACTGAGCTCTGCTAAGGGACAGACTGCCCTCTCA AGTAGGTCCCTGAACCCAGTGCCTCCAGACTGGGAGACACCTCCGGCATGCATCAACAGACAC CTCATAGAGGAGAGCTCCAGCTGGTGGGTGCCCTTAGGACAAAGCTCCAGAGGAAGAAC GTTGAAATGTT[NNN]AACAGTTGAAATGTTAAGATATAGGCTACGCAAATATGCAGACCATT ACTCAGAAATATCCTCCACCTGGCATTGGACAGGTTGTAACAAAACACTGGTATATTAGTGA ACATATAATAGTGGAGGTACAAAGAAATACCTTTGGATGTTACTATTCTAACAGTGTACA AACTCGGTGGGATACATGA	5'- CGGCATGCAT CAACAGACA- 3'	3'- CATGTATC CCACCGAG ATTGT-5'	297

9a	45	4865	L2	Chr15: 582800 64	5q21.3	<b>GTTGTTACTTTATCTGGAGAAATTTCGATTTGTTCCCTCAGGAACGGGGGTGCTACCAAGAA TATTTCAGCCCCCTTAAAGAGCTCTAGCTCAATGATGGATGACTATCCGATTAGCTCTG[NNN]T AGTAGTACCCCCCTCCCTACTGTGCGGGGGTAGCGGGGCCCCGCGCTGTATAGTAGGGCTAATCA ACAGGGCCGTGTGTCACCGTACGGTTTAAC</b>	5'- TGCTACCAAG AATATTCAGC CC-3'	3'- CACGGACC TGTTGATTA GCC-5'	152
10a	45	6852	L1	Chr11: 102911 491	11q22. 2	<b>CTATGAATCTAAAGTTCTTTGAACAAAACTTGCATTGTTAGAGTGATGCTGTG GACTGAAAATTACTTGGTAAATTAAAGTAGCTCTAAAGAGTGTGGTATGCCAATAC[NNN] TACAAGTTAGGGATACATATCGTTGTGCAATCAGTGCTTACCTGTC AAAGGATACTCACCTCCAGAAAAGCAGGATCCATATGATAAAT</b>	5'- GAGTGATGCA TGTGTGACTG A-3'	3'- CTGCTTTTC TGGAGGTG TAGT-5'	160
11a	45	2127	E1	Chr13: 485168 10	13q14. 2	<b>GATGGCTGTCTCCCTATGTCGTTACATCATCCCCCATGCCCTGCATATCTGCATCT AAATTCCCCCTTTATAAGGACACCAATTATGGATTAAGGCCACCCCTAATGACCTCATCTT AACTAACACATCTGCATTGACCCATTCCAATAAGTGTGCACTGTGAGGTACTGGCTGACTT CAACACATAATGTTGCAGGGTCACAATAACCCCTAACAGCATACTTAAAGATCACGAGCAC CCACACCTCTAAAGAGGAAGTATAAAATTGTTGAAAGAGTTTACCTATTGCTACTTAGGA GGCATTTTATTGATATACTTTCAACTTGCACATTTGTTAGGTTATAAAAGATACATAAACAA GCCAATATGTCATC[NNN]ATTGACTATGGGTCTCCAATCCCCACCTTCATCTATTAGAACAT CTATATTAATCCATTGAGACATATTCAATTGGCCTTTGTGCTCTTTATAATGTC CACGACAATCTTTAAATATTGGCCTTGCAGTTACTTTAAAGTCAGCTGCATTACTGTTGC AGTCTGCAATTGGGCATATTGAAATGCCATATCACTTCATGTAAGGTATTACAAATGCC CATTGCAACATGTCGACAATCAAATTACTATGCAATACCATGTTAATAAT</b>	5'- ACATTAAGAT CACGAGCACC C-3'	3'- AAAAGTAA CTGCCAAG CCAAT-5'	322
11b	45	3893	E2?	Chr13: 484915 41	13q14. 2	<b>CCCTATGCCAATTATTGAATTCTAGAACATCACAGGAACCTTCCAAGTCTGTCCTTAATTAG TATTGGGAAAGATCCCACCTCTAGAATAAAAGTAAGGAGCATGGATCCAAGTGATAACT[NNN] AGTATGTAACATTACTATGCTATCTTAGTGTGTTATTGTC TGCAATGTCGGCTGTGCAGTCTGCTATGT</b>	5'- CTCTTCCAAG TCCTGTCCTTA A-3'	3'- GGGACATT GCAGCACAC ATA-5'	169
12a	45	1646	E1	Chr3: 160748 989	3q25.3 3	<b>TTTACATTGATACTACTTATGTTCTGATGTTATTACATCTATGCCATTGATGGCAATA TTGTATAATAGTGGCTACTGAGCAACTATTCCAGCTCGTGGCAGTGGTACAGTTGAAATC AACCATGGTGAGGAGATTCTACCATGGAAATTGCAAACACAGCAGAGCTGAGTCTTACCTC CTCTCTCTAGAGGTCAGGTGTTAACATTACTGCAACACATGGTAAATGTTAGGAGG TTGTATAAGAAGAAGAAAATAAGCAATTAGAAGTTGTTCAAGGTAGCTAGGAATAGGAGG TCACATGGAAATAATAGTTCATGATCAAGAAAAAGACTGG[NNN]GCCACATTATCTTAA TAAAGCTAATATTAAACTCCCCATTACAATCTAACACATTGGATATGGCGTATAACGTTGCTG GTTAATTAAATGTTAAGCCTCTGCTACCGTTGGATACTCAAATATAGCCATTACCAATC TGTACATGTTGTTTATCACTTTAAATTCTAACCAAATCCGAAATGACAGCCATATGTC TTAAATACTGCCAGCATTGCAAGCCTTTGTTACTGCTTGTAAAGCTCCTTGT</b>	5'- GCTAGGAATA GGAGGGTCAC A-3'	3'- ACATGTAC AGATTGGG TAATGGC-5'	219
12b	45	6852	L1	Chr3: 160749 198	3q25.3 3	<b>ATCATGTAAGCCTTGGAGCTGTCAGAGGTGACCAGATATCTCAATCCTGCTCTGAGACTAAT ATACTGTAATTGCTAAAGCTTAAAGTGCACAGGCTTACTGCAAGGTCAAGTATAACAAACAC[NNN] ]TTACAAGTTAGGGATACATATCGTTGTGCAATCAGTGCTGTACCTGTC AAAGGATACTACACCTCCAGAAAAGCAGGATCCATATGATAAAT</b>	5'- TCAATCCTGCT CCTGAGACT-3'	3'- CTGCTTTTC TGGAGGTG TAGT-5'	171
13a	45	1646	E1	Chr11: 102867 368	11q22. 2	<b>GATATGAAAATGATCCTACCTGTTGAAGAAAAAGATCTTACCCACGGTAGTGACAGCAT CAAAACTCAAATTGGGGTCACAGAGAGCTGGTCTGAATTGTCAGGATTGGCAAGCGTTG[NNN] ]CCACATTATCTAACAGCTAATATTAAACTCCCCATTACAATCTAACACATTGAGATATG GGCGTATAACGTTGCTGGTTAATTATGTTT</b>	5'- AAGATCTTATT TCCCACGGTA GT-3'	3'- ACGCCAT ATCCAATG TTAGA-5'	165

14a	45	3390	E2	Chr11: 102831 633	11q22.2	CTCTTGTGCTTTGCTTACTGAGGTTGAGGTCAAGCCTCTGCCTCCAAAGCTTGCG CATGTGTTICATATIGTTCCATTAGAATTICATTCTCCTGCCATCGTTGAAGT[NNN]AG TACCACTGACGACACGGTATCGCTACTCAGATTGTTAACAGCTACAACACGCCTCCACGTCGA CCCCAAAACCGCATCCGTGGGCACCCCAAAA	5'-CGTTCTGCCT CCAAAGCTT-3'	3'-CACGGATG CGGTTTG GG-5'	173
15a	45	2127	E1	Chr8: 860020 65	8q21.3	ACCAAGTGCCTTTTATGAGGAACCAAGTTGCCTCTAACAGGCTACACAAGATCAAATTGTGT TGGCAAGATAGGATTTCTGCATTTCTTGGCAACTTCGTTACGAAGATTATTTAGCTTC CTTCCTCTCAAGGAGAAACTGACAAGACAGGAGCCTTCCACCATCTCCTCGATGAGTTTGG TGAAATTACCAACCTTGGAAAGGAGCACACAGCAGCCCTGACTGCAGATCTCTGAGCATAAAG CAGCTTCTCCTCAGGACAGCTGGGTCTCAGCAGGCACACTGCAAAGAGAGGAGGGAGGAAGAA GCAGGAAATA[NNN]TTGTACTATGGGTCTCAATCCCCACCTTCATCTAATTAGAACATCTATA TTAACATTGAGACATATTGGCTTGTGCTTAAATGTCACACATTACAGC ACAATCTTAAATATTGGCTTGGCAGTTACTTTAAAAATGCAGCTGCATTACTGTTGAGTC TGCTAATTGGGCATATTGAAATGCCATATCACTTCATCTGAAAGTCATTATCAAATGCCATTG CACCATGTCGACAAATCAAAATTACTATCGTCAATACCATGTTGAATAATTGTCAGTCTTGTAT CCACTCAGGT	5'-TGCAAAAGAGA GGGGAGGAAG -3'	3'-AAAAGTAA CTGCCAAG CCAAT-5'	188
15b	45	5114	L2	Chr8: 859471 93	8q21.3	CTAAGAATATAATGCAACTAGAACTCCCATAATTGCTTATGAGAATGCAACATAATAGAGTGA CTTATAAAACTGGAAGTTCATAAAACATTAACATCCCTGCCATATGACCCAGCATTCTA CTCTTACGTTTACTGAAAGAGGGGGCAAAAGACTGTATACAAATGTCGTATACAAATGCT GCAGCAGCTTATTACAAATGCAAAAACAGCTTAAAGCCCAATAGCTTACACTGACAAAT GCATGAAACAATTGCGGTACATCCATAGCAAGGAATATACCTTCTCAGC [NNN]AGGAGTATGGATGGCAATATAATGTCGGGGCAGTATATATAGGTACATCCATGCGA TGTTAATGGTACTGTAACATTACTGTAAGAGGATGCAGCAGTAGAAGGCATGGTCAAGGAATAC TTGGATATGAAATGATTGTATAGTGCTAGGTGAGTGGACGCAGGAGGTGGGAAGTCTGC ATATACATCAAACAGGTCACTATCATTGAGCACTAATTAAAGGCTGCAATTCAATTCTCTG TAGCAGCAATGGGGCTATATCATGGAAAAATGACCTACCCCTATTGTTACCAACTAC GTGAAACATGGTGCCTTGACCCAATCTACTAAATCTAAC	5'-TGAACAAATT GCGGTACATC C-3'	3'-TATGCAGA CTTCCCACC TCC-5'	240
16a	33	4389	L2	Chr5: 533519 27	5q11.2	TTGTTGGCTTCTACTTCCCTGCTCCCATTCTTCCCTGAACTCCCTGTGACACTCCTT AAAAATCACTTGCCTTAAAAATTAGTCTTAAATCTTGTCTTGAATCTGCACTGTG GAATCTAACCTAACGACCACTGGCGAGATCCTGAAAGTACAGGATTGAATGCCAAAGATTAC ATCATTAAATTATAGAATTAAAGGATGTCAAACCTAGAGGGTTCTAGATATCAGGTATCAGAGAG ATGAGTTAGCTCACCTGTATTAGTCCACTACATACTACTAAAGATACTACCAAGAGACTGGAC AATTATAAATGAAAGAAGTTAATTGACTCAGTTGCTTGGCTGGGAGGCCTAGGAAATT ACAATCA[NNN]CACAGGCTCTGGTTCAAGGTGGAAGGACTGGCTATGTACCTATTGGTACTAACCC ACCTACAGCTGCAATCCCCTGCACTACGTCCTCCGGTTAC	5'-AGTTAGCTA CCTGTATTAGT CC-3'	3'-ACATAGCC AGTCCTCC ACC-5'	173
17a	31	3689	E2	Chr2: 23668 560	2p24.1	TTCCCTCCTCCCTGGACTTCTGCAAGGCCAGGTATGACTGTTCTGCTCCCTAGCAAGGTCTGC AACATGCTCTGAGCAGCATCTCCAATGTGGTCACCTGTTGACAGGGGTGCAAAGGTAGGT GCTGGGGTGGGGCAGTCATGGACTTGTCTCCCCACTCGAAAAGCTCTCCTCAGGAAAGCT CTCAAAAGCAGAGCCTGCATCTGACTAGTCTCTGCACCCCCAGTGCCTGGCATAGCTTCTGGCTC[ NNN]ACATAAAAATGCTATTGTAACCTAACATATATAAGTACATCACAAAGAGACGATTGTTA AATA	5'-CCACTCCGAA AAGCTCTCAC -3'	3'-AAAAATCG TCTCTTGT GATGTACT- 5'	161
18a	45	2875	E2	Chr1: 209430 141	1q32.2	AGCTGAACCTGCTGCATTAAACCAATCCTCAGCCACTTGGTCTTCAGGATTCCAGGG ATCCCAGGCAGTAAATTCTGCTGATAATAGGAATTGGTGTGATAAGGTGGGTGCTGAGC[NNN]T ACTAGGAGGCACCACCTGGTGGTCAGTTGGTAATACCATGTTCCCTGCTGAAATAGTATTG CATTTCCACACCGTATAAGTGCCTAACAACTTA	5'-ACCAATCCTC	3'-TTACAGCA	163

							AGCCACTTG-3'	AGGGAACA TGGT-5'	
18b	45	5115	L2	Chr1: 209409 245	1q32.2	CGGGCAAATCACCTGAGGTCAAGACCAGCCTGGCAACATGATGAAACCCCTGTCTCTACTAAAAAATACAAAATTAGCCAGGGGGTTGTGGCACATGCCGTAAATCCCAGCTACTTGGGAGGCTGAAGCAGGAGAATTGCTGAGCCCAGGAGGGCAGGGTTGCAGGGAGCCGAGATTGGCGCATGCAAGCCAGCCTGGGTGACAGGAGTGGAGACTCCGCTCAAAAAAAAAGAAGAAGAAGAATAAGGAAGTCAAATTAGAGGGTTAGAATCTATAAGTTAAGGGTGGGGTGGAGAGATATTATGTAGTCATTGAAATCAGATCACA[NNN]TGTAGATTAGTGGGTCAAAGGGCAACCATGTTACACGTAGTGGTAAACAATAGGGGGTAGGGTACATTTCACCATGATATAAGCCCCTGCTGCTACAGAGGAATTGAATTGCAAGCCTTAATTAGTGTACAAATGATAGTGTACCTGTTGATGTATATGCAGACTTCCCACCTCCTCGTCCACTACACCTAGCACTACACAAATCATTTACATATCCAAGTATTCTTGACCATGCCTTACTGTCATCCTTACAGTAATGTTACAGTACCATTAACATCTGCATGGGATGTAACCTATATACTGCCCGGACATT	5'- GTTAAGGGTG GGTGGAGAGA -3'	3'- GGAGGTGG GAAGTCTG CATA-5'	246
19a	31	4893	L2	Chr6: 802516 36	6q14.1	TTTGTGTGTTAGATTGATTCTAATAACTTAATCCCTACCATACCTTGAAAATGAGAGAACAAACAGCTGTGATGTTGCTGCAAATACCATACACAAACAGAAATACCTAAAGAAAATGATGGCTTTCAAGCCTAAAGAATAATTTTATGAGAACAGAGTATGAAATAATTGAAAGTTAGGGCGGGAGTAGAAGTTGCTCTTGATAATACATGATCTCTGGTTATAAAA[NNN]CAAAACAGCTAATTACATATGAAAACCCCTGCCTATGAAACTGTTAAATGCTGAAGAACATTTACTTTCCAATACATCGCATAATATAGCCCCTGATCCGACTTCTAGATATTAGCATTACATAGGCTGCCCTACCTCACGTTAGGAACACTG	5'- TAGGCCGGGA GTAGAAGTTG -3'	3'- CTACGTGA GGTAAGGG CAGG-5'	210
20a	31	214	E6	ChrX: 437633 26	Xp11.1	CCTGTGGTGCCTAACCCATCATAATGTCTAAAAAATAAAATGAATGGATTCTGTAAATTGTTATTTAATTCTGAATTAGCAATCTGTTACTTACTCAGCAAACAGAAATGCTCCTATGCTGTTTGAGAGGCCAACAAATTGTTGGCATACTCATTGAAAGTGTCAATTGCTCTTCTTACTAGATTAGTTGAGATAGCATTAACTGGTTGGACAGTTT[GNN]AGTAACAGAACAGAGGATTAGATTGCTTACAGATTAAACATGTTAGTATAGGGGACGACACACCACGGAGTGTACAAAATGTTAAGATTATTCAAAAGTAAGTGAATTAGTGTATAGATAGTGTATGGAACACATTAGAAAATTGACAAACAAAGGTATATGTGATTGTTAATTAG	5'- GAATGCTCCT ATGCTGTTTG C-3'	3'- GTGTGGTG TGTGTC TAT-5'	200
21a	45	1646	E1	Chr8: 109485 377	8q23.1	CTGGGAAGGCCCTGCTCCTGTAAGTGGAAAACCTACCTTCAGATCATGCTTATTCTTCTTCTAACAAATTAGTTTATCTTCATTATTACTTGCTAAATGCATTACCGCTAAAGACTATATCCATGAAACAATTAGGACAAGGCTTGGATCAAGGATCCTAAATTTCAGCGTGCACAAAGAATCACCTGGAAGCTGCTTCAAACAGAAAGTGTGAGGCCCTACCCCTATAAAATGCTAAACAGTAAGCTTGCTGAATTGCTGAATGAATGAAAGCACATAGCGGGACTATCGGTAACCA[NNN]GCCACATTATATCTTAATAAGCTAATATTAACTCCCCATTACAATCTAAACATTGGATATGGCGTATAACGTGCTGGTTAAATGTTAAGCCTCTGCTACCGTTGGATTAACTCCAATATAGCCATTACCCAACTCTGTACATGTTGTTTACTCACTTTAAAATTCTAACCAAATCCGTAATGACAGCCCATATATGTCTTAAATACTGCCAGCATTGCAAGCCTTTGTTACTTGTGTAAAGCTCCATTAGTCTGTAATACTGCAATGCGGATCTACATTCTGCAATTGTCACTACTATCCCCACCACTA	5'- GAAAGTGTG AGCCTTACCT -3'	3'- TGGTAAT GGCTATAT TTGGAGT-5'	248
21b	45	2875	E2	Chr9: 125230 778	9q33.3	GAATGGTTTGTGTAATAGTTCCATCTGTCGGCCGGGGCGCAGTGGCTCACACCTGTAATCCCGCACTTGGGAGGCCGAGGTGGCGGATCACAAGGGTCAAGGAGATTGAGACCATCCTGGCTA[NNN]GGCACCCACCTGGTGGTCAGTTGGTAATACCATGTTCCCTGCTGTAAATAGTATTGATTTCACACAGTATAAGTTGCAATACTTA	5'- TAGTTCCATC TGTGGCCG-3'	3'- TTACAGCA AGGGAACA TGGT-5'	161

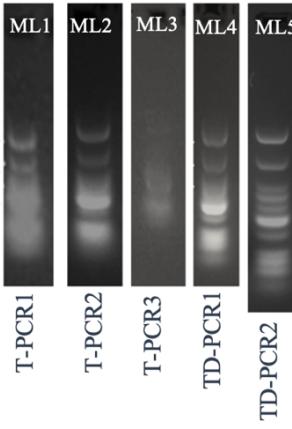
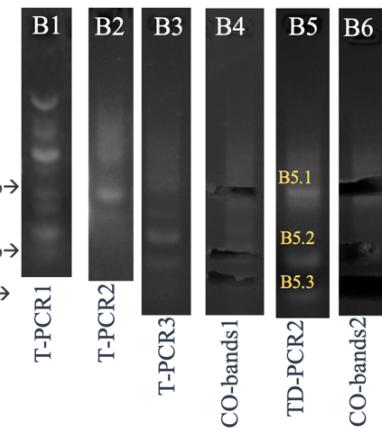
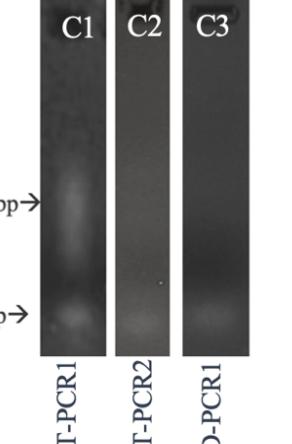
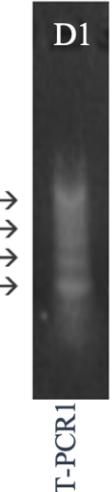
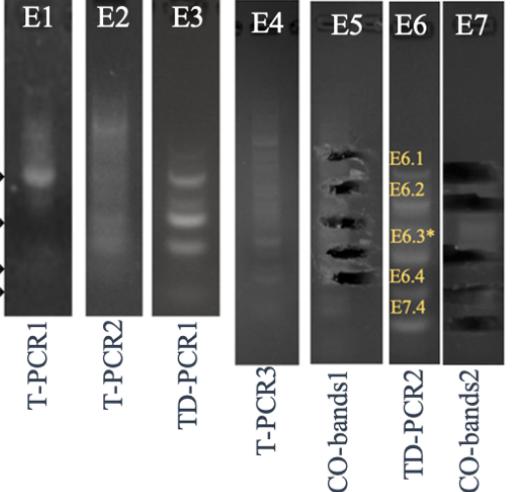
21c	45	5394	L2	Chr8: 109595 503	8q23.1	ACCCAACCCTCTTATAGGGCTGTGGCTCACGAATAACACATGGTGAGAGCTGTTATCTCCACA TCAGTTCATATTGTGTTCTCCAGCTTTGCTAGCACACGTGGGCTGAGACCATAGTC[NNN]A ATGTTACAGTACCAATTAAACATCTGCATGGGATGTACCTATATACTGGCCCGGACATTATATTG CCATCCCATACTCCATGTGGCTAGTACATCT	5'- TGGCTCACGA ATAACACATG G-3'	3'- GGCAATAT AATGTCCG GGCC-5'	173
21d	45	4358	L2	Chr9: 125230 411	9q33.3	GAECTCAAGGGATCCACCCGCCCTGGCCTCTAAAGTGCTGGGATTATAGGCGTGAACCACTGCTC CCGGCTCATACCTGTTATTGAGTGGAGGGATGAAGAAATGAACATGTTCTGCATTGCTCAGACT ACAAGAACTGGAACAATGGGCAGAAGGCATGGGAAGTCCTGTAGTGTATCGGGCTGACTTCC CAAGTCTGAATTATCTGGCTCTATGACAGCTGAGGCCAGGGACACTGGCTTGATTCTTGCA TTGGCTAGAATTGGCCACACCGGGTATTCTCCCTTACT[NNN]TGATAGAATTTCAGTG GTCTAGCCTTGAATAATTGGCTGGCCTTGGCATGGTACCGGCAGTGGTCTGGAGGCCGTA CGGGCTATGTACCCCTTCTTACGGGGCAGGTCAAAACTGTTGGATGTTGGCCCCACTAGGCCA CCTGTGGTTATTGAACCTGTAGGGCCTACTGATCCATCTATTGTACGTTGGTAGAGGATTCCAGT GTTGTTGCTCTGGTGTCCGGTCCCACATTACCGAACCTCTGGTTGAAATTACGTCTTCT GGTACTACCACACCAGCTGTGGACATCACA	5'- AGTGTATCGG GCTGACTTCC- 3'	3'- CACTGCCG GTACCAAT GC-5'	200

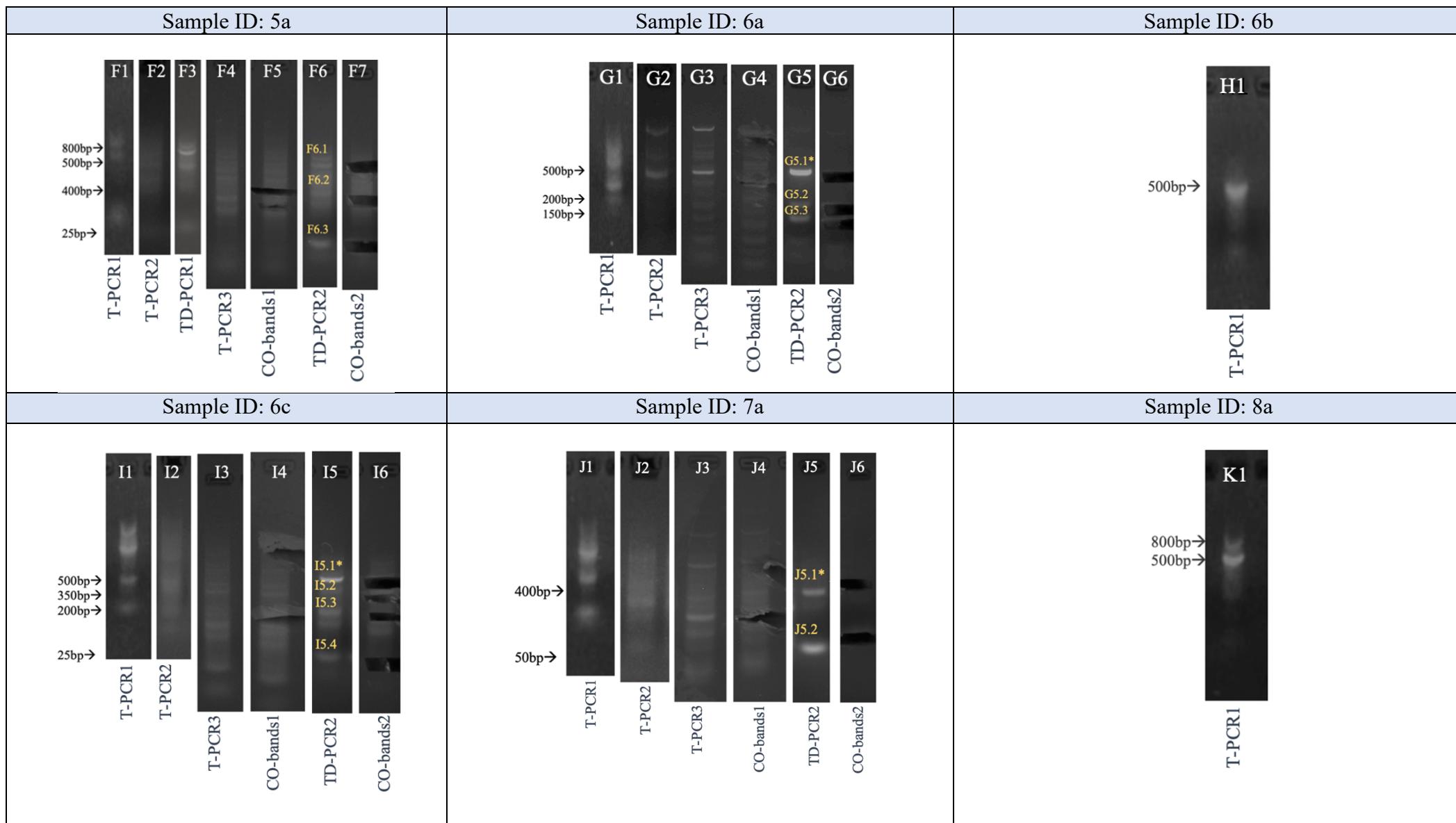
<sup>1</sup>Shows an overview of the 31 qualified HPV integrations from 21 samples, their reported breakpoint in the human and HPV genomes, designed DNA templates and primer pairs (forward and reverse). The blue color represents the human-specific sequence while the orange color represents the HPV-specific sequence. The Forward-primer is human-specific while the Reverse-primer is HPV-specific.

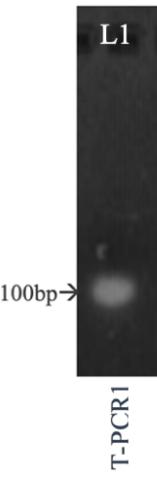
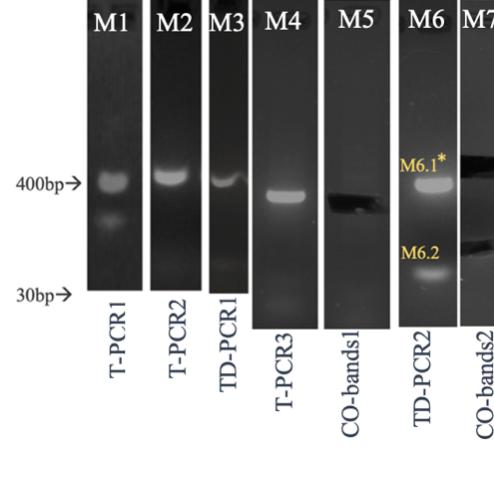
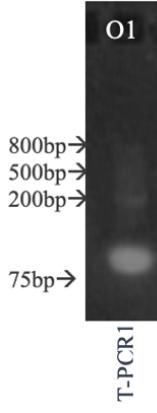
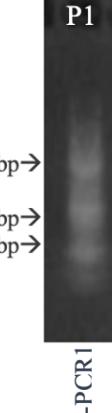
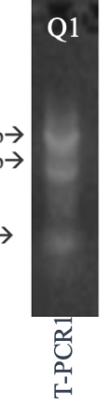
Abbreviations: E= early, HPV= Human papillomavirus, bp= base pairs

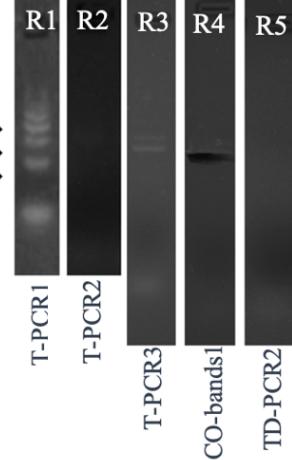
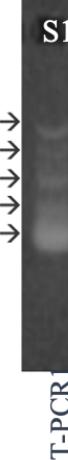
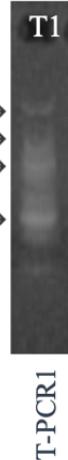
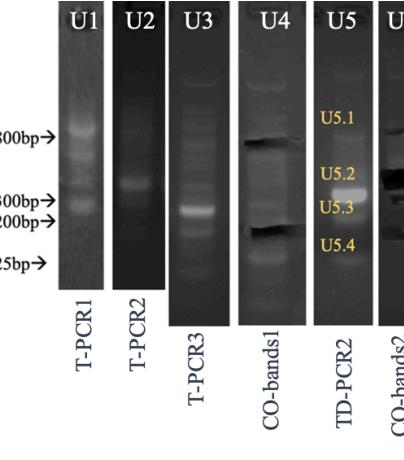
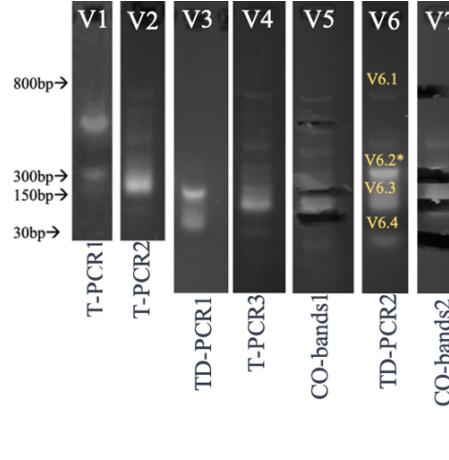
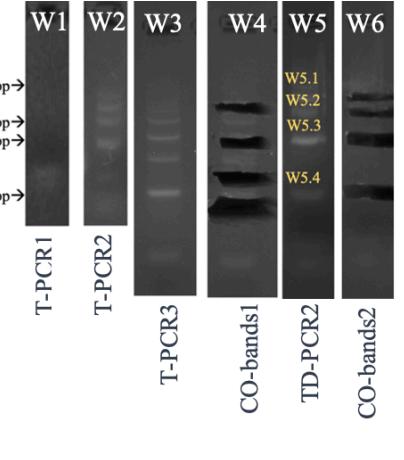
## Appendix 4. Agarose gel runs

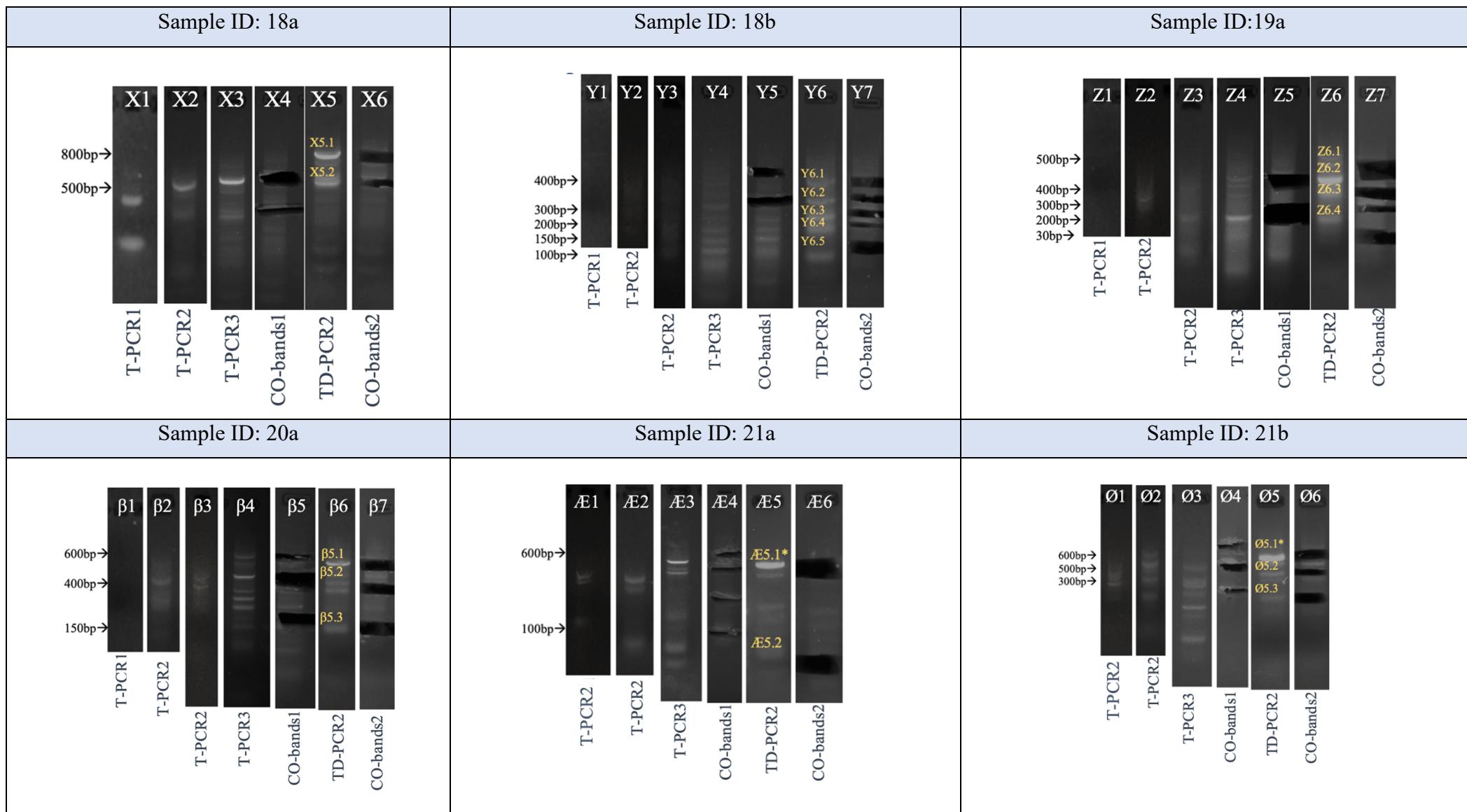
Supplementary table 4A: The various gel runs in the qualified 31 Human papillomavirus (HPV) integrations

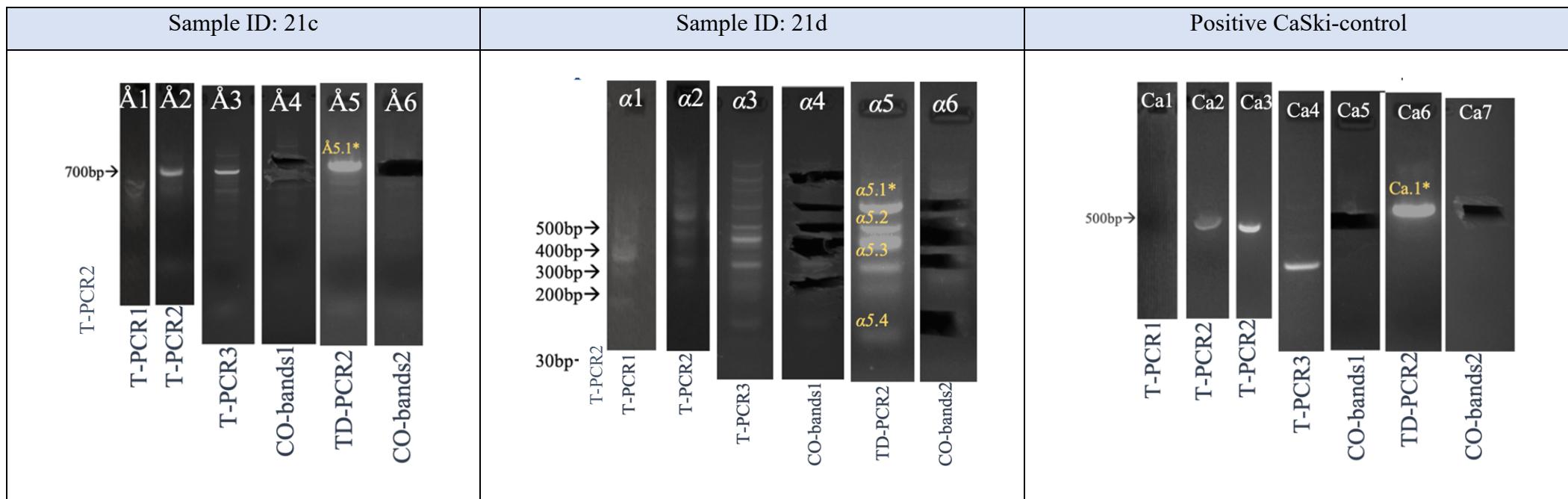
<sup>1</sup> Molecular Weight Ladders (ML)	Sample ID: 1a	Sample ID: 1b
		
Sample ID: 2a	Sample: ID 3a	Sample ID: 4a
		



Sample ID: 9a	Sample ID: 10a	Sample ID: 11a
<p>L1</p>  <p>100bp →</p> <p>T-PCR1</p>	 <p>400bp →</p> <p>30bp →</p> <p>T-PCR1</p> <p>T-PCR2</p> <p>TD-PCR1</p> <p>T-PCR3</p> <p>M1</p> <p>M2</p> <p>M3</p> <p>M4</p> <p>M5</p> <p>M6</p> <p>M7</p> <p>CO-bands1</p> <p>TD-PCR2</p> <p>CO-bands2</p> <p>M6.1*</p> <p>M6.2</p>	 <p>800bp →</p> <p>500bp →</p> <p>150bp →</p> <p>N1</p> <p>T-PCR1</p>
Sample ID: 11b	Sample ID: 12a	Sample ID: 12b
<p>O1</p>  <p>800bp →</p> <p>500bp →</p> <p>200bp →</p> <p>75bp →</p> <p>T-PCR1</p>	 <p>500bp →</p> <p>150bp →</p> <p>100bp →</p> <p>P1</p> <p>T-PCR1</p>	 <p>500bp →</p> <p>200bp →</p> <p>50bp →</p> <p>Q1</p> <p>T-PCR1</p>

Sample ID: 13a	Sample ID: 14a	Sample ID: 15a
 <p>R1 R2 R3 R4 R5</p> <p>800bp→ 500bp→ 200bp→</p> <p>T-PCR1 T-PCR2 T-PCR3 CO-bands1 TD-PCR2</p>	 <p>S1</p> <p>800bp→ 500bp→ 200bp→ 150bp→ 100bp→</p> <p>T-PCR1 T-PCR2</p>	 <p>T1</p> <p>900bp→ 800bp→ 500bp→ 150bp→</p> <p>T-PCR1 T-PCR2</p>
Sample ID: 15b	Sample ID: 16a	Sample ID: 17a
 <p>U1 U2 U3 U4 U5 U6</p> <p>800bp→ 300bp→ 200bp→ 25bp→</p> <p>T-PCR1 T-PCR2 T-PCR3 CO-bands1 TD-PCR2 CO-bands2</p>	 <p>V1 V2 V3 V4 V5 V6 V7</p> <p>800bp→ 300bp→ 150bp→ 30bp→</p> <p>T-PCR1 T-PCR2 T-PCR3 TD-PCR1 TD-PCR2 CO-bands1 CO-bands2</p> <p>V6.1 V6.2* V6.3 V6.4</p>	 <p>W1 W2 W3 W4 W5 W6</p> <p>700bp→ 500bp→ 300bp→ 100bp→</p> <p>T-PCR1 T-PCR2 T-PCR3 CO-bands1 TD-PCR2 CO-bands2</p> <p>W5.1 W5.2 W5.3 W5.4</p>





<sup>1</sup>Displays the various gel runs from the 31 qualified HPV integrations, the molecular weight ladder and the positive CaSki control included in the runs. The agarose gel-bands marked with yellow color are the gel-bands included in the gel extraction method. The gel-bands marked with yellow color and with \* resulted in confirmed HPV integrations. The gel-bands only showing T-PCR1 was confirmed by directly Sanger sequencing and did not require further adjustments (T-PCR2 - CO-bands2). The various gel runs are following:

T-PCR1= Traditional PCR first run, followed by a 2% agarose gel run at 100 V for 60 minutes

T-PCR2= Traditional PCR second run, followed by a 2% agarose gel run at 70 V for 60 minutes

T-PCR3= Traditional PCR third run, followed by a 2% agarose gel run at 70 V for 120 minutes

TD-PCR1 = Touch Down PCR first run, followed by a 2% gel run at 70 V for 60 minutes

TD-PCR2 = Touch Down PCR second run, followed by a 2% agarose gel run at 70V for 135 minutes

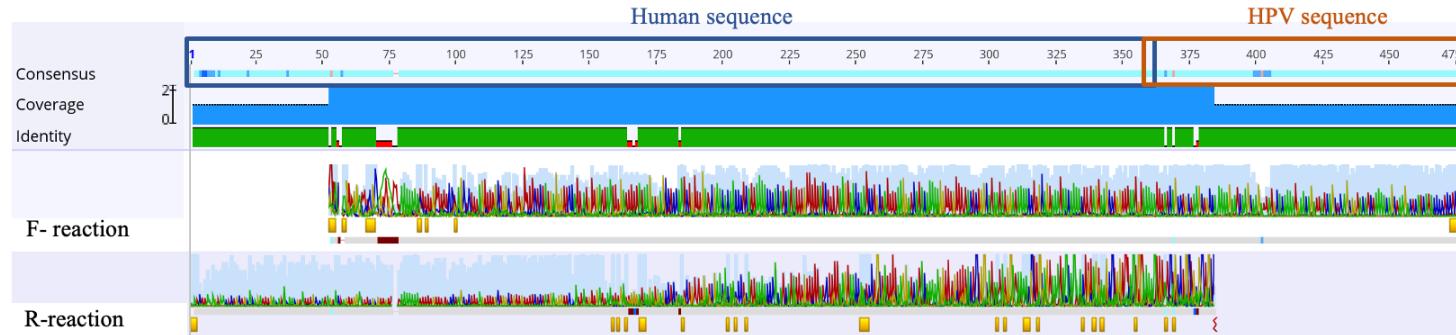
CO-bands1= Cut out bands first time.

CO-bands2= Cut out bands second time

Abbreviations: Min= Minutes, V= Voltage

## Appendix 5. Sanger sequencing chromatograms from confirmed HPV integrations

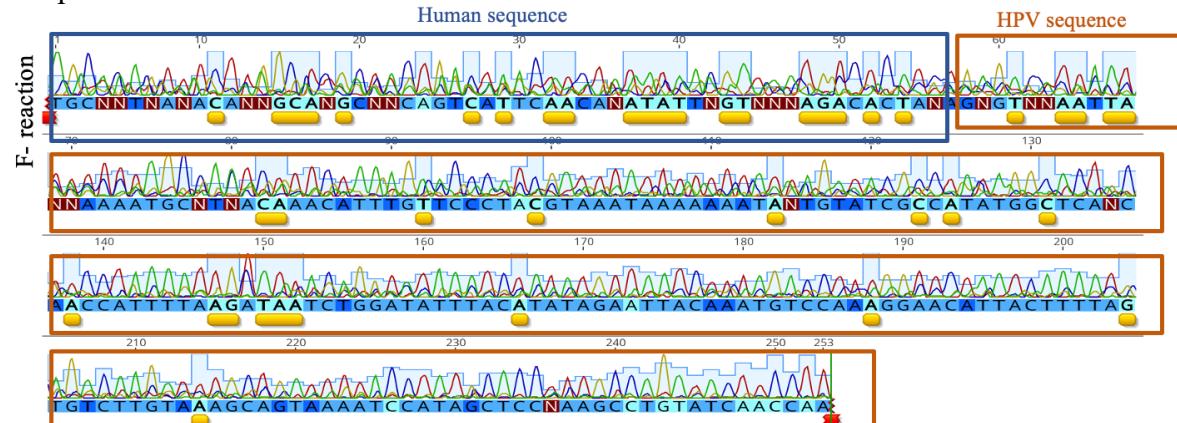
Sample ID: 1a



"TACAATTGATTACAGCAGCTCTGAAAGTCTATCT  
TCGATAAAGCCTATRATCATGAAGGTAAACCGAAG  
AAATTTTTAAAAAGAAAACCATGCACAAATGAA  
AAGCAGGCAGAAGAAAACACATAGTGGGAACCTAAT  
AGAAGTAATTAAAGTAAAATATTGTAGTGCCACCTT  
ACAGAAAGTACGTTCTTCCCACATAATTCCCTCTA  
CTATAAAGAAAATACAAAATTGATTCCCTTACAGCA  
TAAAATTGACTATCTTTGAAAGAAAATCACAG  
GCTATTATAGGTATTCCAAGCTGATTATTCTTTC  
AAATGATTGCCAGGAAATTAG ATT [NNN]  
TTAAATACWGCGCAGATTGCAGCCTTTGTTACTT  
GCTGWAATAGCTCTTGTAAACTGCACT  
GCGGATCTACATTTCTGCATTGCACTACTATCCCCA  
CCAC"

**Supplementary figure 5A:** Shows the continuous Sanger sequence chromatograms of assembled Forward (F)- and Reverse(R)-sequences in sample 1a, with a pairwise identity of 94,9%. The continuous sequence is mapping to the human chromosome 3 and Human papillomavirus (HPV) 45 as reported from the Next Generation Sequencing (NGS) data thereby confirmed. The screenshot is obtained from Geneious v2020.2.2.

Sample ID: 3a



"TGCNNNTANACANNGCANGCNAGTCATTCAAC  
ANATATTNGTNNNAGACACTANAG  
[NNN]NGTNNAAATTANAAAATGCTNACAAACAT  
TTGTTCCCTACGAAATAAAAAAAATANTGTATCGC  
CATATGGCTCANCAACCATTAAAGATAATCTGGA  
TATTTACATATAGAATTACAAATGTCAAAGGAAC  
ATTACTTTAGTGTCTGTAAAGCAGTAAATCCA  
TAGCTCCNAAGCCTGTATCAACCA"

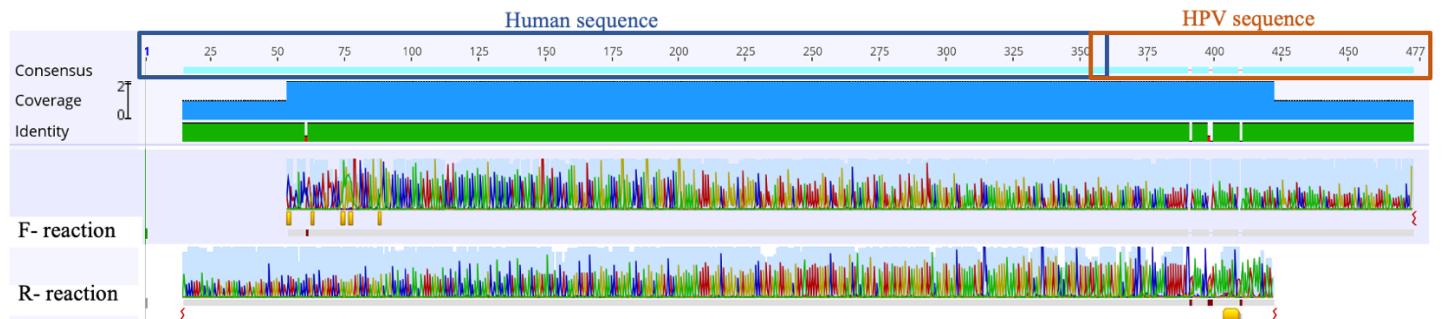
**Supplementary figure 5B:** Shows the Forward(F)- sequence in sample 3a as no continuous sequence was identified. The F-sequence is mapping to the human chromosome 13 and Human papillomavirus (HPV) 31 as reported from the Next Generation Sequencing (NGS) data thereby confirmed. The Reverse (R)-sequence was of low-quality, partly mapping to the human chromosome 13 (chromatogram not shown). The screenshot is obtained from Geneious v2020.2.2.

## Sample ID: 4a



**Supplementary figure 5C:** Shows the continuous Sanger sequencing chromatogram from assembled Forward(F)- and Reverse(R)-sequences in sample 4a, with a pairwise identity of 97,9%. The continuous sequence is mapping to the human chromosome 1 and Human papillomavirus (HPV) 45 as reported from the Next Generation Sequencing (NGS) data thereby confirmed. The screenshot is obtained from Geneious v2020.2.2

## Sample ID: 6a

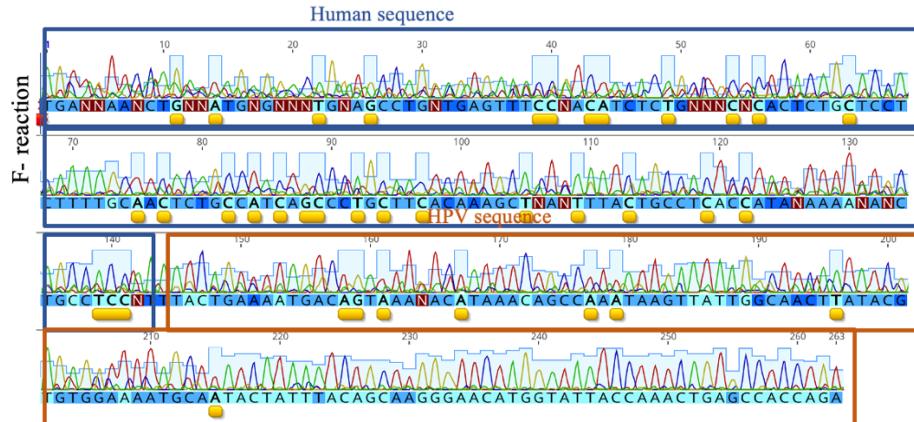


**Supplementary figure 5D:** Shows the continuous Sanger sequencing chromatogram from assembled Forward (F)- and Reverse(R)-sequences in sample 6a, with a pairwise identity of 98,8%. The continuous sequence is mapping to the human chromosome 5 and Human papillomavirus (HPV) 45 as reported from the Next Generation Sequencing (NGS) data thereby confirmed. The screenshot is obtained from Geneious v2020.2.2

“TAGAGGTCAAGGCTGCCATAAGCCAAAT CGCACCACTGCACTCAGTCTGGGTGACAC AGTGAGAGCCTTCTCAAAGTAATAACTG GCCTCTGATTGTGGGGCCTGGATAAGCAGC TCTGAAATTACACAGGGCTG[NNN]TCATTTA TTTCATATACTGGATTACCATTTTATCAAAT TGGAAATGCATGTGAAATGTAAATACCG TCACNNNNNCNMTGRNGTGCAGTGGTGCN ATTTGGCTNATNGSRNNCNTN”

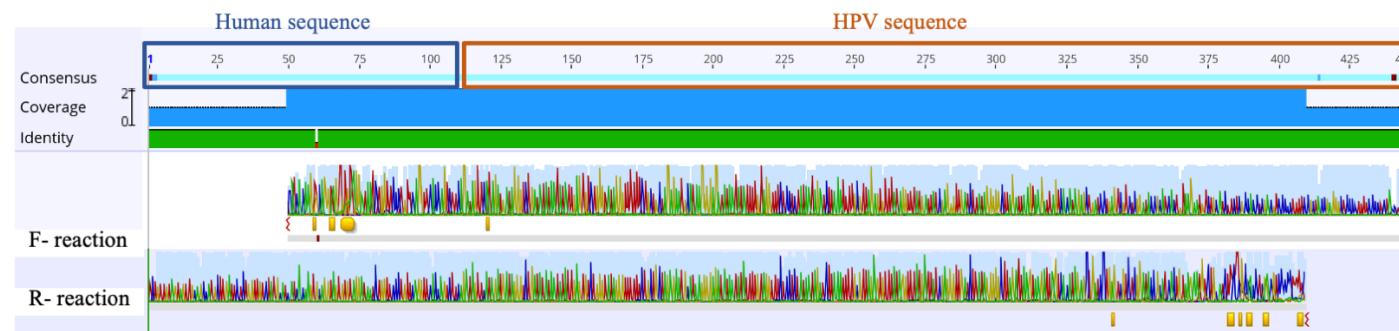
“TCCCAAATCTCTCAGCGAGTTCACGATGGAGCCA GGACACAACCTCCAGCACAGAACACGTGGGAGCA CATGGTCTGCTCCCCTCACTTCTAAATGTATCACATC CTTCTCTTGTGCTCTCCATTCTCATGATTCTATC CACTTGGCAGACATTGGTCCATGTGTCCTACTGGG CTGGACACTGTGTTGTAAAAGTGGCAGGGGGAG GGGCCAGAAGTGAAGGGCAATAAAAAGGAGCAA GGATGCATCTGAGACTGAAGTAGAGGATGGGAGA ATCTGGAGTGCAGTTTAGACCTTGGGTTGAG GTAGCTGCAGGAGCTTAGGCA GATAAT[NNN]GATGTAAGGTCTGGACCATGTCTTTT CAAAAAAAC-ATTTCCA-ATTTTATCA- TTTATTCATATACTGGATTACCATTTTATCAAAT GGAATGCATGTGAAATGTAAATACCG”

## Sample ID: 6b



**Supplementary figure 5E:** Shows the chromatogram from the Forward (F)- Sanger sequence in sample 6b as no continuous sequence was identified. The F-sequence is mapping to the human chromosome 5 and Human papillomavirus (HPV) 45 as reported from the Next Generation Sequencing (NGS) data thereby confirmed. The Reverse(R)- sequence was of low-quality, partly mapping to the human chromosome 5 (chromatogram not shown). The screenshot is obtained from Geneious v2020.2.2.

## Sample ID: 6c

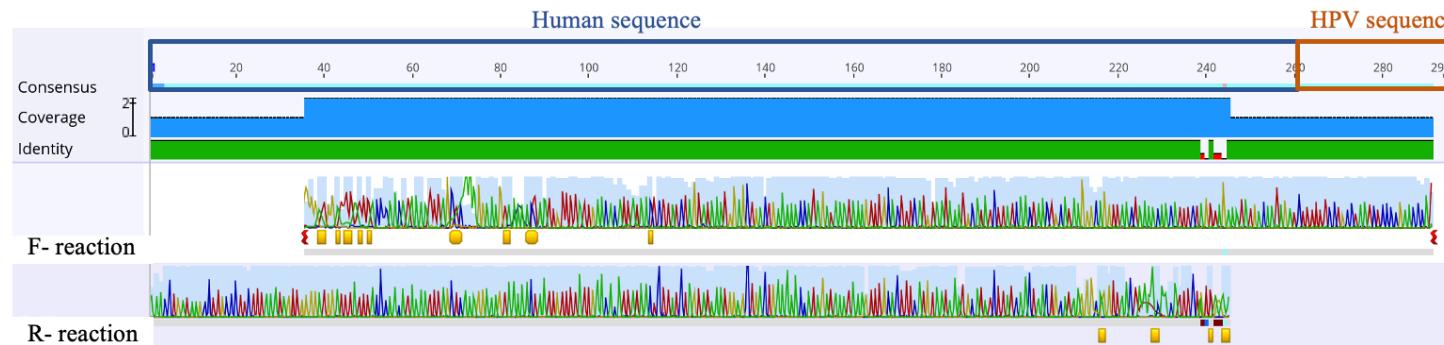


**Supplementary figure 5F:** Shows the continuous Sanger sequencing chromatogram in sample 6c from assembled Forward(F)- and Reverse(R)-sequences, with a pairwise identity of 99,8%. The sequence is mapping to human chromosome 3 and Human papillomavirus (HPV) 45 as reported from the Next Generation Sequencing (NGS) data thereby confirmed. The screenshot is obtained from Geneious v2020.2.2.

"TGANNAANCTGNNAATGNGNNNNTGNAGCTGTNTGAGTTTCCNACATCTCTGNNNCACACTCTGCTCCCTCTTTTGCAACTCTGCCATCAGCCCTGCTTCACA AAGCTNANTTTACTGCCTCACCATANAAAANANCTGCCTCC[NNN]NTTTACTGAAATGACAGTAAANACATAAACAGCCAATAAGTTATTGGCAACTTATACGTGTGAAAATGCAATACTATTTACAGCAAGGAACATGGTATTACCAAACTGAGCCACCAAG

"CAGAAGATGATGAAGCTGCCTAGCAAGTAACACAGGAAGACATAGCTACTACTCTTGAGCTGGATGAGAGATG GTATTACAGGAACCCAGAACTGGAACTCCTG[NNN]ACAACTGTATGAACTATGTAGTATGGGACAGTATATATTATAAAGATG TATAAGTGAGACAGGGATATGGAAAAAACAGCAGC ATGTGTTAGCTATTGGGGTGTATATTATATAAAGATG GAAACACCACATATTATGTACAATTAAAAGCGAATGTGAGAAATATGGAAATAGTAATACGTGGAAAGTACAA TATGGGGCAATGTAATTGATTGTAATGACTCTATGTG CAGTACAGTGTGACGACACGGTATCGCTACTCAGATT GTTAGACAGCTACAACACGCGCTCACGTCGACCCCCA AAACCGCAGTCGTGGCACCCAAACCCCCACATNNAGACG"

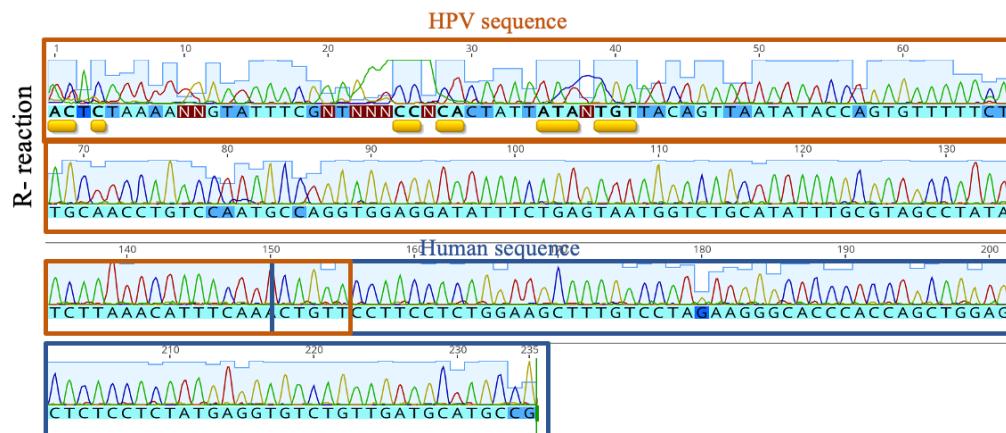
Sample ID: 7a



“TTTCCAGTATTGCCAAGCCAAACAATTATTAAGGGGGTAGGTAAATAATCCCTCTATTAAATGCACTTGGTTGATTTCTATTAAGAAATAAGTCTCTAATATTCAACATACATGATATTAGGTGAGTCAGTCTTCCATGTATTATTTGAAAGCTCAAACAGATTATTAGCTATACAATACCTCATTCAAGCTATTGGGCTGGAATTCACTTCTAAACATGAATGTTATAGTCTTATGTA[NNN]CAAAAAACCAAGGCCATTACACCCGTTCCCTA”

**Supplementary figure 5G:** Shows the continuous Sanger sequencing chromatogram in sample 7a from assembled Forward(F)- and Reverse(R)- sequences, with a pairwise identity of 98,0%. The sequence is mapping to the human chromosome 6 and Human papillomavirus (HPV) 45 as reported from the Next Generation Sequencing (NGS) data thereby confirmed. The screenshot is obtained from Geneious v2020.2.2

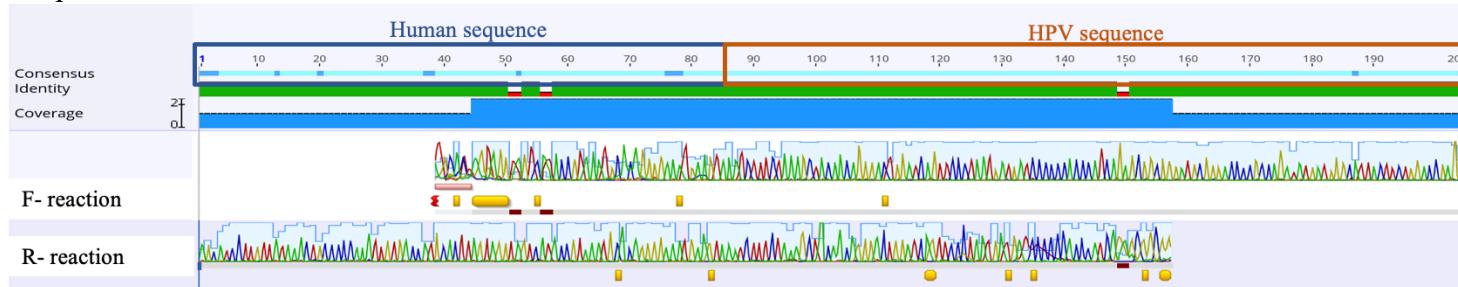
Sample ID: 8a



“ACTCTAAAANNGTATTCGNTNNNNCCNCACTATTATANTGTTACAGTTAATATACCACTGTTTTCCTGCAACCTGTCCAATGCCAGGTGGAGGATATTCGAGTAATGGCTCTGCATATTGCGTAGCCTATAACTGTTNNNCCCTCTGGAGCTTGTCCCTAGAAGGGCACCCACAGCTGGAGCTCCCTATGAGGTGTCTGTTGATGCATGCCG”

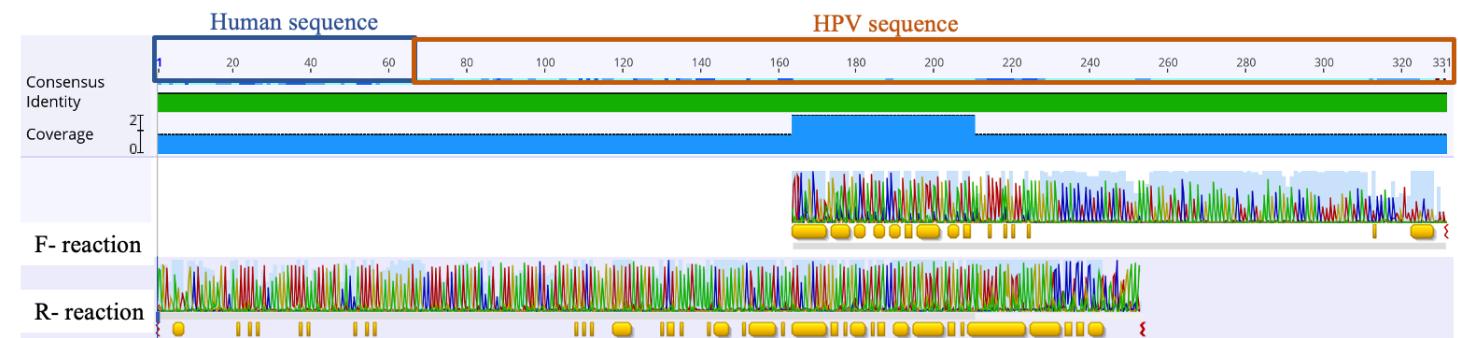
**Supplementary figure 5H:** Shows the chromatogram from the Reverse(R)- Sanger sequence in sample 8a as no continuous sequence was identified. The sequence is mapping to the human chromosome X and Human papillomavirus (HPV) 45 as reported from the Next Generation Sequencing (NGS) data thereby confirmed. The Forward(F)- sequence was of low-quality, partly mapping to HPV 45 (chromatogram not shown). The screenshot is obtained from Geneious v2020.2.2.

Sample ID: 9a



**Supplementary figure 5I:** Shows the continuous Sanger sequence from assembled Forward(F)- and Reverse(R)- sequences in sample 9a, with a pairwise identity of 96%. The continuous sequence is mapping to the human chromosome 15 and Human papillomavirus (HPV) 45 as reported from the Next Generation Sequencing (NGS) data thereby confirmed. The screenshot is obtained from Geneious v2020.2.2

Sample ID: 10a



**Supplementary figure 5J:** Shows the continuous Sanger sequence from assembled Forward(F)- and Reverse(R)-sequences in sample 10a, with a pairwise identity of 100,0%. The continuous sequence is mapping to the human chromosome 11 and Human papillomavirus (HPV) 45 as reported from the Next Generation Sequencing (NGS) data thereby confirmed. The screenshot is obtained from Geneious v2020.2.2

“TGCTACCAAGAAATTTCAGCCCCCTTA  
AGAGTCCTAGCTCAATGATGGATGACTATC  
CGATTAGCTCTGGGGTTAGTTGAT[NNN]  
TTACAAACATTGCATCTCTGGGTCAAGGT  
ACTGAACCCATTAGTAGTACCCCCCTCCCT  
ACTGTGCGGCGGGTAGCGGGTCCCCGCGT  
TATAGTAGGGCTAATCACAGGTCCGTGA”

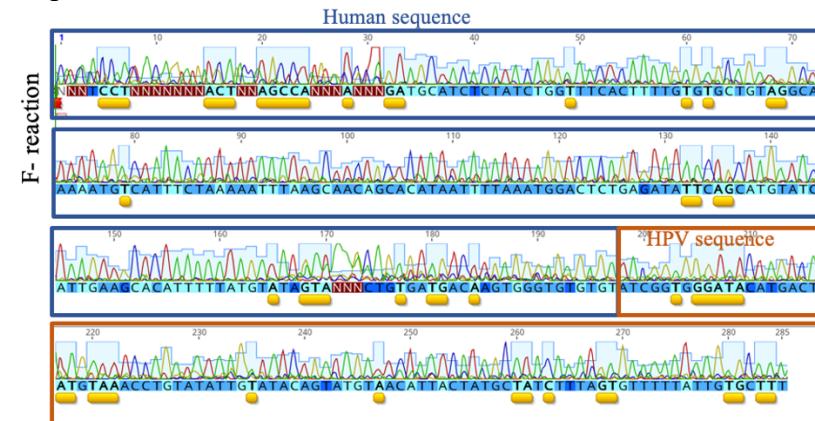
“TGAGTGATGCATGTGACTGAAAATTACTT  
GGTAAAATTTAACAGTAGCTCTAAAGAGTG  
GTGATG[NNN]CCAATACATATGATCCTACT  
AAGTTAACGACTATAGTAGACATGTGGAGG  
AATATGATTACAGTTATTTTCAGTTGTC  
ACTATTACTTAACTGCAGAGGTTATGTCATA  
TATCCATAGTATGAATAGTAGTATATTGGAA  
AAGGGGAATTGGTGTACCTCCACCACCTA  
CTACAAGTTAGTGGATACATATCGTTTG  
CAATCAGTTGCTGTACCTGTCAAAAGGATA  
CTACACCTCCAGAAAAGCAGNAN”

### Sample ID: 11a



**Supplementary figure 5K:** Shows the continuous Sanger sequence from assembled Forward(F)- and Reverse(R)-sequences in sample 11a, with a pairwise identity of 97,4%. The continuous sequence is mapping to human chromosome 13 and Human papillomavirus (HPV) 45 as reported from the Next Generation Sequencing (NGS) data thereby confirmed. The screenshot is obtained from Geneious v2020.2.2

### Sample ID: 11b

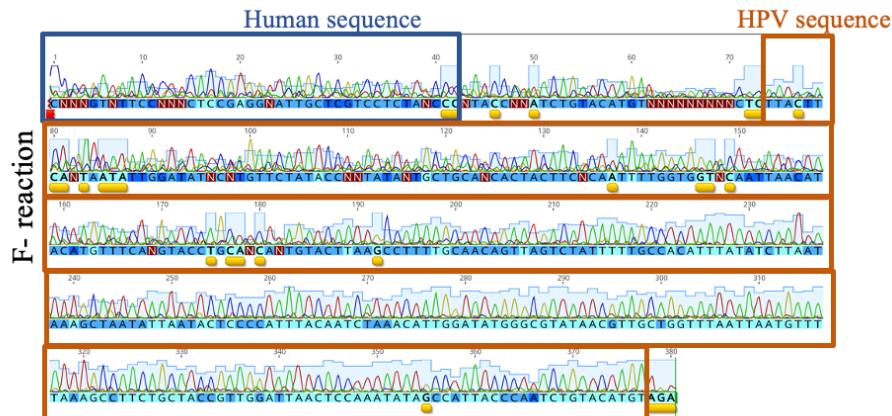


**Supplementary figure 5L:** Shows the chromatogram from the Forward(F)- Sanger sequence in sample 11b as no continuous sequence was identified as a result of low-quality Reverse(R)-sequence (chromatogram not shown). The F-sequence is mapping to the human chromosome 13 and Human papillomavirus (HPV) 45 as reported from the Next Generation Sequencing (NGS) data thereby confirmed. The screenshot is obtained from Geneious v2020.2.2.

"GCTATTTCCCACAGCATCAACAAACTTAC  
 ACATTATTTAAAGCAACAAAGATTTCTAA  
 GGCCTCTCTTCCATCAGTTATAATCAA  
 CCTTACCGGAATGCGGCTGGGTATAAC  
 TGAGTGGACTATCCACTTACCTTACAGA  
 TCTGAACACCTACTTG[NNN]GTGTTCTCAC  
 ATATAACAGTATATGCAATTTTTTGGTG  
 TTCTTTAAGAAATCCTTAGTGCCTTAA  
 AAAGCTAATAAAATTCTACTCCCTGATATCT  
 TAGGAATTGACTATGGGCTCCAATCCCC  
 ACCTTCATCTATTTAGAACATCTATATTAA  
 ATCCATGAGACATATTCAATTGGCGTTTT  
 GTGCTCTTTATAATGCTACACATTACAGC  
 ACAATCTTTAAATATTNGGCCTGGCAGT  
 TACTTTA"

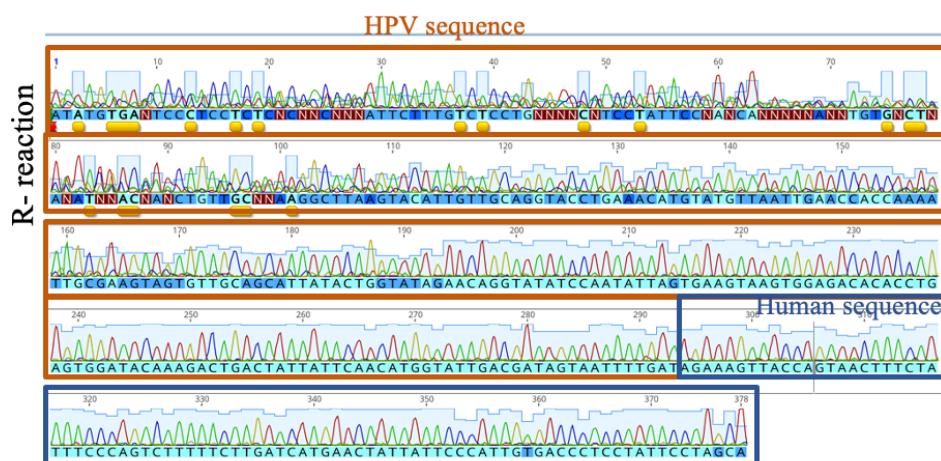
"NNNTCCTNNNNNNACTNNAGCCANNAN  
 NNGATGCATCTCTATCTGGGTTCACTTTGTG  
 TGCTGTAGGCAAAATGTCATTCTAAAAAT  
 TTAAGCAACAGCACATAATTAAATGGACT  
 CTGAGATATTCACTGATGATCATTGAAGCAC  
 ATTTCATGATAGTANNCTGTGATGACAA  
 GTGGGTGTGTG[NNN]TATCGTGGGATACAT  
 GACTATGAAACCTGTATATTGTATACAGTA  
 TGTAACATTACTATGCTATCTTAGTGTGTTT  
 ATTGTGCTT"

Sample ID: 12a



**Supplementary figure 5M:** Shows the chromatogram from the Forward(F)- Sanger sequence in sample 12a as no continuous sequence was identified. The sequence is mapping to more than one human chromosome (chr14, 12, 1, 17, 16) and Human papillomavirus (HPV) 45. The screenshot is obtained from Geneious v2020.2.2.

“CNNNGTNTTCCNNCTCCGAGGNATTGCTC  
GTCCTCTANCC[NNN]CNTACNNATCTGTC  
ATGTNNNNGTCNNCTTACTTCACTAATAT  
TGGATATACCTGTTCTATACCATTATAATGC  
TGCAACACTACTCNCAATTGGTGGTTCA  
ATTAACACATGTTCANGTACCTGCAACA  
ATGTACTTAAGCCTTTGCAACAGTTAGTCT  
ATTTTGCCACATTATATCTAATAAAGCTA  
ATATTAATACTCCCCATTACAATCTAAACA  
TTGGATATGGCGTATAACGTTGCTGGTTA  
ATTAATGTTTAAAGCCTTCTGCTACCGTTG  
GATTAACTCCAAATATAGCCATTACCCAATC  
TGTACATGTAGA”



“ATATGTGANTCCCTCTCNCNNNNATTCT  
TTGTCTCTGNNNCNTCTATTCCNANCANN  
NNANNTGTGNCTNANATNNACNANCTGTTGCN  
NAAGGCTTAAGTACATTGTTGCAGGTACCTGAA  
ACATGTATGTTAATTGAACCAACAAATTGCGA  
AGTAGTGTGCAGCATTAACTGTTAGTAAACA  
GGTATATCCAATATTAGTGAAGTAAGTGGAGA  
CACACCTGAGTGGATACAAGACTGACTATTAT  
TCAACATGGTATTGACGATAGTAATTGGAT[N  
NN]AGAAAGTTACCAAGTAACCTTCTATTCCCA  
GTCTTTCTTGATCATGAACATTATTCCCAATT  
GTGACCCCTCTATTCTAGCA”

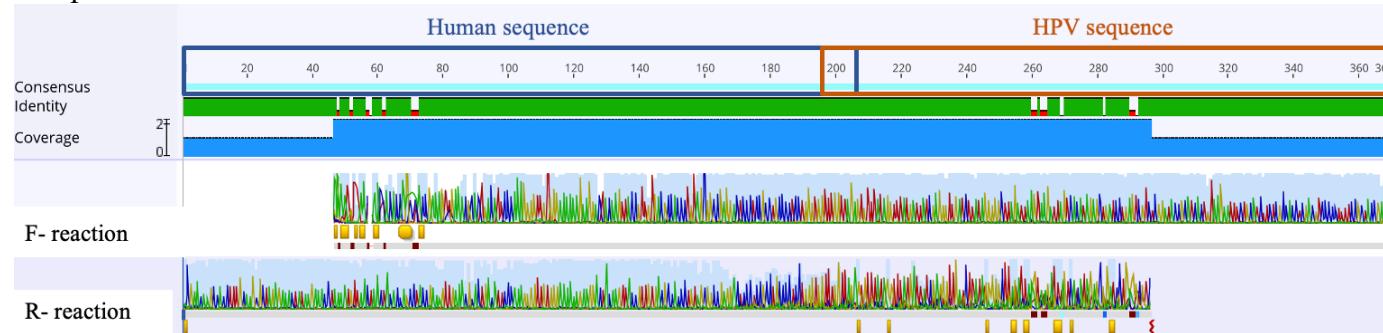
**Supplementary figure 5N:** Shows the chromatogram from the Reverse(R)- Sanger sequence in sample 12a as no continuous sequence was identified. The sequence is mapping to the human chromosome 3 and Human papillomavirus (HPV) 45 as reported from the Next Generation Sequencing (NGS) data thereby confirmed. The screenshot is obtained from Geneious v2020.2.2.

### Sample ID: 12b



**Supplementary figure 5O:** Shows the continuous sequence of assembled Forward(F)- and Reverse(R) sequences in sample 12b, with a pairwise identity of 96,2%. The continuous sequence is mapping to more than one human chromosomes (including chr 3) and Human papillomavirus (HPV) 45 as reported from the Next Generation Sequencing (NGS) thereby confirmed. The screenshot is obtained from Geneious v2020.2.2.

### Sample ID: 14a

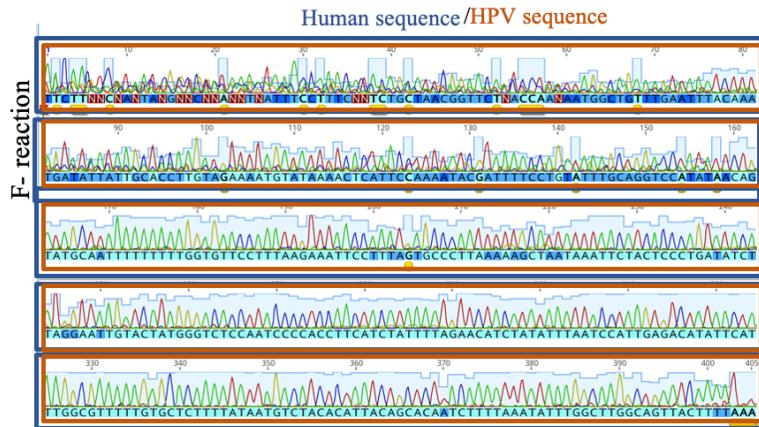


**Supplementary figure 5P:** Shows the continuous Sanger sequence of assembled Forward(F)- and Reverse(R)-sequences, with a pairwise identity of 94,8%. The continuous sequence is mapping to the human chromosome 11 and Human papillomavirus (HPV) 45 as reported from the Next Generation Sequencing (NGS) data thereby confirmed. The screenshot is obtained from Geneious v2020.2.2.

“TCAATCCTGCTCTGAGACTAAATATACTGT  
AATTGCTAAAGCTTAAGTGACAGGCATTAG  
TCAAGGTCAAGTATAACAAACACTTACTGCC  
CTA[NNN]GTACATATATATATACACCATA  
GTACATATACTATGCACACCATAAGTACAATT  
TTTCAGTTGTGCACTTAACTTAACTGCAGA  
GGTTATGTCAATATCCATAGTATGAATAGT  
AGTATATTGAAAATTGGAATTGGTGTAC  
CTCCACCACCTACTACAAGTTAGTGGATAC  
ATATCGTTTGCAATCAGTTGCTGTACT  
GTCAAAAGGAACTACACCTCAGAAAAGC  
AG”

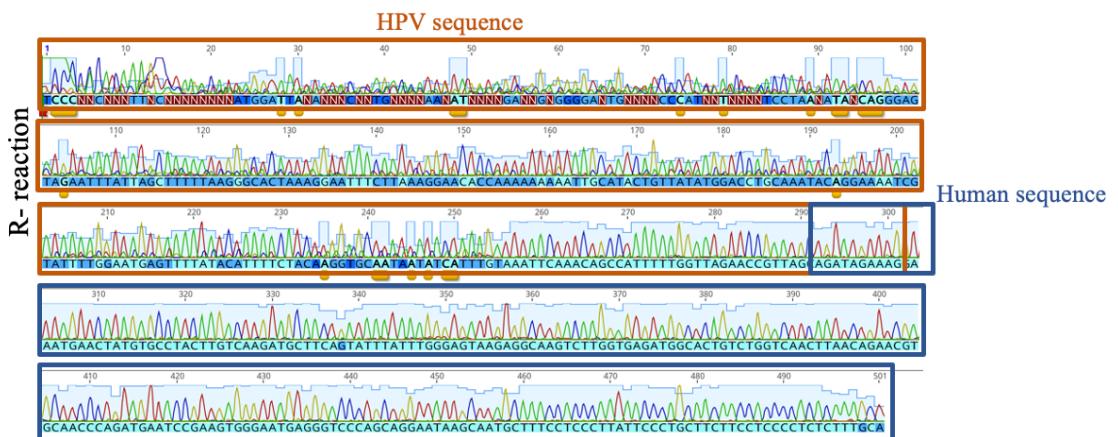
“TCGTTCTTGCCTCAAAGCTTGCATGTG  
TTTCATATTGTTCCATTAGAATTCAATTCA  
TCTCCTGCCATCCGTTGAAGTTTCGCTCAG  
TCTAACCTTTGGGAAGAGGGTTTTGT  
GGTCATCTTGAATCGAAGAGTCGCCTCCA  
GATGCGGCCGTGGTCACCCACCCAGGCC  
CACCAGGGAGG AAAGGAA [NNN]  
TGTGAGAAATATGGAATAGTAATACGTGG  
GAAGTACAATATGGGGCAATGTAATTGAT  
TGTAAATGACTCTATGTGCACTCAGATTGTTAAC  
ACACGGTATCGCTACTCAGATTGTTAAC  
GCTACAACACGCCTCACGTGACCCCCAA  
AACCGCATCCGTGA”

Sample ID: 15a



**Supplementary figure 5Q:** Shows the chromatogram from Forward(F)-Sanger sequence in sample 15a as no continuous sequence was identified. The sequence is mapping to more than one human chromosome (including Chr 8) and to several Human papillomavirus (HPV) types (including HPV45). The screenshot is obtained from Geneious v2020.2.2.

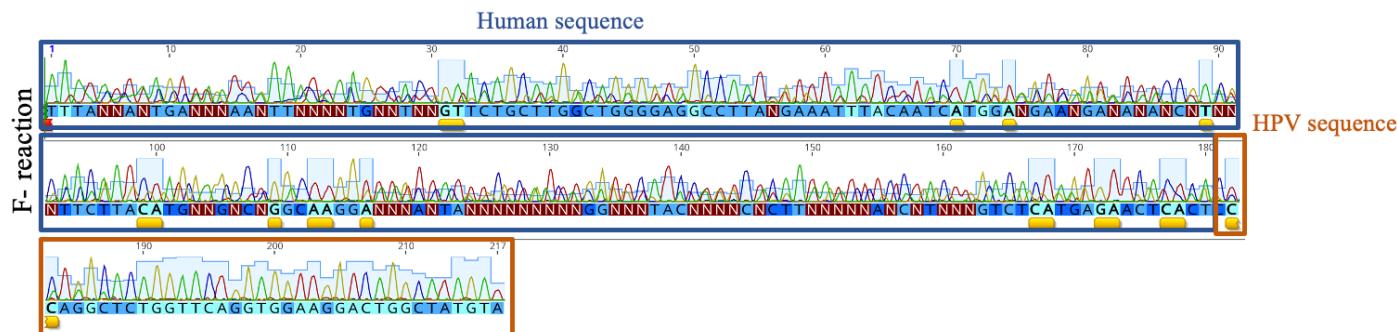
"TTCTTNNCNANTANGNNCNNANNTNATTCTTTCNNCTGCTAACGGTCTNACCAANAATGGCTGTTGAAATTACAAATGTATTATTGCACCTTGTAGAAAATGTATAAAAATCATTCCAAAATACGATTTCCTGTATTGCAAGGTCATATAACAGTATGCAATTTCCTTGTAGTGCCTTAAAGCTAATAAATTCTACTCCCTGATATCTTAGGAATTGTAATGGGTCTCCAATCCCCACCTTCATCTATTAGAACATCTATATTAAATCCATTGAGACATATTCTATTGGCGTTTGTGCTCTTATAATGTCACACATTACAGCACAACTTTAAATATTGGCTTGGCAGTTACTTTAAA"



“TCCCNCCNNNTNCNNNNNNNATGGATTANANN  
NCNNTGNNNNAANATNNNNGANGNGGGGANTGN  
NNNCCCATNNNTNNNTCTAACANATANCAGGGAGTA  
GAATTATTAGCTTTAAGGGCACTAAAGGAATT  
CTTAAAGGAACACCAAAAAAAATTGCATACTGTT  
ATATGGACCTGCAAATACAGGAAAATCGTATTTG  
GAATGAGTTTATACATTTCTACAAAGGTGCAATAA  
TATCATTGTAAATTCAAACAGCCATTGGTTAG  
AACCGTTAGCAGATAGAAAGG[NNN]AAATGAAC  
ATGTGCCACTTGTCAAGATGCTTCAGTATTTATT  
GGGAGTAAGAGGCAGTCTGGTGGAGATGGCACTG  
TCTGGTCAACTTAACAGAACGTCACCCAGATGA  
ATCCGAAGTGGGAATGAGGGTCCAGCAGGAATAA  
GCAATGCTTCCCTCCCTATTCCCTGCTTCTCCTCC  
CCTCTCTTGC”

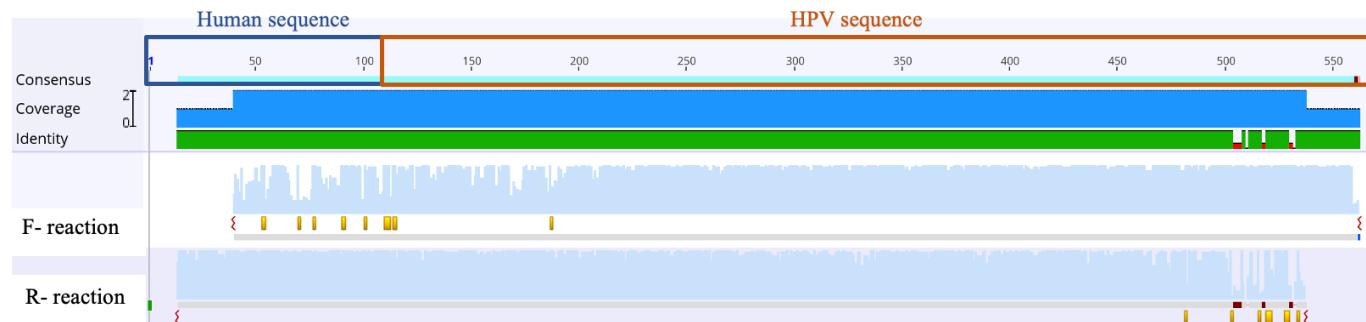
**Supplementary figure 5R:** Shows the chromatogram from Reverse(R)- sequence in sample 15a as no continuous sequence was identified. The R-sequence is mapping to the human chromosome 8 and Human papillomavirus (HPV) 45 as reported from the Next Generation Sequencing (NGS) thereby confirmed. The screenshot is obtained from Geneious v2020.2.2.

## Sample ID: 16a



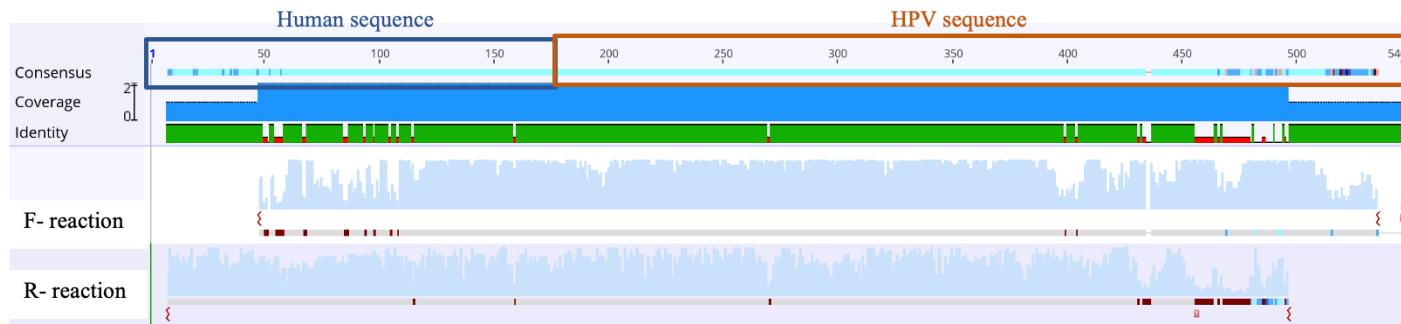
**Supplementary figure 5S:** Shows the chromatogram from Forward(F)- sequence as no continuous sequence was identified. The sequence is mapping to the human chromosome 5 and Human papillomavirus (HPV) 33 as reported from the Next Generation Sequencing (NGS) thereby confirmed. The Reverse(R)- sequence was of low-quality, partly mapping to the human chromosome 5 (chromatogram not shown). The screenshot is obtained from Geneious v2020.2.2.

## Sample ID: 21a



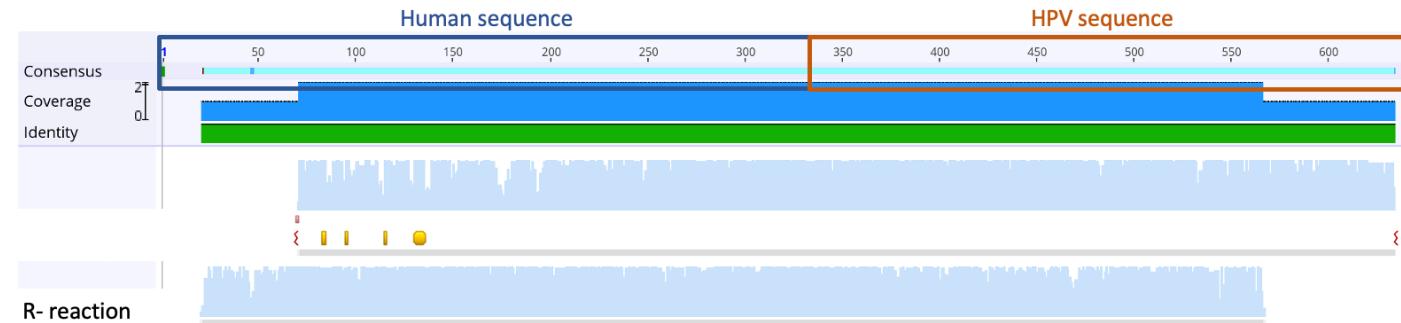
**Supplementary figure 5T:** Shows the continuous Sanger sequence from assembled Forward(F)- and Reverse(R)-sequences with a pairwise identity of 98,4%. The continuous sequence is mapping to the human chromosome 8 and Human papillomavirus (HPV) 45 as reported from the Next Generation Sequencing (NGS) data thereby confirmed. The screenshot is obtained from Geneious v2020.2.2.

Sample ID: 21b



**Supplementary figure 5U:** Shows the continuous Sanger sequence chromatogram from assembled the Forward(F)- and Reverse(R)-sequences with a pairwise identity of 89,2%. The continuous sequence is mapping to the human chromosome 9 and Human papillomavirus (HPV) 45 as reported from the Next generation sequencing (NGS) thereby confirmed. The screenshot is obtained from Geneious v2020.2.2.

Sample ID: 21c

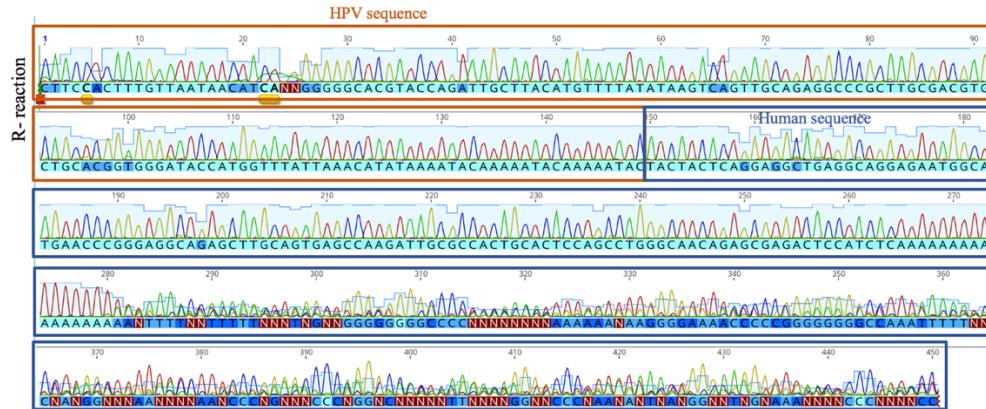


**Supplementary figure 5V:** Shows the continuous Sanger sequence chromatogram from assembled Forward(F)- and Reverse(R)-sequences with a pairwise identity of 100,0%. The continuous sequence is mapping to the human chromosome 8 and Human papillomavirus (HPV) 45 as reported from the Next Generation Sequencing (NGS) thereby confirmed. The screenshot is obtained from Geneious v2020.2.2.

“TTAGTTCCATCTGCGCCGGCGCAGTGGCTACA  
CCTGTAATCCCAGCACTTGGAGGCCGAGGTGGCG  
GATCACAAGTCAGGAGATTGAGACCATCTGGCTAA  
CACGGTAAACCTGTCTACTAAAAATACAAAAAA  
TTAGCCGGCGTGGTGGTGG[NNN]GACATCTTATAT  
AATATAACCCCCAATAGCTAACACATGCTGTTT  
TCCCATATCCCTGTCTCACTTATAATATACTGTGCT  
CATACTACATAGTTCATACAGTTGTCCTTGTGCCATC  
AAAGTATACTGTCACGGTTTACCCCTTTTAAAC  
ACTGCGACGGTTCTGTATTCCATAGTTCCCGCATGTA  
TCTTGAGTGTCCATTCCATTGTTACTTGCTTGT  
GCAAGGCCCTTAAGGCCATTGCAAGTCAATAGCTT  
ATGTGTTGTTGAAATGTTACTAGGAGGCACCA  
CCTGGTGGTTAGTTGGKNA TNCCNTGTCCTGCN  
M”

“CTCACGAATAACACATGGTAGAGACTGTTATCTCCAC  
ATCAGTTCATATTGTGTTCTCAGCTTTGCTAG  
CACACGTGGCTGAGACCATAGTCTGCCTCAGAGA  
GAGGAAATGTCTTAGGAGGGACTCAAGTTAGGTGCT  
AAGAGAGGTCAACAATGGAAAGACTGTTCTGCTG  
GAAGCATGAAGACTTCATGGAGGAGGCACACTGAAC  
TAGCCAGGTTGAAGCCATTACAACAATTCAAGGAC  
TGTAGCATGCGAGCTGGAGACAGCACAGTGTGAA  
GGGAACAGCACACA[NNN]GAGATGTTAATGCTGAAA  
GAAACAAATAGGGGGTAGGGTACATTTCACCATGAT  
ATAAGCCCCATTGCTGCTACAGAGGAATTGAATTGC  
AGCCTTACTTAGTGCTACAGATGATAGTGACCTGTT  
GATGTATATGCAGACTTCCACCTCTGCGTCCACTAC  
ACCTAGCACTAAACAAATCATTACATATCCAAG  
TATTCCCTGACCATGCCTTCACTGCTGCATCCCTTA  
CAGTAATGTTACAGTACCATTAACATCTGCATGGGAT  
GTACCTATATACTGGCCCGGA”

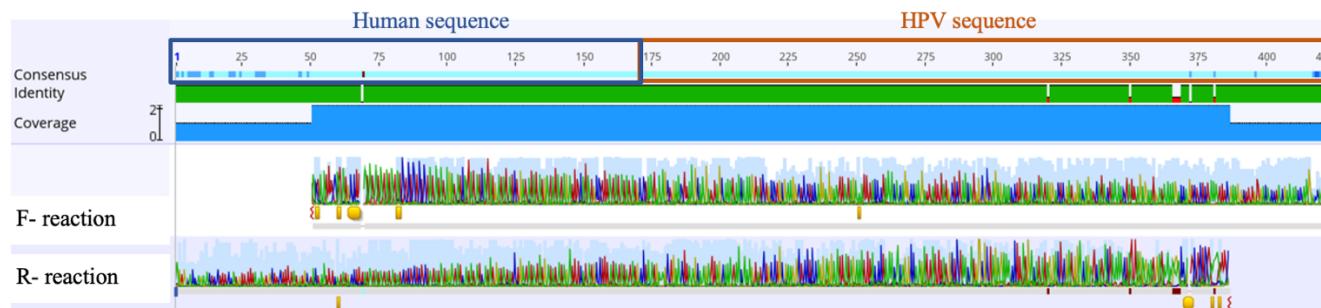
Sample ID: 21d



“CTTCCCACTTGTAAATAACATCANNGGGGCACG TACCAGATTGCTTACATGTTTATATAAGTCAGT GCAGAGGGCCGTTGCGACGTGCTGCACGGTGGG ATACCATGTTTATTAACATATAAAATACAAAAA ATACAAAATAC[NNN]TACTACTCAGGAGGCTGA GGCAGGAGAATGGCATGAACCCGGGAGGCAGAG CTTGCAGTGAAGCAAGATTGCGCCACTGCACTCC AGCCTGGGCAACAGAGCAGACTCCATCTCAAAA AAAAAAAAAAAAAANTTTNNNTTTNNNTNGNN GGGGGGGGCCNNNNNNNNNNNNAAAAAANAAGGGG AAAACCCCCGGGGGGCCAAATTNNNCNANG GNNAANNNNAANCCNGNNCCNGNCNNNN NTNNNNNGNNCCNAANANTNANGNNTNGNA AANNNNCCNNNNCC”

**Supplementary figure 5W:** Shows the chromatogram from the Reverse(R)- sequence in sample 21d as no continuous sequence was identified. The R-sequence is mapping to the human chromosome 9 and Human papillomavirus (HPV) 45 as reported from the Next Generation Sequencing (NGS) data thereby confirmed. The Forward(F)-sequence was of low-quality, partly mapping to the human chromosome 9. The screenshot is obtained from Geneious v2020.2.2.

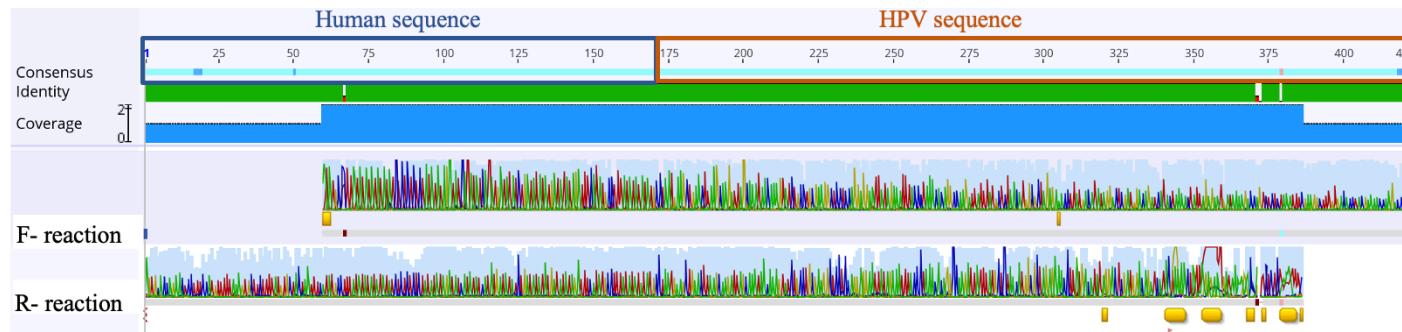
CaSki-cell line -1



“TGAGCTCTGTTACCAAAACCTAGAAAAATTGAG CAACAAATAGAGCATCATTAAACTATAGCCAANT ATATATATATATACACACACATATATGTATAC TATATACTATAGTATATACAGTATATAGTATATA TGTAACATAGCCAATATATATAGCCAT[NNN] TAGTTGCAGTTCAATTGCTTGTAAATGCTTTATCTTT GATACAGCCAGGGTGGCACCCCTGGGTAAATA TGTTAAATCCCATTCTCTGGCCTTGTAAATAATAG CACATTCTAGGCGCATGTGTTCAATAGTCTATAT GGTACACGTAGGTCTGACTATCATTTCATAATGTGT TAGTATTTGTCTGACACACATTAAACGTTGGCA AAGAGTCTCCATCGTTTCCTGTCCTCGA”

**Supplementary figure 5X:** Shows the continuous Sanger sequence from assembled Forward(F)- and Reverse(R)-sequences in CaSki-cell line1 with a pairwise identity of 98,1% obtained when directly sequenced. The continuous sequence is mapping to the human chromosome X and Human papillomavirus (HPV) 16 as excepted. The screenshot is obtained from Geneious v2020.2.2.

## CaSki cell-line-2



"TGAGCTCCTGTTACCAAACCTAGAAAAATTGA  
 GCAACAAAATAGAGCATCATTAACATAGCCAA  
 ATATATATATATACACACACATATATGTA  
 TACTATACTATAGTATACAGTATATAGTA  
 TATATGTAACACTAGCCAATATATATA[NNN  
 JGCCATTAGTTGCAGTTCAATTGCTTGAATGCTT  
 ATTCTTGATACAGCCAGCGTTGGCACCATGCTT  
 GGTAAATATGTTAAATCCCATTCTCTGGCCTTG  
 TAATAATAGCACATTCTAGGCGCATGTGTTCCA  
 ATAGTCTATATGGTCACGTAGGTCTGTACTATNM  
 WTTTCMWAATKKKTWRTATTTGTGYGACAC  
 ACVTTAACGTTGGCAAAGAGTCTCCATCGTTT  
 CCTTGTCTCG"

*Supplementary figure 5Y:* Shows the continuous Sanger sequence from assembled Forward(F)- and Reverse(R)-sequences in CaSki-cell line2 with a pairwise identity of 98,2% obtained when sequenced the gel-elute. The continuous sequence is mapping to the human chromosome X and Human papillomavirus (HPV) 16 as excepted. The screenshot is obtained from Geneious v2020.2.2