

**Viral genomics by next-generation sequencing –
investigating intra-host genomic events in human
papillomavirus and improving intra-hospital
outbreak investigations of SARS-CoV-2**

Alexander Hesselberg Løvestad



PhD programme: Health Sciences
Faculty of Health Sciences (HV)
Department of Life Sciences and Health
OsloMet – Oslo Metropolitan University

Spring 2023

CC-BY-SA versjon 4.0

OsloMet Avhandling 2023 nr 2

ISSN 2535-471X (trykt)
ISSN 2535-5414 (online)

ISBN 978-82-8364-454-8 (trykt)
ISBN 978-82-8364-484-5 (online)

OsloMet – storbyuniversitetet
Universitetsbiblioteket
Skriftserien
St. Olavs plass 4,
0130 Oslo,
Telefon (47) 64 84 90 00

Trykket hos Byråservice

Trykket på Trykket på Scandia 2000 white, 80 gram på materiesider/200 gram på coveret

ACKNOWLEDGEMENTS

The work for this doctoral thesis was carried out at the Department of Life Sciences and Health, Oslo Metropolitan University (OsloMet), the Department of Microbiology and Infection Control, Akershus University Hospital (Ahus) and the Department of Research, Cancer Registry of Norway (KRG). The PhD scholarship was funded by the Faculty of Health Sciences, by a research grant. Additional funding for each study is greatly acknowledged and given in each individual paper. I was part of Health Sciences PhD programme at the Faculty of Health Sciences, OsloMet, that supported the PhD project.

To start off, I would like to thank my three supervisors Ole Herman Ambur (OsloMet), Trine B. Rounge (KRG, University of Oslo) and Irene Kraus Christiansen (Ahus). As a person who had not worked with viruses or cancer previously, your supervision made sure I was able to quickly learn the ropes and get into the world of HPV. You all made me feel welcome at your respective departments and your different backgrounds and skills really complement each other in a way that makes you an excellent trio of supervisors. A big thanks to Sonja Lagström as well, your help in the beginning of my PhD project was very important for me to get started on my project. I would also like to thank Adina Repesa, I don't think anyone could have asked for a better and more hardworking master student to co-supervise their first time.

Jean-Marc Costanzi and Milan Stosic, the newest members of the research group, deserve a big thanks as well. You've made the last half of my PhD a lot more enjoyable by working together on our projects, but mostly on our spare time when we've met to eat great food and drink beer, too fancy drinks and Chartreuse. This appreciation extends to Paula Istvan who have been an important part of these late nights.

I would also like to thank all my previous and current colleagues at the FoU section at Ahus, aNita, Karin, Hanne B, Julie, Gro, Kristiane, Diana, Truls and Chris. Especially Hanne Kristiansen and Mona Hansen, who helped me out a lot when starting out in a new lab environment and with their long experience working with HPV. At the beginning of 2020, when the SARS-CoV-2 pandemic hit Norway, I was approached by Hege Vangstein Aamot (Ahus) and Ole Herman Ambur and asked if I wanted to help establish whole genome sequencing and variant analysis of SARS-CoV-2 at Ahus. I said yes, and suddenly my PhD project changed quite a bit. I am very grateful that I could be a part of it, and I learned a lot I would not have learnt any other way. It was immensely fun to feel that I contributed to the extreme situation that we suddenly found ourselves in, so thank you for including me.

I would also like to thank my colleagues at the Cancer Registry of Norway for bioinformatic help and good talks as well as NORBIS national research school in bioinformatics, biostatistics and systems biology for allowing me to take their courses and expanding my knowledge and competence.

Lastly, I would like to thank my family and friends for always being there and supporting me. A special thanks to Mette Årslund who supplied me with tasty dinners, great supporting company and a place to sleep when my apartment building became bedbug-infested close to my thesis deadline.

ABSTRACT

Human Papillomaviruses (HPVs) are one of the oldest human pathogens and the most common sexually transmitted pathogenic infection worldwide. Most HPV infections are cleared by the immune system, but some infections persist and can progress to HPV-induced cancer. Almost all cervical cancer cases are caused by persistent infections with high-risk HPVs, affecting more than 500.000 women worldwide and causing more than 250.000 deaths per year. Compared to HPVs, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is one of the most novel human pathogens, making the jump to human populations late in 2019. Since then, SARS-CoV-2 have spread globally and caused the deaths of more than six million people. While the histories of these viruses differ substantially, they have in common their immense impact on global health.

One of the most important tools at our disposal to defend ourselves against viral pathogens is next-generation sequencing (NGS). NGS allows us to investigate genomic events affecting intra-host viral populations found within infected persons and how they contribute to disease severity. It also allows for the rapid retrieval of viral genomic information that can be used to understand and track transmissions of pathogens.

In this thesis, NGS is applied to cervical cell samples positive for high-risk HPVs to study how viral intra-host genomic events can contribute to infection persistence and progression to cervical cancer. In total, five high-risk HPV types responsible for ~90% of all cervical cancer cases are investigated with a focus on intra-host minor nucleotide variation (MNV) and integration into human chromosomes. The results show differences between the HPV types, and that these differences extend to the closely related HPV types. Overall, the studies shed light on molecular differences between the HPV types that can reflect type-specific mechanistic routes of HPV-induced cancers, while also presenting much needed knowledge of the lesser studied high-risk HPV types.

Additionally, NGS was applied to SARS-CoV-2 positive samples from healthcare workers and patients from Akershus University Hospital to increase the resolution of outbreak investigations. When genomic information was used in combination with contact tracing data, one suspected intra-hospital outbreak was refuted, and another potential outbreak was discovered. The study shows the benefit of including viral whole genome sequencing data when doing outbreak investigations.

The thesis highlights the power of NGS to understand viral pathogens, be it viruses we have had a shared history with since time immemorial or novel viruses we only recently encountered.

SAMMENDRAG

Humant papillomavirus (HPV) er en av de eldste sykdomsfremkallende virus hos mennesker og verdens vanligste seksuelt overførbare infeksjon som forårsaker sykdom. De fleste HPV-infeksjoner blir klarert av immunforsvaret, men et fåtall blir persistente og kan føre til HPV-indusert kreft. Nesten alle tilfeller av livmorhalskreft skyldes persistente HPV-infeksjoner, noe påvirker mer enn 500.000 kvinner globalt og fører til mer enn 250.000 dødsfall årlig. På den annen side ble severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) introdusert til menneskelige populasjoner sent 2019, og er med det en av de nyeste sykdomsfremkallende virusene. Siden da har SARS-CoV-2 spredt seg globalt og forårsaket mer enn seks millioner dødsfall. Disse virusene har vidt forskjellige historier som sykdomsfremkallende virus hos mennesker, men de har til felles deres enorme byrde på den globale helsen.

Et av våre viktigste verktøy for å beskytte oss mot patogene virus er neste generasjons sekvenseringsteknologi (NGS). NGS lar oss undersøke genomiske hendelser som påvirker virale populasjoner innad i infiserte individer og hvordan disse påvirker forløp og alvorlighetsgrad av infeksjonene. Det lar oss også raskt tilegne oss kunnskap om de virale genomene og deres tendens til endring som kan brukes til å spore virusmitte.

I denne avhandlingen brukes NGS på livmorshalsprøver positive for høyrisiko HPV for å studere disse genomiske hendelsene innad i en vert og hvordan dette kan bidra til persistens og utvikling av livmorhalskreft. Totalt fem høyrisiko HPV-typer som til sammen forårsaker ~90% av alle tilfeller av livmorhalskreft har blitt undersøkt med et fokus på lavforekomstmutasjoner og integrasjoner i menneskekromosomer i enkeltindivider. Resultatene viser at det er forskjeller mellom HPV-typene og at disse forskjellene er konservert mellom nært beslektede HPV-typer. Studiene kaster lys over molekylære forskjeller mellom HPV-typene som kan reflektere type-spesifikke mekanistiske veier mot å utvikle livmorhalskreft, og presenterer også ny kunnskap om mindre studerte høyrisiko HPV-typer.

I tillegg ble NGS brukt på SARS-CoV-2 positive prøver fra helsearbeidere og pasienter fra Akershus Universitetssykehus for å øke oppløsningen på utbruddsoppløring. Når genomisk informasjon fra viruset ble brukt i kombinasjon med tradisjonell smittesporing. Av fem mistenkte utbrudd ble et utbrudd avkreftet, og et helt nytt mulig utbrudd ble oppdaget. Studiet viser fordelene ved å inkludere viral helgenomsekvensering når man gjør utbruddsoppløring.

Avhandlingen fremhever hvor kraftig NGS er som verktøy for å forstå patogene virus, enten det er virus vi har delt historie med siden urtiden eller helt nye virus vi aldri har møtt før.

TABLE OF CONTENTS

| | |
|---|----|
| LIST OF ABBREVIATIONS..... | 3 |
| LIST OF ORIGINAL PAPERS | 5 |
| 1. INTRODUCTION..... | 7 |
| 1.1 Old and new viruses | 7 |
| 1.2 Molecular biology of HPV and SARS-CoV-2 | 8 |
| 1.2.1 <i>Genome structure of HPV</i> | 8 |
| 1.2.2 <i>Genome structure of SARS-CoV-2</i> | 10 |
| 1.2.3 <i>HPV variants</i> | 12 |
| 1.2.4 <i>SARS-CoV-2 variants</i> | 14 |
| 1.2.5 <i>Intra-host variation</i> | 18 |
| 1.3 Life cycle and pathogenicity/pathogenesis | 20 |
| 1.3.1 <i>HPV life cycle and pathogenicity</i> | 20 |
| 1.3.2 <i>SARS-CoV-2 life cycle and pathology</i> | 23 |
| 1.4 Prevention and treatment | 25 |
| 1.4.1 <i>HPV vaccination, screening and treatment</i> | 25 |
| 1.4.2 <i>SARS-CoV-2 vaccination, preventive measures, and treatment</i> | 27 |
| 1.5 Molecular mechanisms of HPV-induced cancer | 29 |
| 1.6 Viral surveillance and genomic epidemiology of SARS-CoV-2 | 32 |
| 2. THESIS AIMS | 35 |
| 3. MATERIALS AND METHODS | 36 |
| 3.1 Sample material and study design | 36 |
| 3.2 DNA/RNA extraction and HPV genotyping | 37 |
| 3.3 DNA concentration and Ct-value | 37 |
| 3.4 Library preparation and sequencing | 37 |
| 3.5 Study I and II sequencing data analysis | 39 |
| 3.5.1 <i>Sequence alignment</i> | 39 |
| 3.5.2 <i>Sequence variation analysis</i> | 40 |
| 3.5.3 <i>Mutational signature analysis</i> | 40 |
| 3.5.4 <i>Detection of integration sites and deletions</i> | 40 |
| 3.5.5 <i>Validation of integration sites</i> | 41 |
| 3.6 Study III sequencing data analysis | 41 |
| 3.6.1 <i>Sequencing analysis of SARS-CoV-2 amplicon data</i> | 41 |
| 3.6.2 <i>Phylogenetic analysis, Nextstrain Clade assortment and pangolin lineage assignment</i> | 42 |

| | |
|--|----|
| 3.6.3 <i>Outbreak assessment</i> | 42 |
| 3.7 Statistical analyses | 43 |
| 3.8 Ethical aspects | 43 |
| 4. SUMMARY OF RESULTS | 44 |
| 4.1 Study I | 44 |
| 4.2 Study II | 45 |
| 4.3 Study III | 46 |
| 5. DISCUSSION | 47 |
| 5.1 Methodological considerations | 47 |
| 5.1.1 <i>Sample material</i> | 47 |
| 5.1.2 <i>Library preparation and sequencing</i> | 48 |
| 5.1.3 <i>Bioinformatic analyses</i> | 49 |
| 5.1.4 <i>Statistics</i> | 50 |
| 5.2 Discussion of results | 51 |
| 5.2.1 <i>HPV intra-host variation and integration frequencies</i> | 51 |
| 5.2.2 <i>Nanopore whole genome sequencing of SARS-CoV-2 to investigate intra-hospital transmission</i> | 53 |
| 5.3 Significance of results and future perspectives | 55 |
| 6. CONCLUSIONS | 58 |
| 7. REFERENCES | 60 |
| 8. PAPERS I-III | 94 |

LIST OF ABBREVIATIONS

| | |
|--------|--|
| ACE2 | Angiotensin-converting enzyme 2 |
| AIS | Adenocarcinoma <i>in situ</i> |
| APOBEC | Apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like |
| ARDS | Acute respiratory distress syndrome |
| ASC-H | Atypical squamous cells, cannot exclude high-grade lesion |
| ASC-US | Atypical squamous cells of undetermined significance |
| BAM | Binary Alignment/Map |
| cDNA | Complimentary DNA |
| CIN | Cervical intraepithelial neoplasia |
| CRG | Cancer-related gene |
| Ct | Cycle threshold |
| ER | Endoplasmic reticulum |
| ERGIC | ER-to-Golgi intermediate compartments |
| GISAID | Global Initiative on Sharing All Influenza Data |
| GLM | Generalized linear model |
| gRNA | Genomic RNA |
| HCW | Healthcare worker |
| HPV | Human papillomavirus |
| HR-HPV | High-risk HPV |
| HSIL | High-grade squamous intraepithelial lesion |
| HSPG | Heparin sulphate proteoglycans |
| IARC | International Agency for Research on Cancer |
| Indel | Insertion or deletion |
| KB | Kilobase |
| LBC | Liquid-based cytology |
| LEEP | Loop electrosurgical excision procedure |
| LR-HPV | Low-risk HPV |
| LSIL | Low-grade squamous intraepithelial lesion |
| MERS | Middle East respiratory syndrome |
| mRNA | Messenger RNA |
| MNV | Minor nucleotide variant |
| MSA | Multiple sequencing alignment |

| | |
|------------|---|
| NCR | Non-coding region |
| NGS | Next generation sequencing |
| NSP | Non-structural protein |
| ORF | Open reading frame |
| ORI | Origin of replication |
| PANGO | Phylogenetic Assignment of Named Global Outbreak |
| PaVE | Papillomavirus Episteme |
| PCR | Polymerase chain reaction |
| QC | Quality control |
| qPCR | Quantitative polymerase chain reaction |
| RB | Retinoblastoma protein |
| RBD | Receptor binding domain |
| RTC | Replication-transcription complexes |
| SAM | Sequence Alignment/Map |
| SARS-CoV-2 | Severe acute respiratory syndrome coronavirus 2 |
| SNP | Single-nucleotide polymorphism |
| sgRNA | Subgenomic RNA |
| sg-mRNA | Subgenomic messenger RNA |
| TaME-seq | Tagmentation-assisted multiplex PCR enrichment sequencing |
| TMPRSS2 | Transmembrane serine protease 2 |
| UK | United Kingdom |
| URR | Upstream regulatory region |
| VLP | Virus-like particle |
| VOC | Variant of concern |
| VOI | Variant of interest |
| WGS | Whole genome sequencing |
| WHO | World Health Organization |

LIST OF ORIGINAL PAPERS

- I. Lagström S, **Hesselberg Løvestad A**, Umu SU, Ambur OH, Nygård M, Rounge TB, Christiansen IK. HPV16 and HPV18 type-specific APOBEC3 and integration profiles in different diagnostic categories of cervical samples. *Tumour Virus Research* 2021;12:200221. DOI: <https://doi.org/10.1016/j.tvr.2021.200221>.
- II. **Hesselberg Løvestad A**, Repesa A, Costanzi JM, Lagström S, Christiansen IK, Rounge TB, Ambur OH. Differences in integration frequencies and APOBEC3 profiles of five high-risk HPV types adheres to phylogeny. *Tumour Virus Research* 2022;14:200247. DOI: <https://doi.org/10.1016/j.tvr.2022.200247>.
- III. **Hesselberg Løvestad A**, Jørgensen SB, Handal N, Ambur OH, Aamot HV. Investigation of intra-hospital SARS-CoV-2 transmission using nanopore whole-genome sequencing. *Journal of Hospital Infection* 2021;111:107–16. DOI: <https://doi.org/10.1016/J.JHIN.2021.02.022>.

1. INTRODUCTION

1.1 Old and new viruses

Viruses are ubiquitous and infect every living organism of the biosphere, and therefore also humans. Humans and viruses share histories going back to time immemorial and have co-evolved. Thus, viruses have played an important role in shaping human evolution and vice versa.

Papillomaviruses are a group of ancient viruses known to infect numerous vertebrates, from fishes to humans[1]. Human papillomavirus (HPV) infections are the most common sexually transmitted infection and 70% of sexually active persons are estimated to acquire HPV during their lifetime[2,3]. While most HPV infections are benign and cleared by the immune system in 6-24 months, a small percentage of infections persist, potentially lasting decades[4,5]. These persistent infections are considered a necessary cause to develop HPV-induced cancer[6]. HPV-induced cancers place an immense disease burden on global health, representing nearly 5% of all cancers worldwide[7]. Almost all cervical cancers are caused by persistent HPV infections, which affect more than 500.000 women worldwide and cause 266.000 deaths per year[6,8]. Additionally, HPV is associated with a significant proportion of oropharyngeal cancer and cancer in anogenital regions, including penile, vaginal, vulvar, and anal cancers[7,9]. This disease burden is not equally shared among sexes or geographical regions. In total, 8.6% of all cancers in women have HPV as the causative agent, while this number is only 0.8% for men. Additionally, 70% of all the cervical cancer cases occur in less developed countries and account for >85% of all cervical cancer deaths[7].

While their impact on global health is huge, HPVs are not new human pathogens. They are, in fact, one of the oldest human pathogens, thought to have infected ancestral human populations more than 500 thousand years ago[10]. To put that in perspective, the oldest known fossil of modern humans is roughly 315 thousand years old[11].

At the other end of the virus to human infection spectrum, we have severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). SARS-CoV-2 was first discovered in Wuhan, China, in December 2019, which makes it one of the most novel human pathogens. It entered the human population through zoonotic spillover events, most likely from wet markets where live

animals were kept and sold [12,13]. SARS-CoV-2 has been classified as a species of Severe acute respiratory syndrome-related coronavirus in the genus Betacoronavirus of the family *Coronaviridae*[14]. The Betacoronavirus genus seems to have coevolved with bats over tens of millions years, exclusively infects mammals and includes endemic human common cold viruses[15,16], SARS, which had an international outbreak in 2003[17], and Middle East respiratory syndrome (MERS), first detected in 2012, with sporadic outbreaks since then[18,19]. Since the introduction of SARS-CoV-2 into human populations, the virus has spread globally, and with over half a billion cases and over 6.300.000 deaths[20], its global impact has been enormous.

The history of the HPVs and SARS-CoV-2 with humans differ substantially, but they have in common their enormous effect on the global human population. With the rapid development of next-generation sequencing technology, we have acquired new tools to gain insight that can be used to defend and protect ourselves from viral pathogens, both old and new.

1.2 Molecular biology of HPV and SARS-CoV-2

1.2.1 Genome structure of HPV

HPVs are a group of viruses with small, circular double-stranded DNA genomes ~7.9 kb in length[21]. The genome consists of eight open reading frames (ORFs) and two non-coding regions. The ORFs can be split into early (E) region genes (E1, E2, E4-E7) and late region genes (L1, L2), and the two non-coding regions are labelled the upstream regulatory region (URR) and non-coding region (NCR) (Figure 1) [22,23]. The division of early and late region genes is based on the life cycle stages when they are expressed, where the early region genes E1, E2, E4, E5, E6, and E7 are expressed early in HPV life cycle and the late region genes L1 and L2 are expressed towards the end[22].

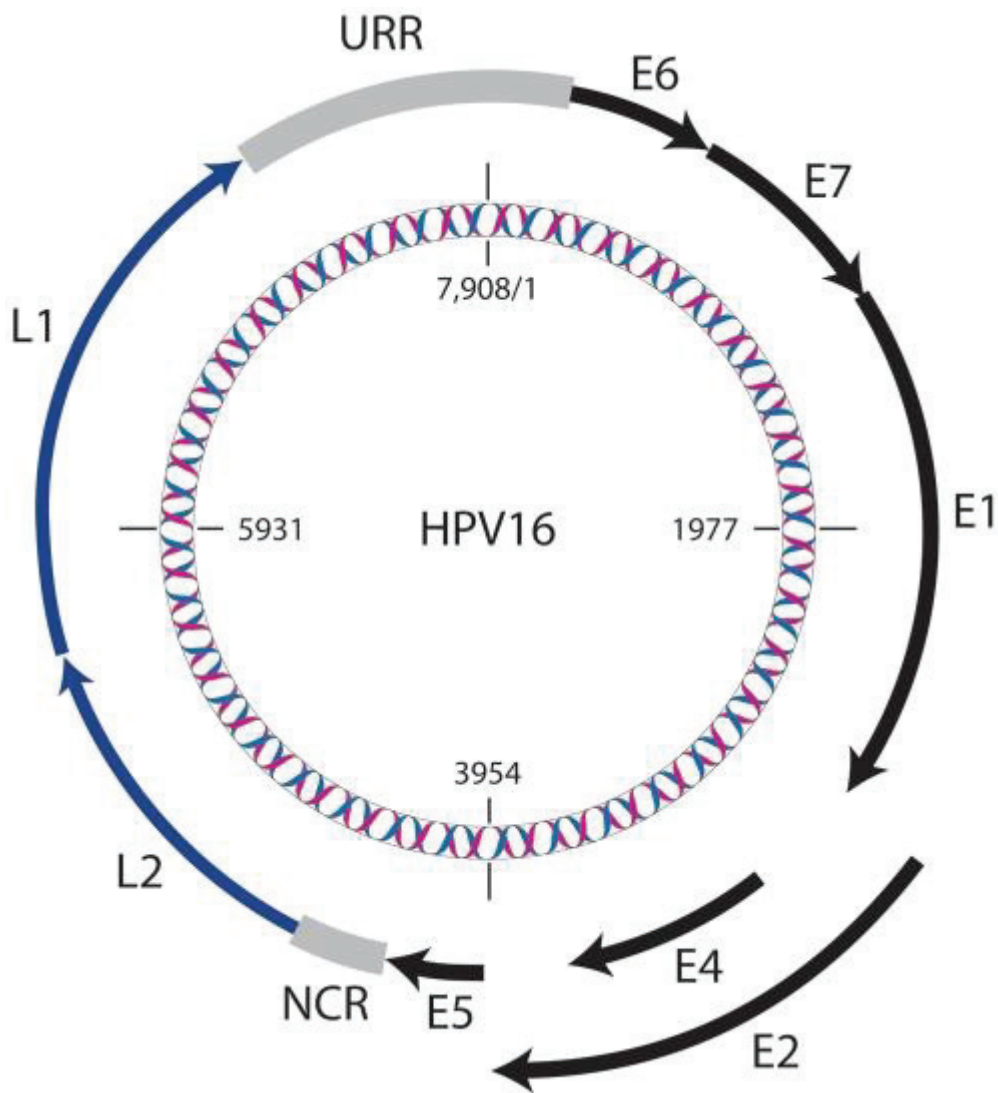


Figure 1: Schematic representation of an HPV genome, here exemplified using HPV16. Early (E) genes coloured black, late (L) genes coloured blue and the Upstream Regulatory Region (URR) and Non-coding Region (NCR) coloured grey. From [24]. Copyright by Smith, Chen, Reimers, van Doorslaer, Schiffman, DeSalle, Herrero, Yu, Wacholder, Wang and Burk. Printed under CC BY 4.0 licencing, <https://creativecommons.org/licenses/by/4.0/>.

Early genes encode non-structural proteins that are involved in virus replication, regulation of viral transcription, immune evasion and modifying the cellular environment to serve the needs of the virus[25]. E1 and E2 are involved in viral replication, where E1 acts as an origin recognition factor and helicase, which is recruited by E2 to the viral origin of replication (ORI) [26]. E2 also functions as a key negative regulator of early viral gene expression, which includes viral oncogenes E6 and E7[27]. The E4 gene is involved in viral genome amplification and

virus synthesis and is also suggested to play additional roles in virus release and/or transmission[28]. E5 is a small protein, whose functions are poorly understood. It does however have oncogenic potential and has been shown to be involved in cell transformation, tumorigenesis and immune modulation[29]. E6 and E7 contribute to the HPV life cycle by driving cell cycle re-entry to allow viral genome amplification in the lower and middle layers of the epithelium, stimulating cell proliferation, repressing tumour-suppressor mechanisms and inhibiting aspects of innate immunity, and are considered to be key factors in HPV-induced malignant cell transformation[30].

The late region ORFs L1 and L2 encodes two structural proteins and together they form the virus capsid[31]. The URR is located immediately upstream of the early region and contains the viral ORI, several transcription factor binding sites and regulates early gene expression[30,32]. NCR is located between the genes E5 and L2, and while little is known about its function, it has a weak promoter activity for the L2 gene[33,34]

1.2.2 Genome structure of SARS-CoV-2

While HPVs have small, circular double stranded DNA genomes, SARS-CoV-2 has a large positive sense single stranded RNA genome of approximately 30 kb length, with a 5'-cap structure and 3'-polyA tail (Figure 2). The genome size of 30 kb makes it among the RNA viruses with the largest genome size. The 5'-cap and poly-A tail allows the genome to perform as an mRNA for translation of polyproteins used in viral replication[35]. The first two thirds of the genome from the 5' end consists of two overlapping ORFs, ORF1a and ORF1b, with ORF1b being translated in a -1 ribosomal frameshift. Together, these two ORFs are translated and cleaved into 16 non-structural proteins that are required for genome replication and transcription. Towards the 3' terminal end of the genome, SARS-CoV-2 encodes four structural proteins, and in addition between six and eleven accessory proteins (depending on the literature)[36–39]. The structural proteins are necessary for the virion to form and consists of the nucleocapsid (N), envelope (E), membrane (M) proteins and spike (S) glycoprotein. The N protein functions in the structural organization of the nucleocapsid by binding the RNA genome into a helix and is involved in the viral replication[35]. The E protein is a small ion channel transmembrane protein which plays an important role in the assembly and release of the virus as well as its virulence[35,40]. The M protein, the most abundant structural protein, is important

for the assembly of the virus particle, the structure of the virus envelope and stabilising the N protein-RNA complex [41].

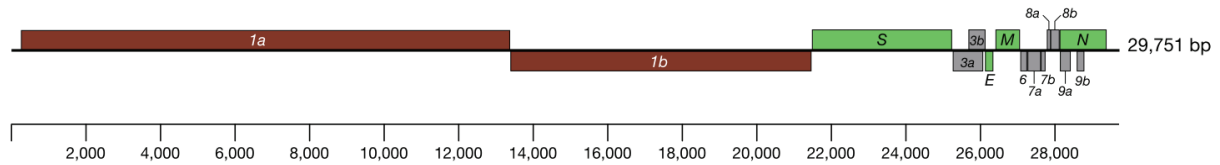


Figure 2: Schematic representation of the SARS-CoV-2 genome. Genes encoding non-structural proteins coloured in red, structural proteins in green and accessory proteins in grey. Modified from [42]. Copyright by Wu, Zhao, Yu, Chen, Wang, Song, Hu, Tao, Tian, Pei, Yuan, Zhang, Dai, Liu, Wang, Zheng, Xu, Holmes, Zhang. Printed under CC BY 4.0 licencing, <https://creativecommons.org/licenses/by/4.0/>.

The most studied structural protein is the S protein due to its importance in the viral binding and entry into host cells by attraction to angiotensin-converting enzyme 2 (ACE2), which is expressed on the surface of respiratory tract cells. ACE2 is also expressed on the cell surface of cells in other tissues and organs, including heart, kidneys and colon, which explains why patients infected with SARS-CoV-2 also experience disorders other than respiratory problems[41,43,44]. The spike protein is a transmembrane protein located on the outer portions of the virus and consists of two subunits, S1 and S2, separated by the S1-S2 cleave site which is cleaved by the host cell furin-like protease. The S1 subunit contains the receptor binding domain (RBD) and is responsible for determining the host virus range while the S2 subunit functions to mediate virus fusion and entry into host cells[41].

Less is known about the accessory proteins, but they have been implicated to play important roles in immunoevasive activities and viral pathogenesis by impairing and suppressing host antiviral responses[36,39].

1.2.3 HPV variants

HPVs are a large and diverse group of viruses, with more than 200 HPV genotypes identified[45], infecting mucosal and cutaneous epithelial tissues[46]. The L1 gene has commonly been used as a yardstick to define HPV types, where types are distinguished by at least 10% nucleotide difference in this gene[47,48]. Genotypes are further divided into lineages (1>10% whole-genome nucleotide difference) and sublineages (0.5>1% difference)[49,50]. Known HPV types are all assorted to one of five major genera: alpha-, beta-, gamma, mu- or nu-papillomavirus[50]. Of the more than 200 HPV types characterized, at least 12 HPV types are categorised as high-risk (HPV16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, and 59) and have been classified as carcinogenic to humans according to the International Agency for Research on Cancer (IARC)[7]. An additional eight HPV types (HPV26, 53, 66, 67, 68, 70, 73, and 82) are considered as probably or possibly carcinogenic by the IARC working group[7]. High-risk HPV (HR-HPV) types 16 and 18 are associated with ~70% of cervical cancer cases, while low-risk (LR-HPV) types, including HPV6 and 11 cause benign diseases such as genital warts. All HR-HPVs belong to the clade *Alphapapillomavirus* (Alpha-PV), where they are assorted to the different subclades, Alpha-5, Alpha-6, Alpha-7, Alpha-9, and Alpha-11 (Figure 3)[1,51]. HR-HPVs, both within and between different subclades, exhibit differences in their carcinogenicity and assumed cell tropism[52,53]. Furthermore, even within HR-HPV types, different sublineages have been shown associated with different risks of persistence and progression to cancer[54–58]. This suggests that different evolutionary histories have given rise to differences in carcinogenic potential and the molecular mechanisms behind why some HPV infections progress more often to invasive cancer than others[59].

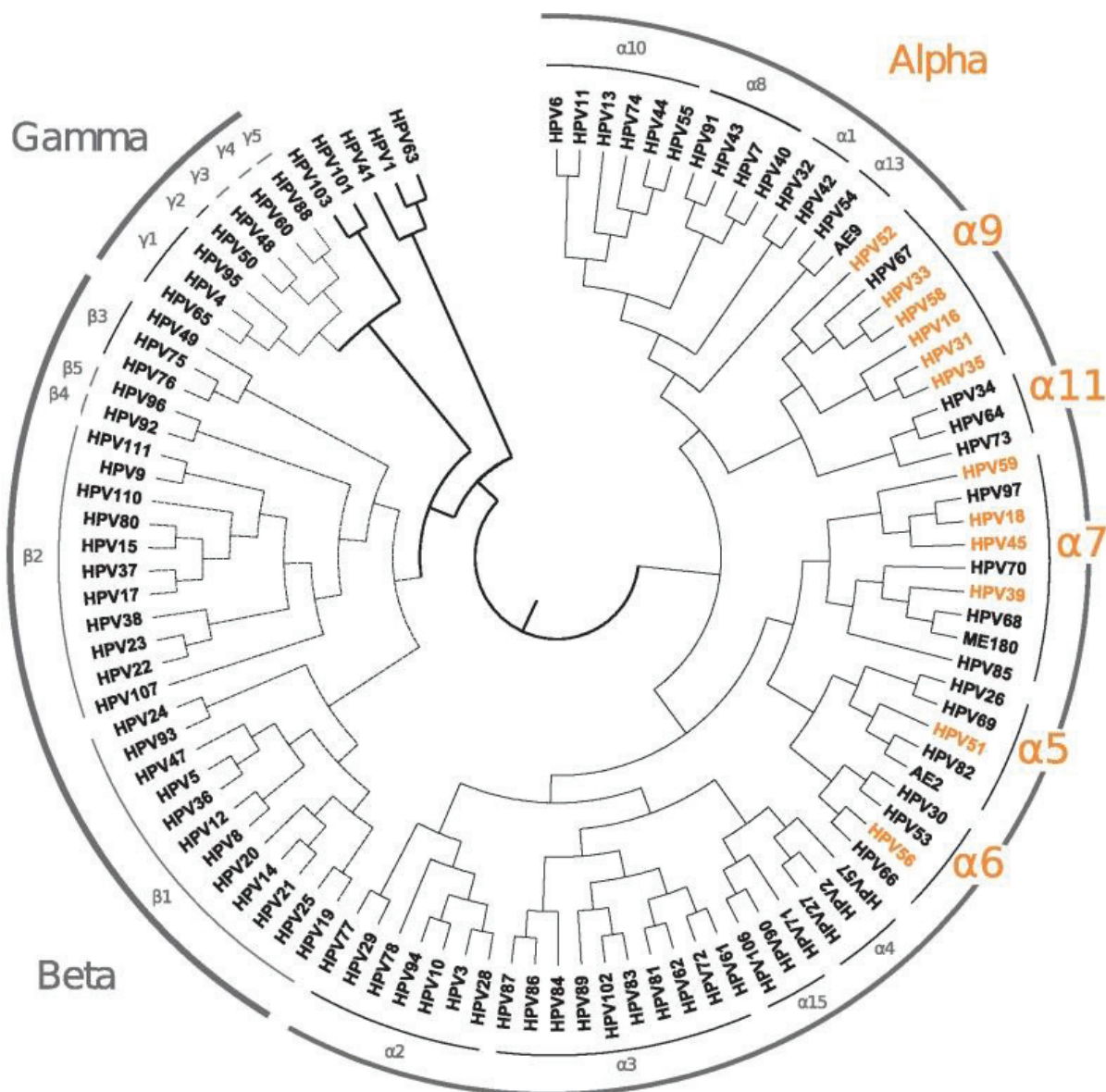


Figure 3: Phylogenetic tree of 100 human papillomavirus types based on the E7, E1, E2, L2 and L1 genes. Alpha-clades containing HR-HPV types are highlighted yellow. From [60]. Copyright 2012 by International Agency for Research on Cancer.

HPV types and their sublineages are not equally distributed worldwide, some types are more prevalent in some geographical regions than others. Owing to their ancient history, distribution of viral sublineages of HPV16 and HPV58 has been found to mirror the migration and dispersal of human prehistoric populations out of Africa, and some extant sublineages have been proposed introduced to the human population by interbreeding with archaic hominins (Figure 4)[10,61]. Geographic distribution of HPV16 variant lineages shows a higher diversity on the African continent and their carcinogenic potential differs in different human populations, while

HPV58 is found in 10-18% of cervical cancers in East Asia and uncommon in other geographical regions[10,54,61]. The differences in carcinogenic potential between different lineages might indicate differences in adaptations to host immune-systems that differ between individuals of different ethnicities[56]. While different sublineages of HPV16 and HPV58 has been shown to have different risks of cervical cancer associated with them, this is not the case for HPV18[55,62].

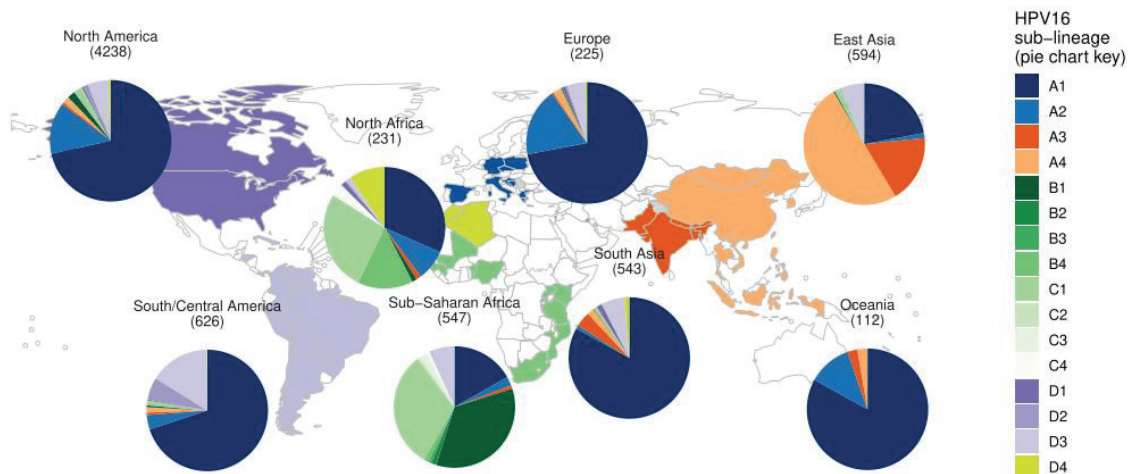


Figure 4: Distribution of HPV16 sublineages by geographic region. The figure is based upon 7116 HPV-16 positive samples. Modified from [56]. Copyright 2019 Elsevier. Printed under CC BY 4.0 licencing, <https://creativecommons.org/licenses/by/4.0/>.

1.2.4 SARS-CoV-2 variants

Since the emergence of SARS-CoV-2 into the human population, it has rapidly diversified and spread, leading to several variants. Several classification systems have been developed to classify SARS-CoV-2 variants, with the most commonly used being the Nextstrain (<https://covariants.org/>) and Phylogenetic Assignment of Named Global Outbreak (PANGO)[63] lineage classification systems. The Nextstrain classification system uses a year-letter nomenclature to label major SARS-CoV-2 clades based on the year they emerged (<https://nextstrain.org/blog/2022-04-29-SARS-CoV-2-clade-naming-2022>), e.g. 19A, 19B, 20A, etc.[64]. For a clade to be named according to the Nextstrain nomenclature, the following criteria must be fulfilled: 1) a clade reaches >20% global frequency for 2 or more months, 2) a clade reaches >30% regional frequency for 2 or more months, 3) a variant of concern (VOC) is recognized by the World Health Organization (WHO), or 4) a clade shows consistent >0.05 per

day growth in frequency where it is circulating and has reached >5% regional frequency. Thus, the Nextstrain classification system provides a long-term overview of the larger scale evolution and diversity of SARS-CoV-2 (Figure 5).

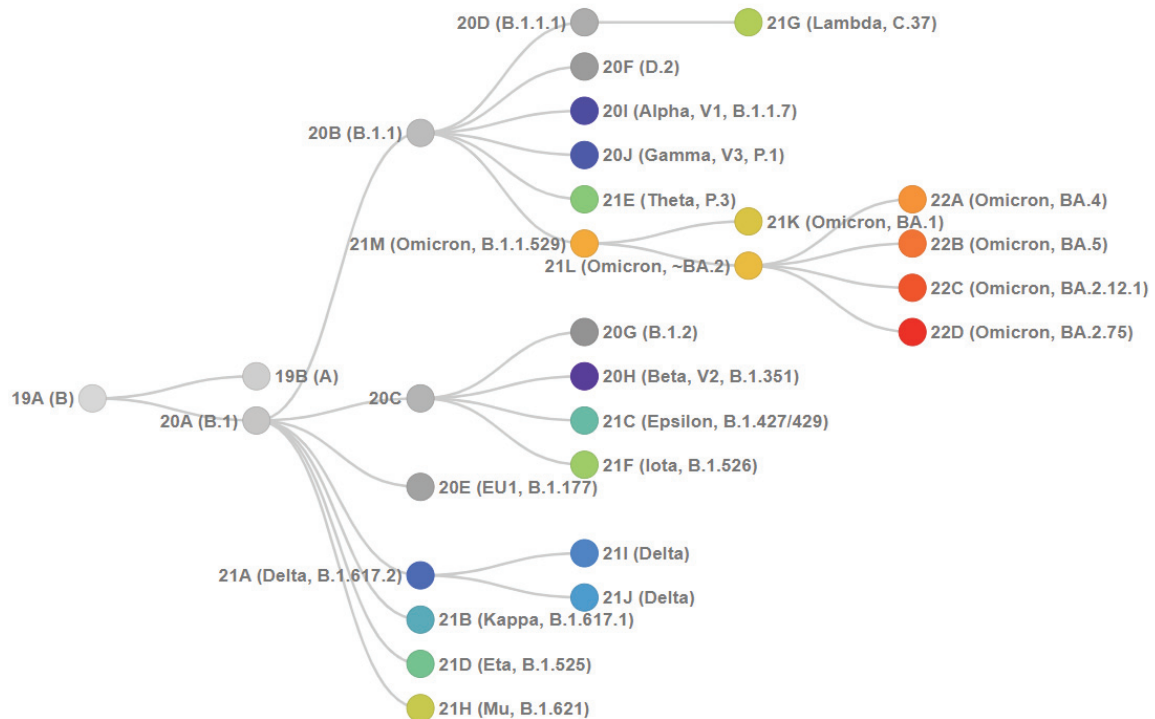


Figure 5: Phylogenetic tree illustrating the hierarchical relationship among SARS-CoV-2 clades according to the Nextstrain classification system. From <https://github.com/nextstrain/ncov-clades-schema>. Printed under CC BY 4.0 licencing, <https://creativecommons.org/licenses/by/4.0/>.

On the other hand, we have the PANGO lineage nomenclature, a hierarchical system containing an alphabetical prefix, followed by up to three number separated by periods to indicate sublineages. (e.g., B.1.1.7). PANGO allows for a higher resolution and specificity, and is therefore the most used classification system of the two[65]. PANGO nomenclature labels transient lineages with local epidemiological significance, which results in many short-lived labels that allows for a more short-term overview[63].

Additionally, we have the classification of variants of concern (VOC) and variants of interest (VOI), labels assigned by the WHO, which follows the Greek alphabet[66]. VOCs are defined as widespread variants that have displayed evidence of the following: 1) increased transmission, 2) increased virulence and/or 3) reduced the effect of public health and social measures or

immunization from vaccination or previous infections or available diagnostics and therapeutics[65,67]. VOIs are variants that contain several mutations similar to those found in VOCs, but that have not spread widely[65,66].

SARS-CoV-2, being a novel human pathogen, have shown several examples of adaptations to its new host, and the evolution and emergence of new SARS-CoV-2 variants has had a profound impact on the dynamics of the pandemic. The earliest known host adaptation is the amino acid replacement D614G in the spike protein to which a transmission advantage has been indicated, and by April 2020 had become the dominant variant[68]. After that the virus evolved relatively slowly for an RNA virus for most of 2020, until the emergence of the first VOCs in late 2020 heralded a new phase of the pandemic[69–71].

In December 2020, the United Kingdom (UK) reported on a novel SARS-CoV-2 variant containing a high number of mutations relative to previous and co-circulating variants, eight of which were found in the spike protein[72]. This novel variant had over a few months grown in frequency to become the dominant circulating variant in the UK[73]. This novel variant was classified as the first VOC by the WHO and named Alpha (20I/B.1.1.7)[66]. Alpha quickly spread and started a new global wave of SARS-CoV-2 infections[73]. After the discovery of the first VOC, it took less than a month until two additional highly mutated variants were discovered that quickly increased in frequency in South Africa and Brazil [74,75]. The two new variants were also classified as VOCs and named Beta (20H/B.1.351) and Gamma (20J/P.1). All three of these VOCs had a substantial number of mutations found within the spike protein, including its S1 receptor binding domain (RBD). In fact, the VOCs shared several mutations, most notably of these N501Y (shared between all three), K417T/N (found in Beta (N) and Gamma (T)) and E484K (found in Beta and Gamma initially, later acquired by Alpha as well) located in RBD[76]. All three also share a nine-nucleotide deletion in ORF1ab, in the portion coding for nsp6[76]. These mutations have all been implicated in increasing transmissibility and in immune escape, thus indicating that these convergently acquired mutations are the result of natural selection acting on the virus in a population with increasing immunity, either through vaccination or previous infections[76].

The global Alpha-wave was eventually displaced by a new VOC during 2021. Delta (21A/B.1.617.2), was first discovered in India late 2020, where it quickly became the dominant variant before spreading globally[77]. The Delta variant displayed several mutations in the spike protein, including several in the RBD, which conferred increased transmissibility and reduced effect of vaccination and immunization through vaccination or previous infection[78].

Delta outcompeted previous VOCs and was the globally dominant variant for a while, until a new variant with a hypermutated spike protein was almost simultaneously discovered in Botswana, Hong Kong and South Africa in November 2021 and then rapidly in many other countries[79]. This novel variant was declared a VOC by the WHO on November 26, 2021, and named Omicron (21K/ B.1.1.529)[80,81]. The Omicron variant contains significantly more mutations than previous VOCs, and they include one insertion, seven deletions, 45 nonsynonymous mutations and 10 synonymous mutations, with many of the nonsynonymous mutations found in the spike protein and its RBD[82]. Of these mutations found in the spike protein, many are novel, while others have been found in previous VOCs, including K417N, N501Y, P681, and deletion at position 69-70[82]. Omicron has shown an even higher transmissibility and immune escape abilities than previous VOCs. However, infection with Omicron has been shown to result in less cases of severe illness, which is hypothesised to be due to a shift in tropism and increased affinity for infecting cells from the upper respiratory tract compared to the lower[83,84]. Omicron has diversified into several lineages (BA.1-5), with BA.2 having been the most dominant variant, but is now (July, 2022) being displaced by BA.2.12.1, BA.4 and BA.5, which all show a further ability to escape neutralizing antibodies, even from past infections with Omicron[85,86].

There are currently several explanations as to how SARS-CoV-2 variants with highly mutated genomes arise. One of the explanations is that chronic infections within immunocompromised persons may give rise to novel variants. Prolonged infection allows for more rounds of viral replication and time for selection to act on the genome, thus explaining the high number of mutations[87]. Several studies have investigated the intra-host variation of SARS-CoV-2 in immunocompromised persons and have discovered that a high number of mutations can be found within the spike protein, including independent acquisition of mutations found in VOCs in several individuals, implying convergent evolution from similar selection pressures[88–91]. Another intra-host event that can result in new variants is recombination between two SARS-CoV-2 variants coinfecting the same host. Several instances of recombination have been reported, including potential Delta-Omicron recombinants[92–94]. The recombination frequency has only been revealed in the later stages of the pandemic with increasing global genomic diversity[93,95,96]. SARS-CoV-2 has also been shown to have a broad potential host range, which includes minks, dogs, cats and numerous other mammals[97]. By introducing SARS-CoV-2 to novel host species the virus can accumulate adaptations within new hosts which can later be reintroduced back to human populations, as was seen in mink farm-related

outbreaks in Denmark and Netherlands[98,99]. Lastly, the virus can spread silently in human populations where the sequencing capacity for viral monitoring is low, allowing the virus to gradually acquire mutations and remain hidden until it has spread and become a dominant variant[100,101].

1.2.5 Intra-host variation

Recent advances in sequencing technology have provided means for cost-effective generation of massive amounts of sequencing data to study genomic diversity of viral populations within infected persons. Previous studies using sequencing technologies with lower throughput have only been able to investigate the viral consensus genome, i.e. one viral genome per sample. The viral consensus genome sequence is usually generated from sequencing data by calling only the most frequent nucleotide in each position of the genome. However, viral infections consist of a large number of viral genomes, all undergoing replication where new mutations can be introduced into the viral intra-host population at a low frequency that might affect their ability to evade host immune responses or disease severity (Figure 6)[102–104]. By only investigating the viral consensus genome, biologically and clinically relevant viral intra-host genomic variability and population diversity can be lost.

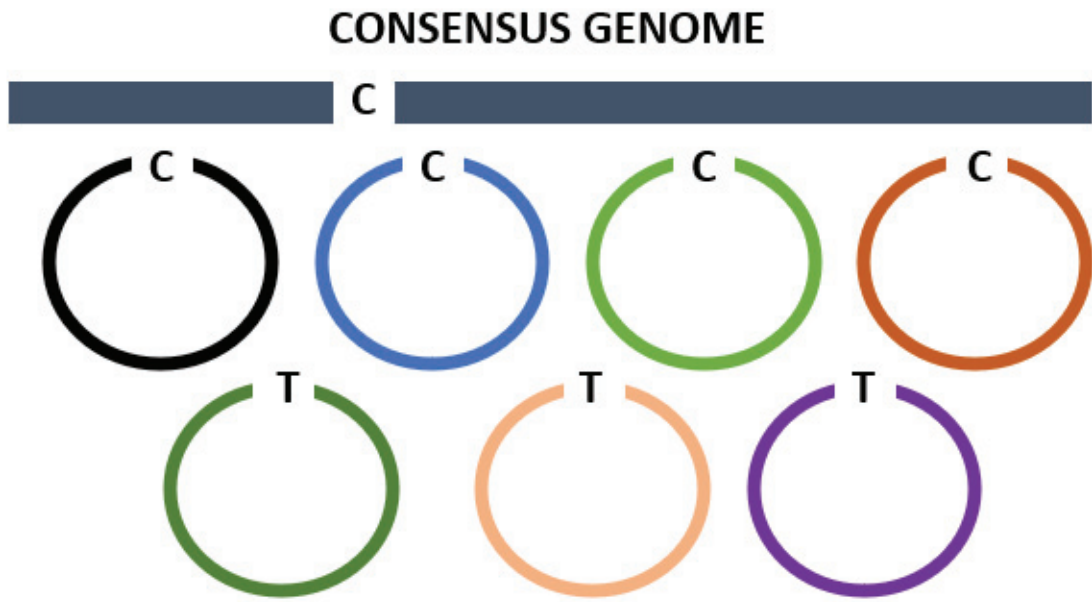


Figure 6: A schematic representation of viral intra-host variation. A consensus genome (black bar) is generated using the most frequent nucleotide in each position. Individual viral genomes (coloured circles) differ in a given position, with the majority having a C, while the rest have a T. The consensus genome will then generate an C in that position of the genome, while the genomic information from the rest of the viral population is lost. By retaining information of variants below the consensus level a more accurate representation of the true viral diversity can be investigated.

Papillomaviruses are slowly evolving viruses, with an estimated mutation rate five times higher than their hosts[105]. It was therefore surprising when deep sequencing of HPV positive samples revealed more low-frequency minor nucleotide variants (MNVs) below the consensus genome level than expected[106–110]. Certain viral genomic positions containing MNVs have been found to be associated with a lower or higher risk of developing precancerous lesions, and the level of MNVs in persistent infections that progress to high-grade lesions or cancer are found to decrease relative to infections that are cleared by the immune system[111–113]. How these MNVs are generated during an HPV infection is presently not completely understood. Although the employment of high-fidelity host polymerases are often used as explanation as to why HPV evolve slowly and replicate with high accuracy[114], a possible source for some of the intra-host variation found could be the recruitment of low-fidelity polymerases during infections[115,116].

Intra-host variation has also been studied in SARS-CoV-2 infections, where the presence of numerous low frequency MNVs has also been found[117–120]. While most SARS-CoV-2 are quickly cleared by the immune system, immunocompromised persons can have persistent infections lasting several months[121,122]. In these persistently infected persons, MNVs of intra-host viral populations has been shown to shift over time as the virus adapted to the host immune system and antibody treatment from convalescent plasma[89–91,123]. These intra-host viral populations have also been found to convergently acquire mutations in the spike protein found in VOCs, including N501Y and deletion of amino acids at position 69-70[89,91]. Furthermore, transmission of intra-host variants between persons from the same household have been investigated, revealing that most MNVs are not shared between members of the same household and a narrow bottleneck in SARS-CoV-2 transmission with few virions causing new infections[118,119,124].

One source of genomic variation found in HPV and SARS-CoV-2 genomes is the activity anti-viral host-defence enzyme apolipoprotein B mRNA editing enzyme catalytic polypeptide-like 3 (APOBEC3)[111,113,120,125,126]. APOBEC3 is a family of proteins which is a part of our innate immune system against viral infections, and they bind to single-stranded DNA and RNA to induce cytosine (C) to thymidine/uridine (T/U) mutations[127,128]. By inducing C>T substitutions in viral genomes, APOBEC3 activity may lower viral fitness by introducing deleterious mutations[126]. APOBEC3 has a preferred trinucleotide context in which they induce C>T mutations, TCN, where N is any nucleotide, except for APOBEC3G which often induces C>T mutations in a CC context[129]. APOBEC3-induced mutations have been observed in HPV positive cervical samples, mainly in low-grade or transient infections, as well as in SARS-CoV-2 and other viruses[111,129–132]. Mutations caused by APOBEC3-activity has also been found in the genomes of human tumours, and more often in HPV-positive tumours[133–135].

1.3 Life cycle and pathogenicity/pathogenesis

1.3.1 HPV life cycle and pathogenicity

The HPV life cycle starts with the infection of the basal cell layer in epithelial tissues through microlesions and is dependent on host cell differentiation through the epithelium layers to complete its life cycle[32,114]. The cervical squamocolumnar transformation zone, which is

the area where columnar epithelium of the endocervix is replaced by squamous epithelial cells of the ectocervix, is especially susceptible to HPV-infections as basal cells are particularly accessible at this site[136]. The virus attaches to the host cell surface through the binding of the viral L1 capsid protein to heparin sulphate proteoglycans (HSPGs). The viral capsid then undergoes several conformational changes, exposing L2 in the viral capsid which allows for the movement of the virus to an, currently unknown, uptake receptor complex and cell entry through endocytosis[137]. The virus genome is then transported to the cell nucleus where viral replication is initiated[114]. In the basal cells, the HPV genomes are maintained as low-copy episomes, and viral replication occurs in parallel to the replication of the cellular genome[114].

As the cells from the basal layers migrate upwards through the epithelium layers, they enter their differentiation process and exit the cell cycle, which also signals the start of the productive phase of the HPV life cycle (Figure 7). As the differentiating cells exit the cell cycle, viral E6 and E7 proteins works in tandem to hijack cell cycle checkpoint mechanisms and promote cell proliferation by inducing unscheduled re-entry into S-phase cell cycle, while preventing apoptosis which normally occurs when cells proliferate when they should not[22]. As the cells migrate through the intermediate layers, E6 and E7 expression is replaced by the expression of E1, E2, E4 and E5, which steps up viral replication, resulting in thousands of viral genomes per cell[30,114]. Lastly, as the cells reach the upper layers of the epithelium, the viral L1 and L2 capsid proteins are expressed and autoassemble into virions, which encapsidates the viral genome. Assembled virions are then released from the epithelial surface[32]. The time from infection to generation and release of infections virus takes approximately 2-3 weeks, the same time it takes for cells in the basal epithelial layer to migrate to the surface layer and desquamate[138].

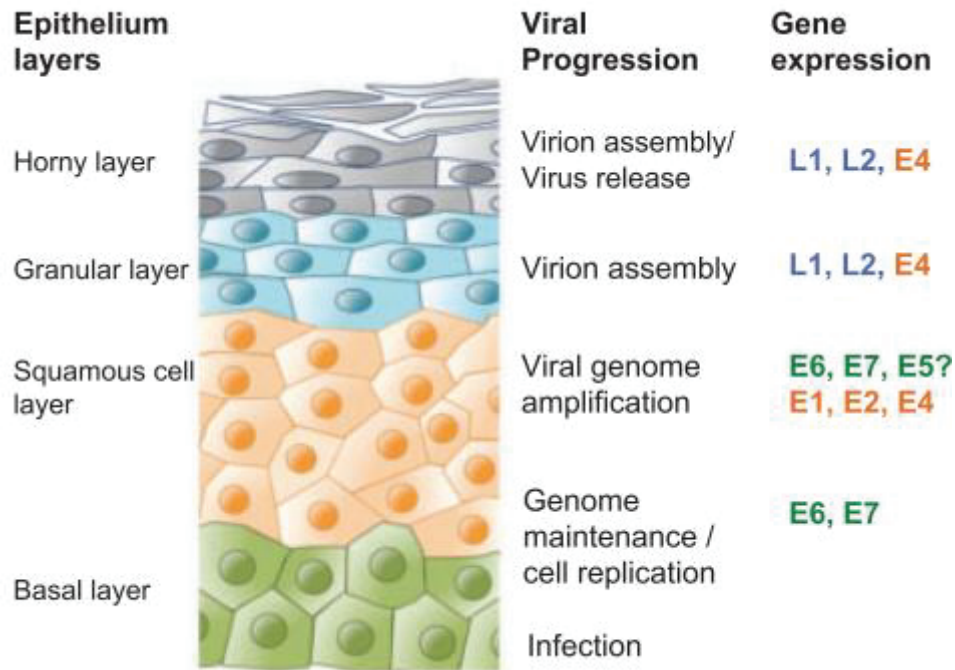


Figure 7: Schematic overview of the general productive HPV life cycle. Squamous epithelium represented on the left and the different stages of the life cycle and which genes are expressed on the right. Modified from [22]. Copyright 2015 by Oxford University Press. Printed under CC BY 4.0 licencing, <https://creativecommons.org/licenses/by/4.0/>.

Normally, ~90% of HPV infections are cleared by the immune system within two years, however, a minority of infections become persistent, lasting several years or decades[139]. Persistent infections can progress to cervical intraepithelial neoplasia (CIN), where CIN1 is an insensitive histopathological sign of HPV infection and CIN2 and CIN3 represents precancerous lesions that can progress to invasive cervical cancer[3]. Persistent infections lasting more than a few years dramatically increase the risk of developing precancerous lesions, and the same is true for the progression of precancerous lesions to invasive cancer which typically occurs gradually over years (Figure 8)[140,141]. Spontaneous regression of precancerous lesions classified as CIN2 and CIN3 occurs, and the regression rate is estimated to be 50-70% and 20-30%, respectively[4,141,142].

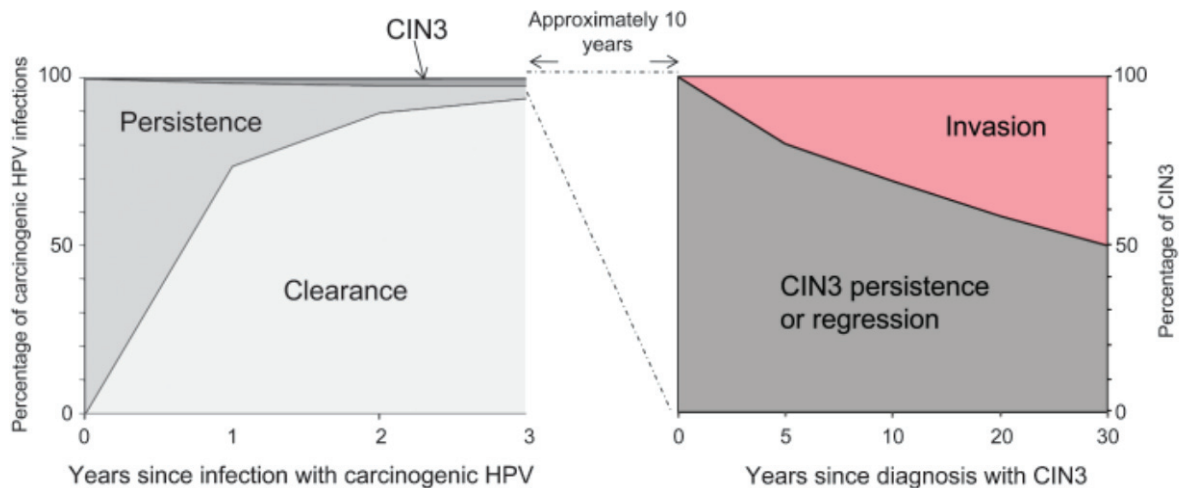


Figure 8: Risk of persistence of HPV infection and estimated risk of progression over time. From [140]. Copyright 2011 by Oxford University Press.

1.3.2 SARS-CoV-2 life cycle and pathology

SARS-CoV-2 transmits through aerosols and respiratory droplets and initially infects epithelial cells in the upper respiratory tract[143], but infections in the lower respiratory tract can happen directly or by inhaling particles from the upper respiratory tract[144]. To enter the cells, the S1 subunit of the viral spike protein binds to ACE2-receptor on the cellular surface and is then cleaved at the S2' site by cellular proteases such as transmembrane serine protease 2 (TMPRSS2)[145]. The cleavage of the spike protein activates the S2 subunit, which allows the virus lipid bilayer to fuse with host lipid bilayer and the release of viral gRNA into the cytosol (Figure 9)[146,147]. When the viral genome is released into the cytosol, host-cell ribosomes are recruited and translates ORF1a and ORF1b, which is then cleaved into 16 non-structural proteins (nsp1-nsp16) that assemble into the replication-transcription complexes (RTCs) initiating viral RNA synthesis and protein expression[148].

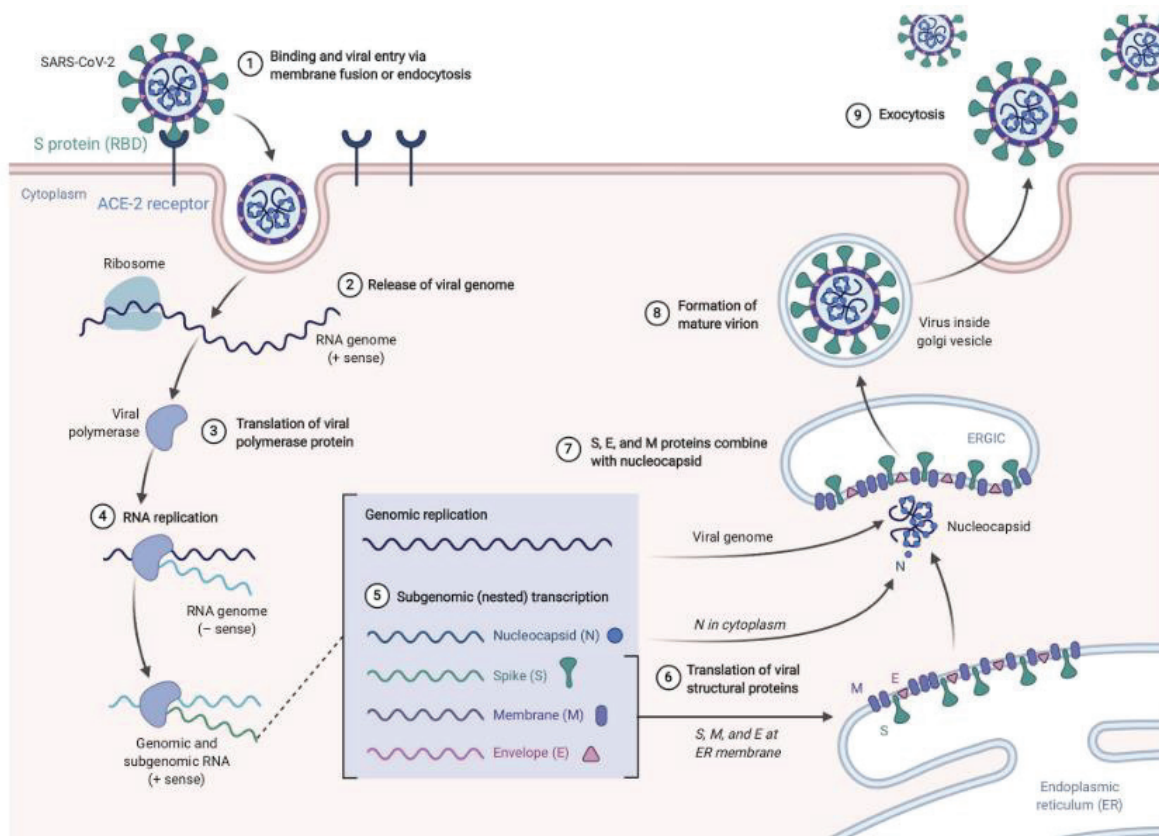


Figure 9: Overview of the SARs-CoV-2 life cycle. Modified from [149]. Copyright by Alturki, Alturki, Connors, Cusimano, Kutzler, Izmirly and Haddad, 2020. Printed under CC BY 4.0 licencing, <https://creativecommons.org/licenses/by/4.0/>.

RTCs remodel host cell membranes into replication organelles, allowing for viral RNA synthesis while evading the host immune system. Using gRNA as a template, the RTC produces a full-length genome complement (the anti-genome) which is used as a template to produce new gRNA, and a set of minus-strand subgenomic RNAs (sgRNAs). Minus-strand sgRNA are derived from the region downstream of ORF1a and ORF1b and directs the synthesis of a nested set of subgenomic mRNAs (sg-mRNAs) which are translated into structural and accessory proteins[148]. Translated structural proteins translocate to the endoplasmic reticulum (ER) membranes and transit through ER-to-Golgi intermediate compartments (ERGIC), where the gRNA and structural proteins are assembled into viral particles. Lastly, virions are secreted from the cells by exocytosis[150,151].

The median incubation time before onset of COVID-19 symptoms is 4-5 days, and most patients experience transient infections with mild to moderate symptoms, including coughing,

fatigue, fever, headache, myalgia and diarrhoea[152,153]. In some cases, when the infection has spread to the lower respiratory tract, it can develop into severe illness and acute respiratory distress syndrome (ARDS), which usually happens one week after the onset of symptoms[144]. COVID-19 ARDS is defined as a form of lung injury characterized inflammation, pulmonary vascular leakage and loss of aerated lung tissue which can be fatal or cause irreversible lung damage[144]. Severe COVID-19 can also affect other tissues than the lung and cause acute cardiac, kidney and liver injuries in addition to cardiac arrhythmias, coagulopathy, and multiorgan failure[154,155]. Most of these severe symptoms are caused by excessive inflammatory responses and immunopathology[144,156]. Around 3-20% of people infected with SARS-CoV-2 require hospitalization and 10-30% of these will require intensive care. The total fatality rate is estimated to be ~1%[144]. These numbers depends on the vaccination status of the population and virulence of SARS-CoV-2 variant causing the infection[157,158]. Additionally, individuals who have had SARS-CoV-2 infections can acquire post-COVID, which lasts for at least 2 months. Common symptoms include, but are not limited to, fatigue, shortness of breath and cognitive dysfunction[159].

1.4 Prevention and treatment

1.4.1 HPV vaccination, screening and treatment

The best method to prevent HPV-induced cancers is vaccination against HR-HPV infections. HPV vaccines are based on the viral L1 capsid protein, which self-assembles into virus-like particles (VLP) that resemble HPV virus particles, but without any genetic material and are non-infectious[23,138]. There are currently three prophylactic vaccines commercially available, and while they all employ L1 VLPs, they differ in the number of HPV types they protect against. The vaccines and the HPV types they protect against are 1) the bivalent Cervarix® (GlaxoSmithKline, London, UK) which protects against HPV16/18, 2) the quadrivalent Gardasil® (Merck, Kenilworth, New Jersey, USA) which protects against HPV6/11/16/18, and 3) the nonavalent Gardasil® 9 (Merck, Kenilworth, New Jersey, USA) which protects against HPV6/11/16/18/31/33/45/52/58[160,161]. While the vaccines do not protect against all HR-HPV types, there is some evidence cross-protection against other HR-HPVs that are genetically related to the genotypes targeted by the vaccines[162–164].

The vaccines have proven to be very efficient at preventing HPV infections if given prior to HPV exposure[165]. Several studies have shown that the vaccine has >90% efficacy against persistent HPV infections and precancerous lesions in women with no HPV infection at the start and end of the three-dose immunization trials[166,167]. Young adolescent girls aged 9-13 have normally been targeted for vaccinations but in the recent years several countries have also started targeting boys[166,168]. While the vaccines have proven to be very effective, most of the over 100 countries that has introduced a national HPV vaccination programme is high or upper-middle income countries, covering only 30% of the global population[169]. Low vaccination coverage in low- or middle-income countries is still a problem, as these are the countries where the cervical cancer burden is the highest and the need for a vaccine the greatest[169,170].

Cervical cancer screening is another method to prevent HPV-induced cervical cancers[23]. With the rollout of HPV vaccines mainly being confined to high or upper-middle income countries, screening is still the most important prevention tool in low- and lower-middle income countries where access to vaccines is limited[170]. Cervical cytology and HPV testing are the two most common methods to screen for cervical cancer. Cervical cytology is a microscopic evaluation of cells from cervical samples, and samples are usually examined using liquid-based cytology (LBC), where cervical epithelial cells are collected with a sample brush before the sample is preserved in a suspension and prepared on a microscopic slide. Precancerous cytological changes are commonly classified according to the Bethesda system, where squamous lesions are classified as low-grade squamous intraepithelial lesions (LSIL) or high-grade squamous intraepithelial lesions (HSIL)[23]. Additional terms for the classification of uncertain results are atypical squamous cells of undetermined significance (ASC-US) and atypical squamous cells, cannot exclude HSIL (ASC-H)[171].

HPV testing in cervical cancer screening allows for the identification of HR-HPV genotypes and most HPV tests are based on detecting HPV DNA. HPV testing has proven to be more sensitive than cytology, and several studies have demonstrated that HPV testing in primary screening allows for earlier detection of precancerous lesions and reduced number of cervical cancers during follow-up, thus allowing for extended screening intervals. However, as most HPV infections are transient, HPV testing is less specific than cytology[172–174].

In the Norwegian cervical cancer screening program, which was implemented in 1995, women between the age of 25 to 69 years are invited for screening. The primary screening method for women aged 25-33 years is cytology, with a screening interval of three years. Between 2018

and 2021, HPV testing replaced cytology as the primary screening method for women aged 34-69 years, with a screening interval of five years[175].

Owing to the relatively low regression levels of high-grade precancerous lesions[141,142], women diagnosed with high-grade cell changes(ASC-H or HSIL) are referred to colposcopy (visual inspection of cervix under magnification) and biopsy in the Norwegian guidelines[176]. Biopsies are used for histological classification of squamous cervical neoplasia which follows the CIN scale. The CIN scale is based on the severity of dysplasia and classifies lesions as CIN1 (mild dysplasia), CIN2 (moderate dysplasia) and CIN3 (severe dysplasia and carcinoma *in situ*) based on the proportion of epithelium replaced by undifferentiated cells[177]. Precancerous lesions in glandular cells are classified as adenocarcinoma *in situ* (AIS)[178]. For women diagnosed with CIN2 or more severe lesions, it is recommended that the abnormal cells are removed by Loop electrosurgical excision procedure (LEEP) which uses a small electrical wire loop to remove abnormal cells from the cervix[179,180]. Not all precancerous lesions will progress to cervical cancer, which means that treatment of precancerous lesions will result in some level of overtreatment[5]. However, the benefit of treatment is considered to outweigh the risks if excessive overtreatment is avoided[181]. All cases of cervical cancer are treated, and the type of treatment depends on how far the cancer has progressed[23].

1.4.2 SARS-CoV-2 vaccination, preventive measures, and treatment

As is the case with HPV, vaccines are the first, and best, line of defence against SARS-CoV-2 infections and hospitalization. Vaccines against SARS-CoV-2 target the S protein, usually in its prefusion conformation, as this has been identified as the main antigenic component responsible for inducing host-immune responses and to confer protective immunity against the virus[182]. The role of the S protein in receptor binding and membrane fusion allows for the generation of antibodies that can block virus binding and fusion and neutralize the virus to stop infections[182]. WHO has put forth a list of vaccines that has met the necessary criteria for safety and efficacy (10 in total) as well as who should get vaccinated[183]. In Norway, all people over the age of 16 are offered a coronavirus vaccine, although children 5-15 years can also be vaccinated if they are in the risk groups for more severe disease or simply according to their parents wish[184]. Three SARS-CoV-2 vaccines are currently offered as part of the Norwegian coronavirus immunisation programme, two based on mRNA technology and a third

which is protein-based[185]. Several other vaccines are available and in development, including those based on viral vectors and VLPs among other technologies[149,186].

The two mRNA vaccines are Comirnaty (Pfizer–BioNTech, NYC, NY) and Spikevax (Moderna, Cambridge, MA) and they utilize lipid nanoparticles to transport prefusion-stabilised S protein-encoding mRNA to the host cells, which then uses the mRNA to produce the S protein and induce an immune response[187]. Both vaccines have proven very effective, with an efficacy over ~95%[188]. The Novavax vaccine (Novavax, Rockville, MD) is a protein-based vaccine where nanoparticles are coated with a recombinant S protein (SARS-CoV-rS), which elicits an immune response against the S protein[149,189]. Novavax has also been proven to be highly effective, with an overall efficacy reported to be 89%[188]. All three vaccines require two doses to complete primary vaccination, while additional booster doses are offered, as it has been shown to give longer-term protection and give broader protection against novel SARS-CoV-2 variants[84].

While the emergence of new SARS-CoV-2 variants with mutations in the S protein that confer characteristics of immune escape have lowered the vaccine efficacies, the vaccines does according to several studies still offer good protection against severe disease and hospitalisation[67,82,84,190].

Another preventive measure to limit SARS-CoV-2 infection in the population is non-pharmaceutical public health and social measures, including quarantining persons with SARS-CoV-2 infection, handwashing, social distancing, wearing of masks and intensive contact tracing around infected persons. A systematic review and meta-analysis reviewed effectiveness of public health measures and concluded that handwashing, mask wearing and physical distancing were associated with reduction in SARS-CoV-2 incidence[191]. Likewise, a study investigating the SARS-CoV-2 transmission clusters in the Nordic countries found that country-specific intervention strategies had the largest impact, with Sweden which had the least strict SARS-CoV-2 intervention policies at the start of the pandemic also had more transmission clusters with larger size and durations compared to Norway, Finland and Iceland (similarities in number and size of transmission clusters between Denmark and Sweden were potentially explained by the higher population densities in Denmark)[192].

Treatment of severe cases of SARS-CoV-2 is mainly confined to oxygen therapy and administration of anti-inflammatory corticosteroids to reduce the immunopathological consequences of SARS-CoV-2 infections[144]. Additionally, neutralizing antibodies targeting

the spike protein can be administered to treat infections[193,194]. Antiviral medication can also be administered, but to have an effect it has to be taken before the onset of severe COVID-19[144].

1.5 Molecular mechanisms of HPV-induced cancer

HPV-induced cancer is a multistep process that relies on several different mechanisms where HPV-infected cells are able to attain several of the hallmarks of cancer, recognized as necessary biological capabilities cells need to acquire to transform into tumours[195]. Much of the carcinogenic potential of HR-HPVs are attributed to the activity of the viral oncoproteins E6 and E7 (Figure 10). The oncoproteins contribute to cellular instability through induction of cell proliferation and inactivation of cell-cycle regulatory and tumour suppressor mechanisms[30]. Together, the oncoproteins maintain an unstable cellular environment that over time causes HPV-infected cells to undergo malignant cellular transformation. E6 is known to interact with many host-proteins, but the most characterised of these is the tumour suppressor p53 which E6 targets for degradation[196]. p53 inhibits cell-cycle progression in response to DNA damage and other cellular stress signals and induces apoptosis in abnormal cells, thus acting as an important tumour suppressor protein[197]. E7 inhibits the activity of the retinoblastoma-associated (RB) proteins whose key function in maintaining cellular stability is to decide whether cells should proceed through the growth-and-division cycle or not[198]. Consequently, the activity of E6 and E7 leads to the formation of cell populations with deregulated cell-cycle and apoptosis mechanisms which accumulate mutations over time[114]. This is one of the mechanisms by which persistent infections with HR-HPVs can induce malignant cellular transformation[30]. While low-risk HPV types also encode for E6 and E7, and their activities are sufficient for the generation of benign warts, they are insufficient to trigger the development of cancer [23].

An additional mechanism by which HPV-induced cancer can develop is by the integration of HPV into human chromosomes[199]. This has been identified as a major driver event in HPV-induced carcinogenesis and the frequency of HPV integrations has been found to increase with lesion severity[200,201]. HPV integration into the host genome has been found to have several detrimental effects. One of these effects is caused by the disruption or deletion of the viral E1 or E2 genes, which regulates the expression of oncoproteins E6 and E7, subsequently leading to overexpression of viral oncogenes and genomic instability (Figure 10) [199]. Overexpression

of E6 and E7 can also happen when multiple copies of HPV are integrated as viral-host concatemers, which leads to additive amplification of viral oncogene expression[199,202].

HPV integrations can also cause nearby structural rearrangements, deletions, somatic mutations, and amplifications of the host genome, which depending on the integration site, can promote carcinogenesis[203–205]. HPV integrations into the host genome has been found distributed in a non-random manner, with several identified hot-spots. Common for these hot-spot regions are their association with known fragile sites and in gene-rich transcriptionally active regions of the chromosomes, containing several oncogenes and tumour-suppressor genes whose disrupted functions might confer a selective growth advantage to the affected cells[206–208]. Microhomology between viral and human genomic sequences has also been observed at integration breakpoints, likely contributing to the non-random distribution of integration-sites[209].

While integration events are not necessary for the development of all HPV-induced cancers, they are often detected in tumours. A study of the Cancer Genome Atlas Research Network observed HPV integrations in >80% of HPV positive cervical cancers[210]. It was found that 76% of the HPV16 and 100% of the HPV18 positive tumours showed integrations, which is in concordance with other studies showing that the integration frequency differs between the HPV types[211,212]. Most studies on integration frequencies have investigated HPV16 and HPV18, but for other HR-HPVs, such as HPV31, 33, 45, 52 and 58, all have been found to have higher amounts of integrated HPV with increasing lesion severity[213–215]. However, there is substantial variation in integration frequencies between different HPV types[211].

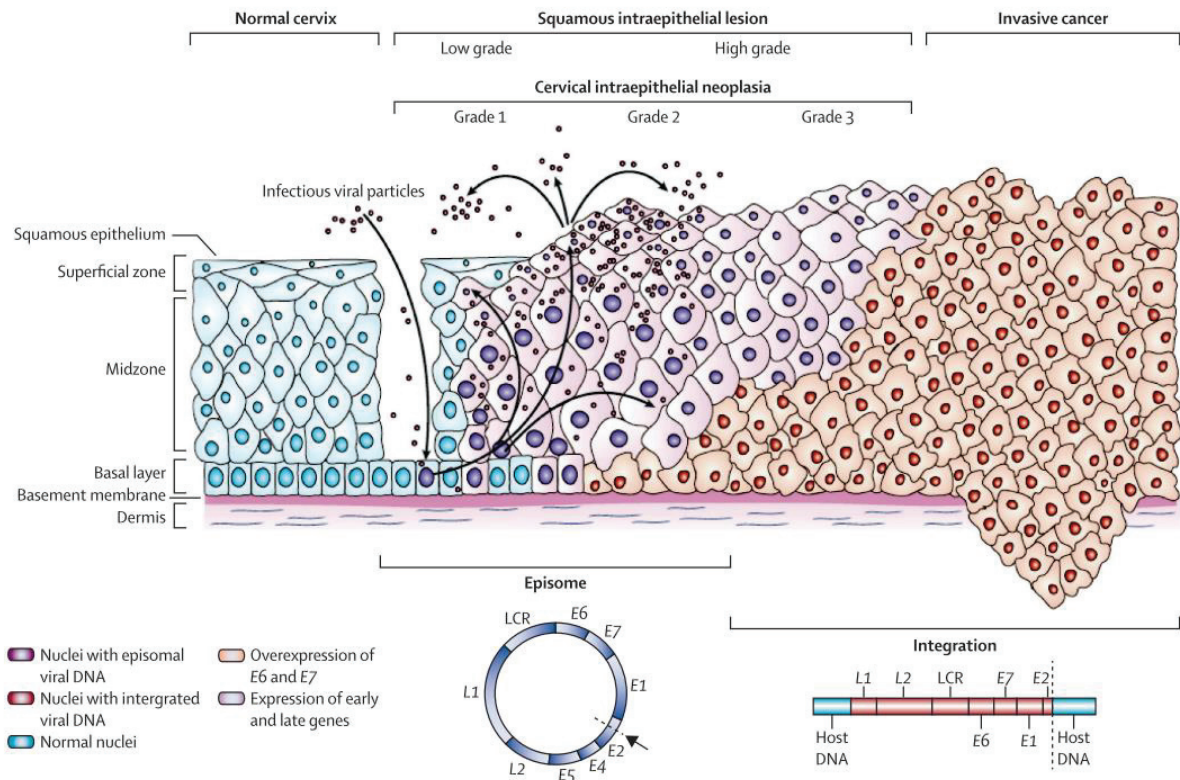


Figure 10: Overview of HPV life cycle in epithelial tissue and gradual gain of cervical intraepithelial neoplasia which eventually progress to invasive cancer. Low-grade intraepithelial lesions are associated with productive HPV infections, and progression to high-grade lesions and invasive cancer is associated with overexpression of viral oncoproteins E6 and E7 and viral integrations into host chromosomes (although viral integration is not always necessary for HPV-induced carcinogenesis). From [216]. Copyright 2007 by Nature Publishing Group.

Integration events during HPV infections are not part of the HPV natural history and considered an evolutionary dead end due to the resulting non-productive infections[199]. These integration events in HPV are therefore considered accidental events and a possible consequence of the physical proximity of the HPV genome and the human genome in the nucleus[217]. Integration events have also been studied in SARS-CoV-2, and an early study claimed to have found evidence of viral-host chimeric sequences based on RNA sequencing data of patient tissues[218]. However, the study has received a lot of criticisms for misinterpreting technical artefacts as biologically relevant and later studies having investigated integration events in SARS-CoV-2 have found no evidence of such [219–221]. Why there are integration events in HPVs and not SARS-CoV-2 could be explained by the fact that SARS-CoV-2 has an RNA

genome and does not replicate via DNA, while HPVs have DNA genomes. Additionally, the genome replication of SARS-CoV-2 occurs in the cytoplasm, while HPV replicates its genomes in the nucleus of infected cells, allowing for the usage of HPV DNA strands to be mistakenly used by host DNA repair mechanisms, like nonhomologous end joining or microhomology-mediated recombination[209,222].

1.6 Viral surveillance and genomic epidemiology of SARS-CoV-2

The SARS-CoV-2 sparked an unprecedented global sequencing effort which would not have been possible without the recent advances in sequencing technology and computational power, making it cost-effective to sequence and analyse a large number of samples[223]. At the time of writing, more than 11.5 million SARS-CoV-2 genomes has been deposited in the Global Initiative on Sharing All Influenza Data (GISAID) database[224], which started as a global initiative to share avian flu data and has now been co-opted for sharing of SARS-CoV-2 data[225]. As a result of this concerted sequencing effort, we now have SARS-CoV-2 genomes from all parts of the world during all time frames of the pandemic. The combination of genomic data with individual-level metadata has allowed us to gain knowledge about what drives the transmission of this virus and has been crucial to understand the dynamics of the pandemic and evaluate the efficacy of control measures[223]. SARS-CoV-2 genomes can now be sequenced within hours or days of a case being identified, effectively allowing for a real-time surveillance of the pandemic and the implementation of rapid public health responses[223].

Genomic epidemiology studies, which links SARS-CoV-2 genomes with metadata to understand disease transmission[226], was crucial in the early phase of the pandemic to understand how the virus spread on a global and more local scale. Early in the pandemic, it has been shown that most viral lineages were very cosmopolitan and present in different parts of the world[227]. However, as countries started to impose international travel restrictions lineages become more continental- and country-specific as importation events started to decline, illustrating the effect of international travel bans[227–229]. It was also shown that international travel ban restrictions has to be imposed early to have an significant effect[230–232].

Genomic epidemiology has also been used to evaluate the effectiveness of non-pharmaceutical interventions, and a study using 29.000 publicly available SARS-CoV-2 genomes from 57

locations reported that locations under early and most stringent interventions experienced less severe COVID-19 morbidity and mortality[233]. Other studies from different locations have also found similar results, where implementation of non-pharmaceutical interventions were associated with reduced viral reproduction number and less numerous and shorter transmission chains[234–236].

Genomic epidemiology studies are also valuable to confirm or refute healthcare-related outbreaks in order to rapidly evaluate if personal protective equipment and infection control guidelines are adequate. Healthcare-associated infections can affect both patients and healthcare workers (HCWs), increasing morbidity and mortality for patients and sickness and morale for HCWs[237]. Additionally, HCWs can be a source of infection in patients, and it was therefore of utmost importance to minimise the risk of nosocomial transmission and assess the burden of nosocomial COVID-19 infections[237]. Several studies have investigated this and examples where the findings have guided local infection control measures are plenty[237–242].

A more general surveillance of viral genomes circulating in the populace has also proven to be a valuable tool in discovering and tracking emerging VOCs and VOIs. By continually sequencing viral genomes, one can discover variants containing mutations that increase transmissibility, that are resistant to therapeutic treatments and/or to which previous infections and vaccines are less effective. This was the case with the emergence of the first three VOCs, Alpha, Beta and Gamma, which was determined to be VOCs because of the presence of mutations in the spike protein associated with increased transmission and ability to evade host immune response in combination with their rapid increase in frequency in the populations[243].

While the recent pandemic has sparked an unprecedented global sequencing effort, it is important to keep in mind that there is a bias in which countries generate the data. Currently, the sequencing effort is biased towards a limited set of countries with high sequencing capacity, mainly regions and countries with already existing infrastructure and competence for large-scale sequencing and computational analysis[101,244,245]. Thus, the sequencing coverage is not equally high in all parts of the world, meaning that we still have regions where undetected transmission of novel variants containing mutations of interest can occur in high frequency and be exported to other parts of the globe.

This discrepancy mirrors the availability of both SARS-CoV-2 and HPV vaccines between high- and upper-middle income countries and low- and lower-middle income countries[246,247]. There has been a narrative that the pandemic has hit Africa less severely

than the rest of the world, termed the “African paradox”[248]. Several studies have found that this is most likely a misinterpretation due to infections not being reported, and the burden and transmission of SARS-CoV-2 has been more severe than previously thought[249–251]. Vaccines are the most potent weapon against the ongoing pandemic, and until the regional differences are equalized, SARS-CoV-2 variants will continue to emerge. WHO has announced a global call to action to eliminate HPV-induced cervical cancer[252], but it will continue to be an ongoing health burden in low- and lower-middle income countries for decades to come until they get better access to HPV vaccines[7,253]. Additionally, the available HPV vaccines does not target all HR-HPV types that are more prevalent in sub-Saharan countries where the HIV-prevalence is also high and associated with increased risk of developing invasive cervical cancer[7,254].

2. THESIS AIMS

The overall aim of this thesis is to use NGS technology and viral genomic analyses on HPV positive cervical samples to study intra-host genomic events and on SARS-CoV-2 positive samples to investigate intra-hospital transmission. The first part of the thesis uses viral genomics to characterise HPV intra-host genomic variation and chromosomal integrations for different HR-HPV types and relate these events to HPV-induced carcinogenesis. Similarities and differences between the HPV types and between diagnostic categories can reveal insight into type-specific molecular mechanisms behind persistent HPV infections and HPV-induced carcinogenesis. The second part uses genomic information from SARS-CoV-2 WGS performed on the Nanopore platform to increase the resolution when performing intra-hospital outbreak investigations.

Study I: To compare HPV minor nucleotide variation and integration profiles in HPV16 and HPV18 positive cervical samples of different diagnostic categories.

Study II: To compare HPV minor nucleotide variation and integration profiles in cervical samples of different diagnostic categories positive for five HR-HPV types and investigate if type-specific profiles adhere to phylogeny.

Study III: Combine SARS-CoV-2 genomic data with epidemiological data to increase resolution of intra-hospital outbreak investigations.

3. MATERIALS AND METHODS

3.1 Sample material and study design

Study I and II

Both study I and II were designed as cross-sectional studies. Sample material used in both studies were collected from a biobank generated in the period January 2005 to April 2008 at Akershus University Hospital, with cytological material and DNA from women attending the cervical cancer screening programme in Norway[255,256]. In both studies, a category of non-progressive infections was defined as: 1) samples with normal cytology (at enrolment and during the preceding two years, and with no previous history of treatment for cervical neoplasia) and 2) ASC-US/LSIL samples from women with no history of cervical abnormality and with no follow-up diagnosis within four-year follow-up. In addition, cytological samples representing each category of progressive disease were included, including women with histologically confirmed CIN2, CIN3, AIS and cervical cancer. For **study I**, all samples positive for HPV16 (n=157) and/or HPV18 (n=75), alone or co-infected with other HPV types, were included in the study, except for HPV16 CIN3 category of which a random selection of 50 samples were included. For **study II**, all samples positive for HPV31 (n=117), HPV33 (n=104) and/or HPV45 (n=66), alone or co-infected with other HPV-types, were included in the study.

Study III

Study III were also designed as a cross-sectional study. Sample material included in the study were collected from healthcare workers (HCWs) and patients from wards with suspected outbreaks of Covid-19 between March 10th and July 1st, 2020. Possible outbreaks were defined as wards with two or more infected HCWs who had had close contact (>2m without PPE for >15 minutes 24 hours (48 hours from June 2020) before the onset of symptoms) and who tested positive less than three weeks apart. If a ward had a suspected outbreak, all viral isolates from HCWs (n=21) in those wards were included in the study. Viral isolates from two patients were also included in the study, based on reported or suspected breaches of infection control practices between patient and HCW. HCWs (n=8) with no reported close contact to other cases in the hospital and from wards with no suspected outbreaks, as well as anonymized patients (n=10), were also included in the study to better assess the local SARS-CoV-2 genetic diversity.

3.2 DNA/RNA extraction and HPV genotyping

Study I and II

The extraction of nucleic acids and HPV genotyping are described in the original studies [255,256]. In brief: either miniMag or easyMag (BioMerieux Inc., France) were used for extracting nucleic acids; the Amplicor HPV test (Roche Molecular Systems, Pleasanton, CA) followed by Linear Array HPV assay (Roche Molecular Systems, Pleasanton, CA) were used for testing for the presence and genotype of HPV DNA, respectively. In addition, results using the PreTect HPV Proofer E6/E7 mRNA test (PreTect AS, Klokkestua, Norway) were available for all samples, providing HPV 16, 18, 31, 33 and 45 genotype information.

Study III

Total nucleic extraction was done using the easyMag (BioMerieux Inc., France) extraction protocol. RT-PCR to detect the SARS-CoV-2 virus E-gene was done using the method published by Corman et al.[257].

3.3 DNA concentration and Ct-value

Study I and II

DNA concentration was measured using Quant-iT™ Broad-Range dsDNA Assay Kit (Thermo Fisher Scientific, Waltham, MA).

Study III

For study II, cycle threshold (Ct) values were determined using the RT-PCR protocol published by Corman et al.[257].

3.4 Library preparation and sequencing

Study I and II:

Library preparation of samples used in **study I and II** was done using the Tagmentation-assisted multiplex PCR enrichment sequencing (TaME-seq) protocol as described in [258] and [259]. Briefly, sample DNA is tagmented and subjugated to two separate multiplex PCR enrichment reactions using specific primer pools designed to hybridize with the forward and reverse strands

of the target HPV genome. These reactions thereby obtain genome coverage from both strands while also allowing for the detection of HPV integrations into human chromosomes. Following the protocol, if the original sample DNA concentration was >2.5 ng/ μ l it was diluted to 2.5 ng/ μ l, or else it remained undiluted. Sample DNA was then tagmented using Nextera DNA library prep kit (Illumina, Inc., San Diego, CA), with the following modifications: 1) reaction volume was reduced to 20 μ l, 2) DNA input amount varied from 0.96 ng to 20 ng based on the original DNA concentration of the sample, 3) incubation was performed at 55°C for 4 minutes. To purify the tagmented DNA, DNA Clean & Concentrator™-5 columns (Zymo Research, Irvine, CA) were used according to the manufacturer's instructions or ZR-96 DNA Clean & Concentrator™-5 plates (Zymo Research, Irvine, CA) according to the Nextera® DNA Library Prep Reference Guide (15027987 v01). Tagmented DNA was then split into two pools before undergoing PCR amplification for target enrichment with one primer pool containing the forward HPV primer pool and another containing the reverse HPV primer pool. Amplification was performed in 20 μ l reactions containing 5 μ l of tagmented DNA, 10 μ l of Qiagen Multiplex PCR Master Mix (Qiagen, Hilden, Germany), 2 μ l of Q-solution (Qiagen, Hilden, Germany), 0.75 μ M of HPV primer pool, 0.5 μ M of i7 index primer, and 1 μ l of i5 Nextera index primer (Illumina, Inc., San Diego, CA). The cycling conditions were as follows: initial denaturation at 95 °C for 5 minutes; 30 cycles at 95 °C for 30 seconds, at 58°C for 90 seconds and at 72 °C for 20 seconds; final extension at 68 °C for 10 minutes.

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

Study III

Library preparation of samples from **study III** was done following the nCoV-2019 sequencing protocol v2 developed by the ARTIC network, which uses a tiling amplicon scheme, using the V3 primer set (https://www.protocols.io/view/ncov-2019-sequencing-protocol-v2-bp216n26rgqe/v2?version_warning=no)[260]. First, samples with Ct-values 15-18 and 12-15

were diluted 10-fold and 100-fold in water, respectively. Samples then underwent reverse transcription in 10 µl reactions, containing 2 µl LunaScript® RT SuperMix (NEB, Ipswich, MA) and 8 µl sample RNA with the following incubation program: 25°C for 2 minutes; 55°C for 10 minutes; 95°C for 1 minute. Samples were split in two reactions before the cDNA were subjected to PCR amplification for target enrichment using the V3 primer pools designed by the ARTIC network. Amplification was done in 25 µl reactions containing 12.5 µl Q5® Hot Start High-Fidelity 2X Master Mix (NEB, Ipswich, MA), 4 µl V3 primer pool, 6 µl nuclease-free water and 2.5 µl sample cDNA. The cycling conditions were as follows: 98°C for 30 seconds, 30 cycles at 98°C for 15 seconds and at 63°C for 5 minutes.

The ends of the amplified cDNA were prepared for barcoding using NEBNext® Ultra™ II End Repair/dA-Tailing Module (NEB, Ipswich, MA), following manufacturer's instructions before barcodes were ligated using NEBNext® Quick Ligation Module (NEB, Ipswich, MA), following manufacturer's instructions. Sequencing was performed on the Nanopore GridION sequencing platform (Oxford Nanopore Technologies, Oxford, UK).

3.5 Study I and II sequencing data analysis

Because of the differences in the bioinformatic analysis between **study I** and **II** and **study III**, the following section will first explain all the bioinformatics methods done for the sequencing data analysis of **study I** and **II** and then for **study III**. For **study I** and **II**, data analysis of sequencing data was performed using a collection of in-house Python scripts executed using the Snakemake workflow management system[261]. The scripts can be accessed on Github: <https://github.com/jean-marc-costanzi/TaME-seq/>.

3.5.1 Sequence alignment

Raw reads were trimmed for Illumina adapters, HPV primers, base quality (-q 20) and minimum read length (-m 50) using cutadapt (v1.10). Trimmed reads were mapped using HISAT2 (v2.1.0) to human (GRCh38/hg38) and HPV reference genomes obtained from the PaVE database. In **study II**, 1000 bp overhangs were added to reference HPV genomes to account for the circular structure of HPV genomes. Mapping statistics and sequencing coverage were counted using Pysam package in **study I**, while BCFtools was used in **study II**.

3.5.2 Sequence variation analysis

Nucleotide counts mapping to HPV reference genomes and average mapping quality values were retrieved from BAM files and variant calling was performed using an in-house R (v3.5.1) script. For a nucleotide to be called, it had to be observed ≥ 2 times in positions with $\geq 100x$ depth and mean Phred quality score < 20 (**study I**) or < 30 (**study II**). MNVs with frequencies $> 1\%$ were called. Samples with mean sequencing depth $< 300x$ were excluded from the analysis. For **study I**, the non-synonymous to synonymous substitutions (dN/dS) ratio was calculated to indicate positive or negative selection affecting protein coding genes. HPV NCR have homopolymeric T tracts (HPV16:4156–4173 and 4183–4212, HPV18:4198–4234, HPV31:4072-4077 and 4145-4167, HPV33: 4149-4167 and 4186-4195, HPV45:4184-4219) that were removed from the analysis due to the possibility of high frequencies polymerase or sequencing errors being introduced in these regions.

3.5.3 Mutational signature analysis

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

3.5.4 Detection of integration sites and deletions

Integration detection is described previously in[258]. Briefly, the detection is a two-step analysis using discordant and junction reads. Paired-end reads mapped using HISAT2 with one read mapping to target HPV reference genome and the other read to human chromosome were identified as discordant read pairs. Unmapped reads were re-mapped using LAST (v876) aligner (options -M – C2) to identify junction reads mapping to both HPV and human genomes and were used to determine the exact position of HPV-human integration breakpoints. Reads sharing identical start and end coordinates were interpreted as PCR duplicates and not considered. Positions that were covered by ≥ 2 unique discordant read pairs or by ≥ 3 unique junction reads were considered potential integration breakpoints. Certain repetitive regions of the human genome had frequently erroneously reported human integration breakpoints and were filtered out.

3.5.5 Validation of integration sites

A subset of the reported integration breakpoints from the integration analysis were chosen for validation by sanger sequencing. Primer pairs were designed to flank the integration breakpoints, with one primer binding site binding to the human genome and the other to the target HPV genome. SAM flags were retrieved from discordant read pairs and used to verify the genome orientation so designed primers would hybridize to correct DNA strands (Table 1).

PCR amplification was done on samples chosen for validation and amplified PCR products were sequenced on the ABI® 3130xl/3100 Genetic Analyzer 16-Capillary Array (Thermo Fisher Scientific Inc., Waltham, MA) using BigDye™ Terminator v1.1 cycle sequencing kit (Thermo Fisher Scientific Inc., Waltham, MA). Chromatograms were analysed in Geneious Prime (v2020.2.2) and BLAST or BLAT searches.

Table 1. Read and genome orientations at the integration breakpoints, SAM flags and +/- DNA strand used for primer design.

| 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 | - | + |

3.6 Study III sequencing data analysis

3.6.1 Sequencing analysis of SARS-CoV-2 amplicon data

The sequencing data analysis used in **study III** was done using the ncov-2019 bioinformatics pipeline developed by the ARTIC network. The pipeline is a collection of tools for analysing SARS-CoV-2 sequencing data generated using the ARTIC protocol to filter reads and further analyses them, finally generating output files consisting of consensus genomes, QC-metrics, and VCF-files for each individual sample.

Demultiplexed reads from the fastq-pass folder were filtered using the `artic guppyplex` command, filtering out reads with length <400 and >700 . Filtered reads were then passed on to the `artic minion` command, where the `medaka` pipeline was used with depth normalisation of 200x.

3.6.2 Phylogenetic analysis, Nextstrain Clade assortment and pangolin lineage assignment

To broaden the genetic background for the subsequent analysis, available SARS-CoV-2 genomes were downloaded from GISAID. The downloaded SARS-CoV-2 genomes consisted of 73 Norwegian, 6 Chinese and 250 genomes from European countries that had been identified from routine contact tracing as sources of SARS-CoV-2 importation events into Norway, with collection dates no later than July 1st, 2020. A multiple sequence alignment (MSA) of sequenced and downloaded samples was generated using MAFFT (V7.450), before visual inspection to remove low-quality sequences from the MSA. FastTree (V2.1.11) was used to generate phylogenetic trees, using GTR substitution model. To visualize and annotate the phylogenetic tree, an in-house R-script using the `ggtree` package (V.2.2.1) was used.

A combination of phylogenetic placement and presence of clade-defining signature mutations were used to assign Nextstrain clades according to the Nextstrain nomenclature [262]. Samples were assigned Pangolin lineages using the Pangolin COVID-19 Lineage Assigner online tool [263].

3.6.3 Outbreak assessment

To assess if samples were part of the same transmission chain, study unique variants were investigated. Study-unique variants were defined as SARS-CoV-2 nucleotide variants that met the following two criteria: 1) variants that showed no geographic distribution and 2) with two or more co-occurring mutations not found together in any other genome in the GISAID database. Samples sharing study-unique mutations and either had reported close contact or worked in the same ward were considered part of an intra-hospital outbreak.

3.7 Statistical analyses

In **study I**, statistical analysis was performed using non-parametric Kruskal-Wallis test. A Shapiro-Wilk test of normality was performed to confirm that the data did not follow normal distribution.

In **study II**, a non-parametric Chi-square of independence was used to determine whether integration breakpoints in HPV genes occurred more often than would be expected by chance. A generalized linear model (glm) was used to understand the relationship between dependent variables (MNVs, APOBEC3 mutations, number of integrations in samples, integration breakpoints) and the independent variables (HPV type, Alpha-clade, diagnostic category). Number of integrations and MNVs were tested using a negative binomial distribution while a binomial distribution was used for the other tests. Following this, multiple comparisons of means using Tukey HSD was done using the package multcomp[264] to test differences between different categories.

In both studies, a p-value $p < 0.05$ was considered statistically significant. All statistical analyses were done in R (v3.5.1).

3.8 Ethical aspects

This thesis describes the genetic analysis of viruses obtained from human sample material. The HPV studies were approved by the Regional Ethical Committee (2017/447) and by the Akershus University Hospital's Data Protection Official (2017-109). The SARS-CoV-2 data were recorded as part of the hospital's routine for outbreak investigations, as authorized by the institutional infection control programme and the Norwegian regulation of infection control in the healthcare service (FOR-2005-06-17-610) and was approved by the Akershus University Hospital's Data Protection Official (2020_62). All analyses of sensitive data were undertaken on the secure platform provided by Services for sensitive data (TSD), University of Oslo.

4. SUMMARY OF RESULTS

4.1 Study I

In this study, HPV16 and HPV18 positive cervical cell samples were sequenced using the TaME-seq protocol to investigate type-specific MNVs and integration patterns. In total, 80 HPV16 and 51 HPV18 samples passed the filtering criteria of 300× sequencing depth and were analysed. Samples were stratified into the diagnostic categories non-progressive (HPV16 n=21, HPV18 n=12), CIN2 (HPV16 n=27, HPV18 n=9), CIN3/AIS (HPV16 n=27, HPV18 n=30) and cervical cancer (HPV16 n=5, HPV18 n=0). 1.05 billion reads were analysed and on average the samples had 77.7% of the genome covered by a minimum depth of 100×.

3747 MNVs were called in 131 samples, with no significant differences in amount or frequency between the diagnostic categories or HPV types. HPV18 E4 gene showed most overall variation compared to all other genomic elements in all diagnostic categories, while HPV16 E7 had a dN/dS ratio <1 in the cancer category. HPV16 non-progressive and CIN2 samples had APOBEC3-related C>T nucleotide substitutions, while this was not observed for HPV18 samples.

Integration frequencies were found to differ between the HPV types, with HPV18 positive samples having a significantly higher proportion of samples with at least one integration (30/51) compared to HPV16 (10/80). When combining the integrations for the HPV types, a significant part of the integration breakpoints were found to be located in HPV genes E1 and E2. The percentage of breakpoints in the human genome that were in, or close to, cancer-related genes were found to increase according to diagnostic severity, with 34%, 38% and 65% in non-progressive, CIN2 and CIN3/AIS, respectively, as well as in all cancer samples. In addition to adding to the growing evidence of HPV intra-host variation, the study gives insight into dissimilar genomic alterations between the two HPV types, which might reflect mechanistic differences in how they induce cell transformation.

4.2 Study II

Study II builds upon **study I**, by sequencing cervical cell samples from women positive with HPV31, 33 and 45. Samples from study I (HPV16 and HPV18) were re-analysed using an updated pipeline, and the study allowed for the comparison of MNV and integration patterns between HPV-types belonging to the two distinct phylogenetic clades Alpha-7 (HPV18 and 45) and Alpha-9 (HPV16, 31 and 33). 354 samples, stratified into the diagnostic categories non-progressive, CIN2 and CIN3+, passed the filtering criteria of an average sequencing depth $>300\times$ (HPV16 n = 77, HPV18 n = 49, HPV31 n = 84, HPV33 n = 88, HPV45 n = 56).

A total of 10664 MNVs were called in the 354 analysed samples. While no significant differences in the amount of MNVs between the diagnostic categories were found, HPV45 had significantly more MNVs in all diagnostic categories compared to the other four HPV types. Most variation was found within the E4 gene from Alpha-7 positive samples in the CIN2-category. Alpha-9 samples showed APOBEC3-related C>T nucleotide substitutions, something that was not observed for Alpha-7. For HPV16-positive samples the APOBEC3-related substitutions were observed to decrease with an increase in diagnostic severity.

In total, 154 integration sites were reported, with 85% (131/154) of the reported integration sites belonging to Alpha-7 positive samples. For Alpha-7, the proportion of samples with integrations increased with increase in diagnostic severity (21% non-progressive, 33% CIN2, 61% CIN3+). The proportion of Alpha-7 samples with at least one integration was 42.8% compared to 6.4% for Alpha-9 samples. Alpha-7 samples also had on average more integrations per sample in all diagnostic categories compared to Alpha-9. Within the Alpha-7, HPV45 only had integrations reported in the CIN3+ category while HPV18 had integrations in all diagnostic categories, and in Alpha-9 HPV16 had more integrations compared to HPV31 and HPV33.

For Alpha-7 and Alpha-9 combined, integration breakpoints in the HPV genome revealed that 38%, 36% and 51% of the breakpoints were in the E1 or E2 genes for the non-progressive, CIN2 and CIN3+ categories, respectively. Integration breakpoints in the human genome revealed that 41% (12/29), 40% (10/25) and 59% (59/100) of reported breakpoints were $\pm 10\text{kb}$ of human cancer-related genes in the non-progressive, CIN2 and CIN3+ samples, respectively. Overall, the inclusion of three HR-HPV types from the Alpha-7 and Alpha-9 clades reveals that the type-specific patterns found in **study I** extends to the more closely related HPV types. The results broaden our understanding of the molecular mechanisms behind HPV-induced cancers and sheds light on similarities and differences between the HPV types investigated.

4.3 Study III

Contact tracing was conducted around 68 HCWs from 38 wards at Akershus University Hospital, identifying five wards where intra-hospital outbreaks could not be ruled out. Following the ARTIC nCoV-2019 sequencing protocol v2 (https://www.protocols.io/view/ncov-2019-sequencing-protocol-v2-bp2l6n26rgqe/v2?version_warning=no), 46 samples from Akershus University Hospital were sequenced on the Oxford Nanopore GridION, where 21 samples from HCWs and 2 patient samples came from wards with suspected intra-hospital outbreaks. An additional 8 HCWs from wards where there was no suspicion of intra-hospital outbreaks, and 10 anonymized patient samples were added to increase the comparative background. 5 samples were sequenced two times to assess the reproducibility of the method. All re-sequenced samples had the same variants called in the parts of the genome where they shared coverage.

After filtering out sequences with <80% genome coverage, the analysed samples consisted of 24 HCWs (18 from wards with suspected outbreaks), 2 patients and 7 anonymous samples. The remaining samples had an average genome coverage of 95.5%. In total, 273 variants were called relative to the reference genome over 62 sites, and the average number of variants per sample was 8.3. To assess whether suspected intra-hospital outbreaks should be confirmed or refuted, the presence of study-unique mutations in sample consensus sequences were investigated. The five suspected outbreaks were termed outbreak A-E and resolved or refuted based on the combination of contact tracing data and viral genomes.

With the inclusion of whole genome sequencing data, one outbreak (A) was refuted due to the viral genomes belonging to different clades, while two other outbreaks (B and E) were confirmed due to shared study-unique mutations. Outbreak C and D was more difficult to resolve due to having similar viral genomes but not necessarily having study-unique mutations in common. However, the combination of epidemiological data and similarities in viral genomes makes it likely that they would be part of the same transmission chain, although it can not be confirmed with confidence. An additional outbreak (F) was discovered when WGS data was included, which included two HCWs with identical viral genomes who had worked in the same ward, but with no reported close contact. Overall, the study illustrates the usefulness of combining SARS-CoV-2 WGS data with epidemiological data to increase the resolution when doing outbreak investigations.

5. DISCUSSION

5.1 Methodological considerations

5.1.1 Sample material

Sample material used in **study I** and **study II** consisted of clinical cervical cell samples which had previously been collected, DNA extracted and analysed for HPV, including genotyping. Previously extracted DNA were stored in a -80°C freezer, which is considered proper long-term storage of DNA to maintain DNA quality and integrity[265]. TaME-seq uses HPV type-specific primer pools to amplify and enrich target HPV genome. HPV genotyping do sometimes give incorrect results[266], which would cause the type-specific primers to not work, causing neither amplification nor sequencing would give results. Not all samples were at the recommended input DNA concentration for the Nextera tagmentation reaction; however, the reaction has been shown to perform well with low amounts of input material[267]. Furthermore, since TaME-seq enriches HPV target sequences, the viral load in the samples affects the performance of TaME-seq. Low viral load has been shown to correlate strongly with the mean sequencing coverage of samples[110]. High-grade lesions have been shown to have a higher HPV viral load than normal or low-grade lesions[268], and low viral load could have given low coverage or a failed sequencing reaction in some samples. The sample material was clinical cervical cell samples of different diagnostic categories positive for different HPV-types. While a study design with an equal number of samples for all diagnostic categories for all HPV types would have been preferred, this was not obtainable. This is due to difference in prevalence of the different HPV types and the genotype distribution between the diagnostic categories.

Sample material used in **study III** consisted of extracted RNA from naso-/oropharyngeal samples positive for SARS-CoV-2 from healthcare workers and patients, stored in -80°C. The ARTIC-protocol used to enrich for SARS-CoV-2 sequences is amplicon-based, and like TaME-seq performs better with higher viral load. It is possible that samples with low viral load is the main factor responsible for failed sequencing reactions and low genome coverage, as we observed better sequencing performance in samples with Ct-values <33. Because we did not have resources to sequence viral genomes from all patients who had been cared for by the HCWs or all HCWs from the same wards as those infected, we were not able to assess the degree of cryptic nosocomial transmission.

5.1.2 Library preparation and sequencing

The library preparation protocol used in **study I** and **study II** was TaME-seq. The protocol uses a combination of Nextera transposome tagmentation and multiplexed target-specific primers to amplify HPV sequences as well as hybrid HPV-human sequences[258]. This allows for the simultaneous investigation of intra-sample viral genomic variation as well as viral integration into host chromosomes while producing relatively few off-target sequences[258]. The method is PCR-based, using HPV type-specific primers that are designed to evenly cover the individual HPV type genomes and suboptimal primer design can result in lower genome coverage and sequencing depth, poor sequencing alignment and generally less than optimal performance. Furthermore, PCR reactions can introduce technical errors besides those introduced by differences in primer efficiency. The sequences themselves can affect the PCR efficiency, where high GC content and the formation of secondary structures can reduce the amplification efficiency of certain sequence regions, in addition to stochasticity which might cause low copy number sequences not to be amplified[269,270]. PCR reactions also uses DNA polymerases to synthesise new DNA strands which are inherently error-prone, generating both single nucleotide substitutions and indels, which might be difficult to disentangle from true biological variation[271]. Sequencing was performed on the Illumina platform, one of the most common sequencing platforms in use, able to generate the most sequence data[272]. For this dataset, Illumina was the most suitable due to its ability to generate massive amounts of sequencing data for a relatively low cost per base, as well as its low error rate which makes it suitable to study low-frequency MNVs[273].

In **study II**, the ARTIC V3 library preparation protocol was used. The ARTIC protocol amplifies SARS-CoV-2 genomes using tiled, multiplexed primers. As the protocol is PCR-based, the same errors discussed above for TaME-seq can be introduced by this method. Additionally, it has been shown that the primers are sensitive to substitutions or deletions in the primer-binding regions, and the primer set has been redesigned several times as SARS-CoV-2 lineages have acquired mutations. The sequencing was done on the Nanopore platform. Nanopore sequences are known to have a high per-base error rate[273], but as errors are normally distributed randomly (with the exception for homopolymeric regions) and the method

is designed to generate consensus genomes from many amplicons, the consensus genomes generated contain close to no sequencing errors[274].

5.1.3 Bioinformatic analyses

To analyse the TaME-seq sequencing data from **study I** and **study II**, an in-house bioinformatics pipeline using several Python and R scripts, executed using the Snakemake workflow management system[261]. The aim of the bioinformatics analyses was to investigate intra-host MNV and integration patterns of HPV in cervical cell samples of different diagnostic categories.

The MNV calling was based on the relative highest and second highest coverage of each nucleotide position of the genome to call major and minor alleles, respectively. This was done in a per-sample and reference genome independent manner. This approach allows the identification of the genomic diversity of HPV types and their sublineages in the collection of samples investigated. There is a trade-off in variant calling between retaining true and false positives. By setting the variant calling threshold low, the number of true positives increases, but so does the risk of calling false positives generated from PCR and sequencing errors[275]. Vice versa, setting the calling threshold high will increase the number of true and false negatives[276]. To make sure true low-frequency MNVs were called, a 1% variant calling threshold was set for **study I** and **study II**, and the variant calling pipeline utilizes a stepwise evaluation of MNVs from both the forward and reverse reactions to minimize the risk of calling false positive MNVs. Additionally, since the NCR genomic region is known to contain homopolymeric T tracts which can cause polymerase or sequencing errors, these were excluded from the variant calling analysis to reduce the number of false positive MNVs.

The identification of integrations was done in a two-step analysis using both discordant reads and junction reads. The two-step analysis strengthens each other and allows for the identification of rare and low-frequency integrations with more confidence, as well as identifying the exact breakpoint coordinates in some cases. Removal of PCR duplicates are necessary, as their inclusion would result in too many false positives. This might result in the loss of low-frequency true positive integrations due to having too few reads covering the integration site, but the filtering is essential and strict in order not to report too many false positives. Filtering out reported integrations with human breakpoints in homopolymeric regions are also essential to remove false positive integrations. Validation of reported integration sites

was done using Sanger sequencing, which is considered the most common method to validate NGS results[277]. While many integration sites were validated, some were not. In addition to being potential false positives, these could be due to suboptimal PCR amplification, off-target primer hybridization or genomic structural rearrangements in the vicinity of the integrations site[278,279].

The bioinformatics analysis in **study III** used a pipeline developed by the ARTIC network, the same group that developed the SARS-CoV-2 whole genome sequencing protocol. By using a pipeline developed by the same researchers who developed the WGS protocol and primer scheme, the pipeline was already optimized for the generation of consensus genomes of high quality. The Medaka-pipeline was used instead of Nanopolish, since it is not relying on raw signal data in the form of FAST5-files of considerable size and quicker, while achieving close to similar results[274]. To assess the reproducibility of the variants called from the consensus genomes, five samples were sequenced two times and they all called identical variants where they shared coverage.

5.1.4 Statistics

In **study I**, a Kruskal-Wallis test was used to test for significant differences between HPV16 and HPV18. It is a non-parametric and was chosen due to the non-normal distribution of the test variables, and a significant result indicates group differences, but not which groups that differ.

In **study II**, a non-parametric Chi square test of independence was used to analyse if differences between the groups were significant in respect to integration breakpoints in the HPV genome. A Chi square test of independence is considered one of the most useful statistics for testing hypothesis where the variables are categorical[280]. A generalized linear model (glm) was used when testing dependent and independent variables. The distribution chosen for the glm was chosen according to the dependent variable to avoid overdispersion. A negative binomial distribution was used when testing the number of MNVs and number of integrations, as this is more suitable for overdispersed count data. The rest of the tests uses a binomial distribution, which is most suitable for presence/absence data.

The standard deviation in our dataset was large with relatively small sample sizes, something that can contribute to non-significant statistical results. Larger sample sizes are needed to confirm statistically significant findings.

5.2 Discussion of results

5.2.1 HPV intra-host variation and integration frequencies

Traditionally, research into HPV lineages and sublineages and their related carcinogenic potential has studied genomic variation using consensus genomes. However, by solely investigating genomic variation based on consensus genomes one loses out on variation found below the consensus level found within infected persons. As deep sequencing studies have become more common, it has been revealed that HPV infections contain many low-frequency MNVs which might play a role in sustaining persistence of HPV infections and development of HPV-induced cancers[109–113]. Most studies have investigated MNVs from clinical samples in HPV16 (except for one study including also HPV52 and 58), and less is known of these events in other carcinogenic HR-HPV types. That is why **study I** investigated these events in both HPV16 and HPV18. In **study II** we expand upon **study I** by including three additional HR-HPV types, HPV31, 33 and 45. HPV16, 31 and 33 sort under the Alpha-9 subclade and HPV18 and 45 under Alpha-7, and the aim was to investigate if similarities and differences found in **study I** extends to phylogenetically related HPV types.

Our results from **study I and II** found the presence of numerous MNVs from cervical cell samples of different diagnostic categories positive for HPV16, 18, 31, 33 and 45. When comparing the total number of MNVs and their frequency, it was revealed that HPV45 positive samples had significantly more MNVs at a higher frequency than the four other HPV types. Since MNVs are called in a reference-independent manner, we rule out reference-divergence as a source of inflated number of MNVs. Six samples with indicative patterns of co-infections were also removed from the analysis. However, we cannot completely rule out the possibility that co-infections of different HPV45 sublineages were causing some of the intra-host variation observed. The rest of the HPV types had numbers of called MNVs and frequencies that were not significantly different from each other. Neither the total number of MNVs nor their frequency were statistically significant between the diagnostic categories within an HPV type. When investigating where in the HPV genomes the variation was located, it was revealed that HPV18 and HPV45, which both sort under Alpha-7, have more variation in the E4 gene than HPV types belonging to Alpha-9. The biological significance of these differences is currently not clear, and more research is needed. Results from **study I** showed that the E7 gene in HPV16

cancer samples had very few non-synonymous mutations, indicative of purifying selection to conserve the function of the gene. This is in line with similar studies, suggesting that conservation of E7 function is critical for the development of HPV16-induced cervical cancer[112].

Several studies have investigated APOBEC3-induced C>T mutations in HPV[110,111,113]. APOBEC3-induced mutations have previously been observed in HPV16, 52 and 58, and have been found to decrease with increase in diagnostic severity of samples[111]. **Study I** was to our knowledge the first study investigating APOBEC3-induced substitutions in HPV18 positive samples. The study revealed the presence of APOBEC3-related MNV profiles in low-grade lesions positive for HPV16, but not in high-grade lesions. The result is in concordance with other studies showing that APOBEC3-activity is associated with transient and benign HPV infections[111,113]. This suggests that in high-grade lesions caused by persistent HPV16 infections the virus might be able to evade APOBEC3-activity by unknown mechanisms. No APOBEC3-related MNV profile was observed for HPV18. In **study II** we expanded the analysis to include two HPV types from Alpha-7 (HPV18 and 45) and three Alpha-9 (HPV16, 31 and 33). It was revealed that Alpha-9 types HPV31 and 33 also had APOBEC3-induced C>T mutations in the trinucleotide context TCA, while this signal was absent in both Alpha-7 types. Interestingly, the proportion of APOBEC3-mutations were not found to decrease with lesion severity for HPV31 and 33, suggesting that the ability to evade APOBEC3-activity might be a feature of HPV16 carcinogenesis and not an Alpha-9 specific tendency. HPV31 and 33 have been shown to have a high risk to progress to CIN3, but relative to HPV16, 18 and 45, their risk to progress to invasive cervical cancer is relatively low[281], and it can be speculated that this might be partially explained by their inability of evade APOBEC3-activity. Furthermore, the results indicate that APOBEC3-induced mutagenesis of viral genomes is not a general feature of HR-HPV infections, as it was not detected in Alpha-7 samples.

In line with other studies, our results show a significantly higher integration frequency in Alpha-7 than Alpha-9, and that HPV16 have a higher integration frequency than HPV31 and 33[211,215]. Alpha-7 positive samples with integrations also had more integrations per sample compared to Alpha-9. Additionally, a significant increase in integration frequency with increased diagnostic severity was observed for Alpha-7. Integrations are associated with increased genomic instability and progression to invasive cervical cancer[222,282,283]. Our results show that the percentage of all integrations combined with breakpoints in E1 or E2, which can cause overexpression of oncogenes E6/E7, increases with increased diagnostic

severity. The same pattern is found when investigating the presence of cancer-related genes (CRGs) 10 kb upstream or downstream of the integration site. Integrations are known to cause local genomic structural rearrangements and to affect host gene expression in their vicinity[200,205,284]. Many of the CRGs observed in our dataset are also found in similar studies, indicating that some viral integrations into host chromosomes might confer a selective growth advantage and contribute to HPV-induced carcinogenesis of the affected cells in a location-specific manner[203,283]. Additionally, we observed integrations located close to RCAN2, KLHL29 and MIR205HG twice in independent samples, all genes associated with cancers in some way or another[285–288]. MIR205HG is in several studies implicated in playing a role in the development of cervical cancer[287–289], but the observation of integration close to these two other genes might suggest they could also play a role, an observation that should be pursued further.

Taken together, the results suggest conserved differences between HPV types belonging to Alpha-7 and Alpha-9. What drives the differences in integration frequencies and MNV profiles between the HPV-types is presently not completely understood. Alpha-7 types are more associated with adenocarcinoma and lesions in glandular cells while Alpha-9 types are more associated with squamous cell carcinoma and lesions in squamous cells[281]. The differences observed could be explained by different host-responses between different infected cells and/or differences in HPV genomic variation. It is already known that HPV16 and HPV18 differs in oncogene splice variants, capability to induce p53-degradation, integration frequencies, DNA methylation patterns and tumour gene expression profiles, among other traits[210,211,215,290–292]. **Study I** and **study II** expands upon this knowledge by investigating integration frequencies and MNV and APOBEC3 mutation profiles of HPV16 and HPV18 and revealing that these differences extend to their phylogenetically related high-risk HPV types HPV31, 33 and 45. Although the carcinogenic processes of different HR-HPV types have many similarities, there is growing evidence that they differ in their ability to infect different cervical cells and in molecular mechanisms behind HPV-induced cervical cancer.

5.2.2 Nanopore whole genome sequencing of SARS-CoV-2 to investigate intra-hospital transmission

While **study I** and **study II** were deep sequencing studies investigating HPV intra-host nucleotide variation and viral integration sequenced on the Illumina platform, **study III**

investigated intra-hospital SARS-CoV-2 transmission using consensus genomes sequenced on the Nanopore platform. **Study III** was done during the first wave of the pandemic, at a time when the knowledge of SARS-CoV-2 was in its infancy. There was a pressing need to better understand the main routes of virus transmission, viral reproductive number, severity of infection and co-morbidities, in addition to other characteristics that are easily taken for granted. It was of the utmost importance to prevent nosocomial infections to not increase patient morbidity and mortality[240,293]. Additionally, it was important to prevent infection of HCWs to prevent HCW-patient transmissions and an understaffed healthcare system.

The routine contact tracing identified five suspected intra-hospital outbreaks of SARS-CoV-2 during the time period of the study. Overall, the study highlights the benefit of including whole genome sequencing (WGS) data when doing outbreak investigations. With the inclusion of SARS-CoV-2 WGS data, we gained another level of information on which to confirm or refute these suspected outbreaks. The increase in resolution allowed for the confirmation of two suspected outbreaks and the refutation of another while two remained uncertain. A possible cryptic outbreak not identified during routine contact tracing was discovered with the inclusion of WGS data.

Another aspect the study brings to light is the complexity in determining whether outbreaks are nosocomial or not. Several of the suspected outbreaks had viral genomes of high similarity, but with one or more variants not shared with each other, which makes it difficult to determine whether the infections were due to direct transmission or independently acquired. During the time when the study was conducted there were very few available genomes from Norway online and the viral diversity was very low. This makes it difficult to assess whether two infected individuals with similar, but not identical, genomes were part of the same transmission chain or not[237,294]. Therefore, the presence of shared study-unique variants was emphasised when investigating the suspected outbreaks, while many similar studies use predefined cut-offs of maximum allowed differences to define outbreaks[237,295–297]. However, with the inclusion of contact tracing data this obstacle can to an extent be overcome, and the suspected outbreaks in question can be assumed to be more likely nosocomial than not. The combination of 1. recorded close contact between the suspected cases, 2. similar viral genomes and 3. that HCWs were under strict regulations to avoid infections at work and under national lockdown between March 12th and July 15th, makes it more likely that the cases resulted from nosocomial transmissions.

Several similar studies have been published during the ongoing pandemic, showing the power of combining epidemiological data with viral genomic data[237,295–299]. The studies have highlighted the benefits of real-time outbreak investigations. In-hospital studies are especially helpful when investigating transmission dynamics and genetic variability of SARS-CoV-2 due to the hospital environment being tightly monitored and controlled. In **study III** we observed the gain of new mutations from the same outbreaks. This might be due to the viral genomes acquiring new mutations during the course of infections which are then transmitted to new individuals. Longitudinal studies of SARS-CoV-2 intra-host genomic variation have revealed the presence of numerous MNVs that shift in frequency as an infection progresses[91,120,123]. Additionally, studies that have investigated the transmission bottleneck of SARS-CoV-2 discovered that, even if the transmission bottleneck is narrow, MNVs are transmitted to new hosts[118,119,124,298]. We observed that samples acquired late in the transmission chain relative to suspected primary cases (>8 days) carry novel variants, suggesting that novel variants are generated within infected individuals and the likelihood of them being transmitted increases as the infection progresses. The mutation rate of SARS-CoV-2 is estimated to result in approximately 2-3 mutations per month[300,301], however, the acquisition of mutations in genomes and their transmission is a stochastic process[302]. Another possible explanation could be cryptic transmission between asymptomatic individuals, but as we did not have the resources to sequence the viral genomes of all patients cared for by the HCWs or all HCWs from affected wards, investigating cryptic transmission and possible transmission between HCWs wearing personal protective equipment and patients in general was not possible.

5.3 Significance of results and future perspectives

Future research into MNV and integration profiles of different HPV types should strive to include more samples in all clinical categories, especially so in the normal and low-grade lesion categories. This would allow for a more thorough exploration of these genomic events and their differences between different categories. Longitudinal sampling would also allow to investigate these events and how they change over time, allowing us to further investigate and verify if the findings have prognostic value in assessing risk of persistent HPV infections and cervical cancer development. Additionally, follow-up data from the national cancer registry detailing cancer development and interventions would better inform such risk assessment. Furthermore, TaME-seq can be used in epidemiological studies of HPV infections to study genomic diversity per se and in vaccine surveillance studies. The intra-host events investigated in **study I** and

study II could have an impact on vaccine efficacy and could be used in the development of new and better targeting HPV vaccines, both prophylactic and therapeutic. The TaME-seq protocol can easily be adapted to study these events also in other viruses where intra-host MNVs and/or integrations are thought to be of relevance in disease severity or developing resistances to therapeutic treatments or vaccines[303–306].

Study I adds to the growing evidence that HPV16 and HPV18 differ in their molecular biology and suggests that their carcinogenicity may manifest through parallel mechanistic routes to HPV-induced cancer[52,290–292]. **Study II** strengthens the findings by showing that these differences extend to the closely related HR-HPV types from the Alpha-7 and Alpha-9 clades, HPV31, 33 and 45. Studying and comparing similarities and differences of these genomic events between different HR-HPV types can provide knowledge of different molecular mechanisms behind HPV-induced cancers and help explain why some types are more prevalent in cancers than others and more carcinogenic.

Persistent HR-HPV infections are considered necessary for the development of HPV-induced cervical cancer[30]. However, as most infections are eventually cleared by the immune system, persistent infection is in itself an insufficient driver[23]. Intra-host molecular events like the generation of MNVs and integrations can influence whether an infection progresses to precancerous lesions and cancer, and one of the long-term aims of the project **study I** and **study II** is to see if these molecular events can be used in disease risk assessment. Currently, the cervical cancer screening programme's method for preventing cervical cancer is to detect early HPV infections and triage precancerous lesions[307]. It is acknowledged that this regimen causes overtreatment of precancerous lesions that could have regressed in its own given time[181]. Future studies should assess whether MNVs in certain genomic positions or viral integration of certain characteristics are associated with the development of precancerous lesions and cervical cancer to bring us one step closer to having a personalised cervical cancer risk assessment.

Study III gave insight into the resolution one can obtain in outbreak investigations by combining epidemiological data with genomic data. Relying on a yes/no answer from a PCR test and contact tracing information alone makes it difficult to ascertain whether an infection was nosocomial in nature or community-acquired. By including information of all variants present in the viral genomes sampled from individuals within the same suspected transmission

chain, we can more easily refute cases that are not connected and confirm true transmission events. By sequencing on the Nanopore platform, one can gain near real-time knowledge of hospital transmission and effectuate necessary infection control measures to limit spread[237,297]. This is extremely valuable in a hospital setting, as an excess of infected HCWs would strain the healthcare service, and the consequence of HCW-patient transmission can cause unnecessary SARS-CoV-2 related deaths and post-covid. As several studies have shown, intra-host MNVs arise during SARS-CoV-2 infections as well. While the transmission bottleneck is supposedly narrow and transmission of MNVs between individual does not occur often, it would still add useful information one could use to confirm or refute outbreaks[118,119,124,298,308]. However, due to its high error rate, Nanopore would not be an appropriate sequencing instrument to find MNVs[274]. Inclusion of MNVs in outbreak investigations would then necessitate deep sequencing on for example the Illumina platform.

The research presented here laid the foundation for how SARS-CoV-2 outbreak investigations are conducted at Akershus University Hospital today. As the sequencing now is done in-house, the time from detection of suspected outbreaks or infection with variants of concern to implementing necessary infection prevention measures, has drastically been reduced. The sequencing results in themselves are important, as being a part of the global sequencing effort has contributed to understanding the evolution and biology of SARS-CoV-2, the dynamics of the pandemic over time, and contributed with information used to aid and guide public health decisions[228,309]. The ongoing pandemic is far from over, as the dominating Omicron variant and its sublineages have shown. At the time of writing, cases of SARS-CoV-2 infections are increasing in several countries[310]. As with the HPV vaccine, there seems to be an inability to fairly distribute vaccines against SARS-CoV-2 infections between high-and upper-middle income countries and low- and lower-middle income countries[246,247]. Until we achieve equity in vaccine coverage, the world is still at risk of emerging SARS-CoV-2 variants that can have increased virulence to compromise the effects of diagnostics, therapeutics and immunity through infection or vaccines[311].

Overall, the thesis highlights how powerful of a tool NGS sequencing and viral genomics is to gain insight into human viral pathogens. On one hand, we are still able to attain new knowledge of one of the oldest human viral pathogens, HPVs, to further understand their molecular routes of cellular transformation that place such an immense burden on the global health, even with effective vaccines having been developed. On the other hand, NGS has been an extremely important tool in understanding a completely novel viral pathogen. SARS-CoV-2 rapidly

spread to all parts of the world, forced people social isolation for weeks and months at a time and still managed to cause the deaths of millions of individuals. The rapid sequencing of the first SARS-CoV-2 genomes allowed for the early development of vaccines, which started rolling out within the first year of the pandemic and reintroduced the populace to a more normal day-to-day life. The concerted global sequencing effort allowed for quick insights into how the virus spread, guiding public health decision-makers in how to best protect the most people. NGS and the knowledge we gain through it is the first line of defence against novel pathogens, and by building competence in viral genome sequencing and the associated bioinformatic analysis, the healthcare system is more prepared when new pandemics are upon us.

6. CONCLUSIONS

This thesis aimed at using Illumina sequencing and viral genomics to characterise HPV intra-host genomic variation and chromosomal integrations for different HR-HPV types and explore these events in HPV-induced carcinogenesis. Additionally, Nanopore sequencing was used to whole genome sequence SARS-CoV-2 genomes to combine genomic information with epidemiological data to increase the resolution of intra-hospital outbreak investigations.

The studies into intra-host HPV genomic variation revealed numerous MNVs at low frequency. While the amount did not differ between the diagnostic categories within the HPV types, HPV45 were found to have more MNVs at a higher frequency than the other HR-HPVs investigated. Notably, only HPV16, 31 and 33 showed APOBEC3-related nucleotide substitutions, while this was not found for HPV18 and 45. Thus, HR-HPV types from the Alpha-7 and Alpha-9 clades are shown to differ in their ability to trigger APOBEC3-activity. Additionally, samples from Alpha-7 types had a higher integration frequency than Alpha-9. The results add to the growing knowledge of the biological differences and similarities of different HR-HPV types, and that closely related HPV types are more similar in these traits than to those more distant.

The inclusion of SARS-CoV-2 WGS data was shown to be a powerful tool to include in intra-hospital outbreak investigations. Contact tracing alone falsely identified one hospital outbreak

and overlooked another. The rapid inclusion of WGS data can give a better understanding of nosocomial transmissions and aids in guiding local infection prevention and control routines at hospitals.

7. REFERENCES

- [1] Willemsen A, Bravo IG. Origin and evolution of papillomavirus (onco)genes and genomes. *Philos Trans R Soc B Biol Sci* 2019;374. <https://doi.org/10.1098/rstb.2018.0303>.
- [2] Bosch FX, de Sanjosé S. Chapter 1: Human papillomavirus and cervical cancer--burden and assessment of causality. *J Natl Cancer Inst Monogr* 2003;3–13. <https://doi.org/10.1093/OXFORDJOURNALS.JNCIMONOGRAPHS.A003479>.
- [3] Crosbie EJ, Einstein MH, Franceschi S, Kitchener HC. Human papillomavirus and cervical cancer. *Lancet* 2013;382:889–99. [https://doi.org/10.1016/S0140-6736\(13\)60022-7](https://doi.org/10.1016/S0140-6736(13)60022-7).
- [4] Moscicki AB, Schiffman M, Burchell A, Albero G, Giuliano AR, Goodman MT, et al. Updating the Natural History of Human Papillomavirus and Anogenital Cancers. *Vaccine* 2012;30:F24. <https://doi.org/10.1016/J.VACCINE.2012.05.089>.
- [5] Rodriguez AC, Schiffman M, Herrero R, Hildesheim A, Bratti C, Sherman ME, et al. Longitudinal study of human papillomavirus persistence and cervical intraepithelial neoplasia grade 2/3: critical role of duration of infection. *J Natl Cancer Inst* 2010;102:315–24. <https://doi.org/10.1093/JNCI/DJQ001>.
- [6] Walboomers JMM, Jacobs M V., Manos MM, Bosch FX, Kummer JA, Shah K V., et al. Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *J Pathol* 1999;189:12–9. [https://doi.org/10.1002/\(SICI\)1096-9896\(199909\)189:1<12::AID-PATH431>3.0.CO;2-F](https://doi.org/10.1002/(SICI)1096-9896(199909)189:1<12::AID-PATH431>3.0.CO;2-F).
- [7] De Martel C, Plummer M, Vignat J, Franceschi S. Worldwide burden of cancer attributable to HPV by site, country and HPV type. *Int Agency Res Cancer (IARC/WHO)* 2017;141:664–70. <https://doi.org/10.1002/ijc.30716>.
- [8] 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. *Int J Cancer* 2015;136:E359–86. <https://doi.org/10.1002/IJC.29210>.
- [9] Bosch FX, Broker TR, Forman D, Moscicki A-B, Gillison ML, Doorbar J, et al.

- Comprehensive control of human papillomavirus infections and related diseases. *Vaccine* 2013;31 Suppl 7:H1–31. <https://doi.org/10.1016/J.VACCINE.2013.10.003>.
- [10] Pimenoff VN, De Oliveira CM, Bravo IG. Transmission between Archaic and Modern Human Ancestors during the Evolution of the Oncogenic Human Papillomavirus 16. *Mol Biol Evol* 2017;34:4–19. <https://doi.org/10.1093/MOLBEV/MSW214>.
- [11] Hublin JJ, Ben-Ncer A, Bailey SE, Freidline SE, Neubauer S, Skinner MM, et al. New fossils from Jebel Irhoud, Morocco and the pan-African origin of *Homo sapiens*. *Nat* 2017 5467657 2017;546:289–92. <https://doi.org/10.1038/nature22336>.
- [12] Worobey M, Levy JI, Serrano LMM, Crits-christoph A, Pekar JE, Goldstein SA, et al. The Huanan market was the epicenter of SARS-CoV-2 emergence. *Zenodo* 2022. <https://doi.org/10.5281/ZENODO.6299600>.
- [13] Pekar AJE, Magee A, Parker E, Moshiri N, Havens JL, Gangavarapu K, et al. SARS-CoV-2 emergence very likely resulted from at least two zoonotic events. *Zenodo* 2022. <https://doi.org/10.5281/ZENODO.6291628>.
- [14] Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA, et al. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol* 2020;5:536–44. <https://doi.org/10.1038/S41564-020-0695-Z>.
- [15] Corman VM, Muth D, Niemeyer D, Drosten C. Hosts and Sources of Endemic Human Coronaviruses. *Adv Virus Res* 2018;100:163–88. <https://doi.org/10.1016/BS.AIVIR.2018.01.001>.
- [16] Wertheim JO, Chu DKW, Peiris JSM, Pond SLK, Poon LLM. A Case for the Ancient Origin of Coronaviruses. *J Virol* 2013;87:7039. <https://doi.org/10.1128/JVI.03273-12>.
- [17] Drosten C, Günther S, Preiser W, van der Werf S, Brodt H-R, Becker S, et al. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N Engl J Med* 2003;348:1967–76. <https://doi.org/10.1056/NEJMOA030747>.
- [18] Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus ADME, Fouchier RAM. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N Engl J Med* 2012;367:1814–20. <https://doi.org/10.1056/NEJMOA1211721>.
- [19] De Wit E, Van Doremalen N, Falzarano D, Munster VJ. SARS and MERS: recent

- insights into emerging coronaviruses. *Nat Rev Microbiol* 2016 148 2016;14:523–34.
<https://doi.org/10.1038/nrmicro.2016.81>.
- [20] World Health Organization. WHO Coronavirus Disease (COVID-19) Dashboard With Vaccination Data | WHO Coronavirus (COVID-19) Dashboard With Vaccination Data. *World Heal Organ* 2021:1–5.
- [21] de Villiers E-M, Fauquet C, Broker TR, Bernard H-U, zur Hausen H. Classification of papillomaviruses. *Virology* 2004;324:17–27.
<https://doi.org/10.1016/J.VIROL.2004.03.033>.
- [22] Bravo IG, Felez-Sanchez M. Papillomaviruses: Viral evolution, cancer and evolutionary medicine. *Evol Med Public Heal* 2015;2015:32–51.
<https://doi.org/10.1093/emph/eov003>.
- [23] Schiffman M, Doorbar J, Wentzensen N, De Sanjosé S, Fakhry C, Monk BJ, et al. Carcinogenic human papillomavirus infection. *Nat Rev Dis Prim* 2016;2.
<https://doi.org/10.1038/nrdp.2016.86>.
- [24] Smith B, Chen Z, Reimers L, van Doorslaer K, Schiffman M, DeSalle R, et al. Sequence imputation of HPV16 genomes for genetic association studies. *PLoS One* 2011;6. <https://doi.org/10.1371/JOURNAL.PONE.0021375>.
- [25] Graham S V. Human papillomavirus: gene expression, regulation and prospects for novel diagnostic methods and antiviral therapies. *Future Microbiol* 2010;5:1493–506.
<https://doi.org/10.2217/fmb.10.107>.
- [26] Frattini MG, Laimins LA. Binding of the human papillomavirus E1 origin-recognition protein is regulated through complex formation with the E2 enhancer-binding protein. *Proc Natl Acad Sci U S A* 1994;91:12398–402.
<https://doi.org/10.1073/PNAS.91.26.12398>.
- [27] McBride AA. The Papillomavirus E2 proteins. *Virology* 2013;445:57–79.
<https://doi.org/10.1016/J.VIROL.2013.06.006>.
- [28] Doorbar J. The E4 protein; structure, function and patterns of expression. *Virology* 2013;445:80–98. <https://doi.org/10.1016/J.VIROL.2013.07.008>.
- [29] Venuti A, Paolini F, Nasir L, Corteggio A, Roperto S, Campo MS, et al. Papillomavirus E5: the smallest oncoprotein with many functions. *Mol Cancer*

- 2011;10:140. <https://doi.org/10.1186/1476-4598-10-140>.
- [30] Doorbar J, Egawa N, Griffin H, Kranjec C, Murakami I. Human papillomavirus molecular biology and disease association. *Rev Med Virol* 2015;25 Suppl 1:2–23. <https://doi.org/10.1002/RMV.1822>.
- [31] Finnen RL, Erickson KD, Chen XS, Garcea RL. Interactions between Papillomavirus L1 and L2 Capsid Proteins. *J Virol* 2003;77:4818. <https://doi.org/10.1128/JVI.77.8.4818-4826.2003>.
- [32] de Sanjosé S, Brotons M, Pavón MA. The natural history of human papillomavirus infection. *Best Pract Res Clin Obstet Gynaecol* 2018;47:2–13. <https://doi.org/10.1016/J.BPOBGYN.2017.08.015>.
- [33] Maki H, Fujikawa-Adachi K, Yoshie O. Evidence for a promoter-like activity in the short non-coding region of human papillomaviruses. *J Gen Virol* 1996;77 (Pt 3):453–8. <https://doi.org/10.1099/0022-1317-77-3-453>.
- [34] Mandal P, Bhattacharjee B, Das Ghosh D, Mondal NR, Roy Chowdhury R, Roy S, et al. Differential Expression of HPV16 L2 Gene in Cervical Cancers Harboring Episomal HPV16 Genomes: Influence of Synonymous and Non-Coding Region Variations. *PLoS One* 2013;8:e65647. <https://doi.org/10.1371/JOURNAL.PONE.0065647>.
- [35] Fehr AR, Perlman S. Coronaviruses: An Overview of Their Replication and Pathogenesis. *Coronaviruses* 2015;1282:1. https://doi.org/10.1007/978-1-4939-2438-7_1.
- [36] Redondo N, Zaldívar-López S, Garrido JJ, Montoya M. SARS-CoV-2 Accessory Proteins in Viral Pathogenesis: Knowns and Unknowns. *Front Immunol* 2021;12:2698. <https://doi.org/10.3389/FIMMU.2021.708264/BIBTEX>.
- [37] Brant AC, Tian W, Majerciak V, Yang W, Zheng ZM. SARS-CoV-2: from its discovery to genome structure, transcription, and replication. *Cell Biosci* 2021 111 2021;11:1–17. <https://doi.org/10.1186/S13578-021-00643-Z>.
- [38] Yadav R, Chaudhary JK, Jain N, Chaudhary PK, Khanra S, Dhamija P, et al. Role of Structural and Non-Structural Proteins and Therapeutic Targets of SARS-CoV-2 for COVID-19. *Cells* 2021, Vol 10, Page 821 2021;10:821.

<https://doi.org/10.3390/CELLS10040821>.

- [39] Shang J, Han N, Chen Z, Peng Y, Li L, Zhou H, et al. Compositional diversity and evolutionary pattern of coronavirus accessory proteins. *Brief Bioinform* 2021;22:1267–78. <https://doi.org/10.1093/BIB/BBAA262>.
- [40] Schoeman D, Fielding BC. Coronavirus envelope protein: current knowledge. *Virology* 2019 161 2019;16:1–22. <https://doi.org/10.1186/S12985-019-1182-0>.
- [41] Astuti I, Ysrafil. Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2): An overview of viral structure and host response. *Diabetes Metab Syndr Clin Res Rev* 2020;14:407–12. <https://doi.org/10.1016/J.DSX.2020.04.020>.
- [42] Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. *Nature* 2020;579:265–9. <https://doi.org/10.1038/s41586-020-2008-3>.
- [43] Chen L, Li X, Chen M, Feng Y, Xiong C. The ACE2 expression in human heart indicates new potential mechanism of heart injury among patients infected with SARS-CoV-2. *Cardiovasc Res* 2020;116:1097–100. <https://doi.org/10.1093/CVR/CVAA078>.
- [44] Guo Y, Wang B, Gao H, Gao L, Hua R, Xu JD. ACE2 in the Gut: The Center of the 2019-nCoV Infected Pathology. *Front Mol Biosci* 2021;8:819. <https://doi.org/10.3389/FMOLB.2021.708336/BIBTEX>.
- [45] Human Reference clones – hpvcenter n.d. https://www.hpvcenter.se/human_reference_clones/ (accessed March 5, 2020).
- [46] Egawa N, Egawa K, Griffin H, Doorbar J, Egawa N, Egawa K, et al. Human Papillomaviruses; Epithelial Tropisms, and the Development of Neoplasia. *Viruses* 2015;7:3863–90. <https://doi.org/10.3390/v7072802>.
- [47] Bzhalava D, Eklund C, Dillner J. International standardization and classification of human papillomavirus types. *Virology* 2015;476:341–4. <https://doi.org/10.1016/j.virol.2014.12.028>.
- [48] PaVE: Papilloma virus genome database n.d. https://pave.niaid.nih.gov/#explore/reference_genomes/human_genomes (accessed February 6, 2020).

- [49] Burk RD, Harari A, Chen Z. Human papillomavirus genome variants. *Virology* 2013;445:232–43. <https://doi.org/10.1016/j.virol.2013.07.018>.
- [50] Bernard HU, Burk RD, Chen Z, van Doorslaer K, Hausen H zur, de Villiers EM. Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology* 2010;401:70–9. <https://doi.org/10.1016/j.virol.2010.02.002>.
- [51] Burk RD, Chen Z, Van Doorslaer K. Human Papillomaviruses: Genetic Basis of Carcinogenicity. *Public Health Genomics* 2009;12:281–90. <https://doi.org/10.1159/000214919>.
- [52] Egawa N, Wang Q, Griffin HM, Murakami I, Jackson D, Mahmood R, et al. HPV16 and 18 genome amplification show different E4-dependence, with 16E4 enhancing E1 nuclear accumulation and replicative efficiency via its cell cycle arrest and kinase activation functions. *PLoS Pathog* 2017;13:e1006282. <https://doi.org/10.1371/journal.ppat.1006282>.
- [53] Tjalma WA, Fiander A, Reich O, Powell N, Nowakowski AM, Kirschner B, et al. Differences in human papillomavirus type distribution in high-grade cervical intraepithelial neoplasia and invasive cervical cancer in Europe. *Int J Cancer* 2013;132:854–67. <https://doi.org/10.1002/IJC.27713>.
- [54] Mirabello L, Yeager M, Cullen M, Boland JF, Chen Z, Wentzensen N, et al. HPV16 Sublineage Associations with Histology-Specific Cancer Risk Using HPV Whole-Genome Sequences in 3200 Women. *J Natl Cancer Inst* 2016;108. <https://doi.org/10.1093/jnci/djw100>.
- [55] Chen AA, Gheit T, Franceschi S, Tommasino M, Clifford GM. Human Papillomavirus 18 Genetic Variation and Cervical Cancer Risk Worldwide. *J Virol* 2015;89:10680. <https://doi.org/10.1128/JVI.01747-15>.
- [56] Clifford GM, Tenet V, Georges D, Alemany L, Pavón MA, Chen Z, et al. Human papillomavirus 16 sub-lineage dispersal and cervical cancer risk worldwide: Whole viral genome sequences from 7116 HPV16-positive women. *Papillomavirus Res* 2019;7:67–74. <https://doi.org/10.1016/j.pvr.2019.02.001>.
- [57] Cornet I, Gheit T, Iannacone MR, Vignat J, Sylla BS, Del Mistro A, et al. HPV16 genetic variation and the development of cervical cancer worldwide. *Br J Cancer* 2013

- 1081 2012;108:240–4. <https://doi.org/10.1038/bjc.2012.508>.
- [58] Chan PKS, Zhang C, Park JS, Smith-McCune KK, Palefsky JM, Giovannelli L, et al. Geographical distribution and oncogenic risk association of human papillomavirus type 58 E6 and E7 sequence variations. *Int J Cancer* 2013;132:2528–36. <https://doi.org/10.1002/IJC.27932>.
- [59] Chen Z, Schiffman M, Herrero R, DeSalle R, Anastos K, Segondy M, et al. Classification and evolution of human papillomavirus genome variants: Alpha-5 (HPV26, 51, 69, 82), Alpha-6 (HPV30, 53, 56, 66), Alpha-11 (HPV34, 73), Alpha-13 (HPV54) and Alpha-3 (HPV61). *Virology* 2018;516:86–101. <https://doi.org/10.1016/J.VIROL.2018.01.002>.
- [60] IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Meeting (2008-2009) : Lyon F, International Agency for Research on Cancer., World Health Organization. Volume 100 B. A review of human carcinogens. *Monogr Eval Carcinog Risks Hum* 2012;100:1–144.
- [61] Chen Z, Ho WCS, Boon SS, Law PTY, Chan MCW, DeSalle R, et al. Ancient Evolution and Dispersion of Human Papillomavirus 58 Variants. *J Virol* 2017;91. <https://doi.org/10.1128/JVI.01285-17>.
- [62] Xu HH, Zheng LZ, Lin AF, Dong SS, Chai ZY, Yan WH. Human papillomavirus (HPV) 18 genetic variants and cervical cancer risk in Taizhou area, China. *Gene* 2018;647:192–7. <https://doi.org/10.1016/J.GENE.2018.01.037>.
- [63] Rambaut A, Holmes EC, O’Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* 2020;5:1403–7. <https://doi.org/10.1038/s41564-020-0770-5>.
- [64] SARS-CoV-2 clade naming strategy for 2022 n.d. <https://nextstrain.org/blog/2022-04-29-SARS-CoV-2-clade-naming-2022> (accessed June 14, 2022).
- [65] Tao K, Tzou PL, Nouhin J, Gupta RK, de Oliveira T, Kosakovsky Pond SL, et al. The biological and clinical significance of emerging SARS-CoV-2 variants. *Nat Rev Genet* 2021 2212 2021;22:757–73. <https://doi.org/10.1038/s41576-021-00408-x>.
- [66] World Health Organization. Tracking SARS-CoV-2 variants. Who 2021:<https://www.who.int/en/activities/tracking-SARS-Co>.

- [67] Han X, Ye Q. The variants of SARS-CoV-2 and the challenges of vaccines. *J Med Virol* 2022;94:1366–72. <https://doi.org/10.1002/JMV.27513>.
- [68] Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* 2020;182:812–827.e19. <https://doi.org/10.1016/j.cell.2020.06.043>.
- [69] Telenti A, Hodcroft EB, Robertson DL. The Evolution and Biology of SARS-CoV-2 Variants. *Cold Spring Harb Perspect Med* 2022;12:a041390. <https://doi.org/10.1101/cshperspect.a041390>.
- [70] Dearlove B, Lewitus E, Bai H, Li Y, Reeves DB, Joyce MG, et al. A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. *Proc Natl Acad Sci U S A* 2020;117:23652–62. <https://doi.org/10.1073/pnas.2008281117>.
- [71] Rausch JW, Capoferri AA, Katusiime MG, Patro SC, Kearney MF. Low genetic diversity may be an Achilles heel of SARS-CoV-2. *Proc Natl Acad Sci* 2020;117:202017726. <https://doi.org/10.1073/pnas.2017726117>.
- [72] COG-UK. COG-UK update on SARS-CoV-2 Spike mutations of special interest Report 1 n.d. https://www.cogconsortium.uk/wp-content/uploads/2020/12/Report-1_COG-UK_19-December-2020_SARS-CoV-2-Mutations.pdf (Dead link, PDF found here: https://www.attogene.com/wp-content/uploads/2020/12/Report-1_COG-UK_19-December-2020_SARS-CoV-2-Mutations.pdf) (accessed August 4, 2022).
- [73] Davies NG, Abbott S, Barnard RC, Jarvis CI, Kucharski AJ, Munday JD, et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* (80-) 2021;372. <https://doi.org/10.1126/science.abg3055>.
- [74] Faria NR, Mellan TA, Whittaker C, Claro IM, Candido DDS, Mishra S, et al. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science* 2021;372. <https://doi.org/10.1126/SCIENCE.ABH2644>.
- [75] Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, et al. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nat* 2021 5927854 2021;592:438–43. <https://doi.org/10.1038/s41586-021-03402-9>.
- [76] Martin DP, Weaver S, Tegally H, San JE, Shank SD, Wilkinson E, et al. The

- emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. *Cell* 2021;184:5189. <https://doi.org/10.1016/J.CELL.2021.09.003>.
- [77] Cherian S, Potdar V, Jadhav S, Yadav P, Gupta N, Das M, et al. SARS-CoV-2 Spike Mutations, L452R, T478K, E484Q and P681R, in the Second Wave of COVID-19 in Maharashtra, India. *Microorganisms* 2021;9. <https://doi.org/10.3390/MICROORGANISMS9071542>.
- [78] Mlcochova P, Kemp S, Dhar MS, Papa G, Meng B, Ferreira IATM, et al. SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. *Nature* 2021;599:114–9. <https://doi.org/10.1038/s41586-021-03944-y>.
- [79] Viana R, Moyo S, Amoako DG, Tegally H, Scheepers C, Althaus CL, et al. Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nat* 2022 6037902 2022;603:679–86. <https://doi.org/10.1038/s41586-022-04411-y>.
- [80] World Health Organization. Tracking SARS-CoV-2 variants n.d. <https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/> (accessed June 20, 2022).
- [81] Ma W, Yang J, Fu H, Su C, Yu C, Wang Q, et al. Genomic perspectives on the emerging SARS-CoV-2 omicron variant. *Genomics Proteomics Bioinformatics* 2022. <https://doi.org/10.1016/J.GPB.2022.01.001>.
- [82] McLean G, Kamil J, Lee B, Moore P, Schulz TF, Muik A, et al. The Impact of Evolving SARS-CoV-2 Mutations and Variants on COVID-19 Vaccines. *MBio* 2022;13. <https://doi.org/10.1128/mbio.02979-21>.
- [83] Hui KPY, Ho JCW, Cheung M chun, Ng K chun, Ching RHH, Lai K ling, et al. SARS-CoV-2 Omicron variant replication in human bronchus and lung ex vivo. *Nat* 2022 6037902 2022;603:715–20. <https://doi.org/10.1038/s41586-022-04479-6>.
- [84] Fan Y, Li X, Zhang L, Wan S, Zhang L, Zhou F. SARS-CoV-2 Omicron variant: recent progress and future perspectives. *Signal Transduct Target Ther* 2022;7:1–11. <https://doi.org/10.1038/s41392-022-00997-x>.
- [85] Hachmann NP, Miller J, Collier AY, Ventura JD, Yu J, Rowe M, et al. Neutralization Escape by SARS-CoV-2 Omicron Subvariants BA.2.12.1, BA.4, and BA.5. <https://doi.org/10.1056/NEJMc2206576> 2022.

- <https://doi.org/10.1056/NEJMC2206576>.
- [86] Cao Y, Yisimayi A, Jian F, Song W, Xiao T, Wang L, et al. BA.2.12.1, BA.4 and BA.5 escape antibodies elicited by Omicron infection. *Nat* 2022 2022:1–3.
<https://doi.org/10.1038/s41586-022-04980-y>.
- [87] Otto SP, Day T, Arino J, Colijn C, Dushoff J, Li M, et al. The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic. *Curr Biol* 2021;31:R918–29. <https://doi.org/10.1016/J.CUB.2021.06.049>.
- [88] McCarthy KR, Rennick LJ, Nambulli S, Robinson-McCarthy LR, Bain WG, Haidar G, et al. Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. *Science* (80-) 2021;371:1139–42.
https://doi.org/10.1126/SCIENCE.ABF6950/SUPPL_FILE/ABF6950_REPRODUCIBILITY-CHECKLIST.PDF.
- [89] Choi B, Choudhary MC, Regan J, Sparks JA, Padera RF, Qiu X, et al. Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. *N Engl J Med* 2020;383:2291–3.
https://doi.org/10.1056/NEJMC2031364/SUPPL_FILE/NEJMC2031364_DISCLOSURES.PDF.
- [90] Avanzato VA, Matson MJ, Seifert SN, Pryce R, Williamson BN, Anzick SL, et al. Case Study: Prolonged Infectious SARS-CoV-2 Shedding from an Asymptomatic Immunocompromised Individual with Cancer. *Cell* 2020;183:1901–1912.e9.
<https://doi.org/10.1016/j.cell.2020.10.049>.
- [91] Kemp SA, Collier DA, Datir RP, Ferreira IATM, Gayed S, Jahun A, et al. SARS-CoV-2 evolution during treatment of chronic infection. *Nat* 2021 5927853 2021;592:277–82.
<https://doi.org/10.1038/s41586-021-03291-y>.
- [92] Lacek KA, Rambo-Martin BL, Batra D, Zheng X, Hassell N, Sakaguchi H, et al. SARS-CoV-2 Delta–Omicron Recombinant Viruses, United States. *Emerg Infect Dis* 2022;28:1442. <https://doi.org/10.3201/EID2807.220526>.
- [93] Lindh E, Smura T, Blomqvist S, Liitsola K, Vauhkonen H, Savolainen L, et al. Genomic and epidemiological report of the recombinant XJ lineage SARS-CoV-2 variant, detected in northern Finland, January 2022. *Eurosurveillance* 2022;27:2200257. <https://doi.org/10.2807/1560->

7917.ES.2022.27.16.2200257/CITE/PLAINTEXT.

- [94] Ou J, Lan W, Wu X, Zhao T, Duan B, Yang P, et al. Tracking SARS-CoV-2 Omicron diverse spike gene mutations identifies multiple inter-variant recombination events. *Signal Transduct Target Ther* 2022 71 2022;7:1–9. <https://doi.org/10.1038/s41392-022-00992-2>.
- [95] Jackson B, Boni MF, Bull MJ, Collieran A, Colquhoun RM, Darby AC, et al. Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. *Cell* 2021;184:5179–5188.e8. <https://doi.org/10.1016/J.CELL.2021.08.014>.
- [96] VanInsberghe D, Neish AS, Lowen AC, Koelle K. Recombinant SARS-CoV-2 genomes circulated at low levels over the first year of the pandemic. *Virus Evol* 2021;7. <https://doi.org/10.1093/ve/veab059>.
- [97] Telenti A, Arvin A, Corey L, Corti D, Diamond MS, García-Sastre A, et al. After the pandemic: perspectives on the future trajectory of COVID-19. *Nat* 2021 5967873 2021;596:495–504. <https://doi.org/10.1038/s41586-021-03792-w>.
- [98] Munnink BBO, Sikkema RS, Nieuwenhuijse DF, Molenaar RJ, Munger E, Molenkamp R, et al. Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. *Science (80-)* 2021;371:172–7. https://doi.org/10.1126/SCIENCE.ABE5901/SUPPL_FILE/ABE5901_OUDE_MUNNINK_TABLE_S1.PDF.
- [99] Larsen HD, Fonager J, Lomholt FK, Dalby T, Benedetti G, Kristensen B, et al. Preliminary report of an outbreak of SARS-CoV-2 in mink and mink farmers associated with community spread, Denmark, June to November 2020. *Eurosurveillance* 2021;26:2100009. <https://doi.org/10.2807/1560-7917.ES.2021.26.5.210009/CITE/PLAINTEXT>.
- [100] Khandia R, Singhal S, Alqahtani T, Kamal MA, El-Shall NA, Nainu F, et al. Emergence of SARS-CoV-2 Omicron (B.1.1.529) variant, salient features, high global health concerns and strategies to counter it amid ongoing COVID-19 pandemic. *Environ Res* 2022;209:112816. <https://doi.org/10.1016/J.ENVRES.2022.112816>.
- [101] Oude Munnink BB, Worp N, Nieuwenhuijse DF, Sikkema RS, Haagmans B, Fouchier RAM, et al. The next phase of SARS-CoV-2 surveillance: real-time molecular epidemiology. *Nat Med* 2021 279 2021;27:1518–24. <https://doi.org/10.1038/s41591->

021-01472-w.

- [102] Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* 2006;439:344–8. <https://doi.org/10.1038/nature04388>.
- [103] Parameswaran P, Wang C, Trivedi SB, Eswarappa M, Montoya M, Balmaseda A, et al. Intra-host Selection Pressures Drive Rapid Dengue Virus Microevolution in Acute Human Infections. *Cell Host Microbe* 2017;22:400–410.e5. <https://doi.org/10.1016/J.CHOM.2017.08.003>.
- [104] Henn MR, Boutwell CL, Charlebois P, Lennon NJ, Power KA, Macalalad AR, et al. Whole Genome Deep Sequencing of HIV-1 Reveals the Impact of Early Minor Variants Upon Immune Recognition During Acute Infection. *PLOS Pathog* 2012;8:e1002529. <https://doi.org/10.1371/JOURNAL.PPAT.1002529>.
- [105] Van Doorslaer K. Evolution of the Papillomaviridae. *Virology* 2013;445:11–20. <https://doi.org/10.1016/j.virol.2013.05.012>.
- [106] Cullen M, Boland JF, Schiffman M, Zhang X, Wentzensen N, Yang Q, et al. Deep sequencing of HPV16 genomes: A new high-throughput tool for exploring the carcinogenicity and natural history of HPV16 infection. *Papillomavirus Res* 2015;1:3–11. <https://doi.org/10.1016/J.PVR.2015.05.004>.
- [107] Oliveira CM de, Bravo IG, Souza NCS e., Genta MLND, Fregnani JHTG, Tacla M, et al. High-level of viral genomic diversity in cervical cancers: A Brazilian study on human papillomavirus type 16. *Infect Genet Evol* 2015;34:44–51. <https://doi.org/10.1016/j.meegid.2015.07.002>.
- [108] Sekizuka MT, Ogasawara T, Kondo Y. Genetic Variation of Human Papillomavirus Type 16 in Individual Clinical Specimens Revealed by Deep Sequencing. *PLoS One* 2013;8:80583. <https://doi.org/10.1371/journal.pone.0080583>.
- [109] Dube Mandishora RS, Gjøtterud KS, Lagström S, Stray-Pedersen B, Duri K, Chin'ombe N, et al. Intra-host sequence variability in human papillomavirus. *Papillomavirus Res* 2018;5:180–91. <https://doi.org/10.1016/j.pvr.2018.04.006>.
- [110] Lagström S, van der Weele P, Rounge TB, Christiansen IK, King AJ, Ambur OH. HPV16 whole genome minority variants in persistent infections from young Dutch

- women. *J Clin Virol* 2019. <https://doi.org/10.1016/J.JCV.2019.08.003>.
- [111] Hirose Y, Onuki M, Tenjimbayashi Y, Mori S, Ishii Y, Takeuchi T, et al. Within-Host Variations of Human Papillomavirus Reveal APOBEC Signature Mutagenesis in the Viral Genome. *J Virol* 2018;92:e00017-18. <https://doi.org/10.1128/jvi.00017-18>.
- [112] Mirabello L, Yeager M, Yu K, Clifford GM, Xiao Y, Zhu B, et al. HPV16 E7 Genetic Conservation Is Critical to Carcinogenesis. *Cell* 2017;170:1164–1174.e6. <https://doi.org/10.1016/j.cell.2017.08.001>.
- [113] Zhu B, Xiao Y, Yeager M, Clifford G, Wentzensen N, Cullen M, et al. Mutations in the HPV16 genome induced by APOBEC3 are associated with viral clearance. *Nat Commun* 2020;11:886. <https://doi.org/10.1038/s41467-020-14730-1>.
- [114] 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 Suppl 5. <https://doi.org/10.1016/J.VACCINE.2012.06.083>.
- [115] Spriggs CC, Laimins LA. Human Papillomavirus and the DNA Damage Response: Exploiting Host Repair Pathways for Viral Replication 2017. <https://doi.org/10.3390/v9080232>.
- [116] Kang S Do, Chatterjee S, Alam S, Salzberg AC, Milici J, van der Burg SH, et al. Effect of Productive Human Papillomavirus 16 Infection on Global Gene Expression in Cervical Epithelium. *J Virol* 2018;92. <https://doi.org/10.1128/JVI.01261-18>.
- [117] Jary A, Leducq V, Malet I, Marot S, Klement-Frutos E, Teyssou E, et al. Evolution of viral quasispecies during SARS-CoV-2 infection. *Clin Microbiol Infect* 2020;26:1560.e1-1560.e4. <https://doi.org/10.1016/J.CMI.2020.07.032>.
- [118] Wang D, Wang Y, Sun W, Zhang L, Ji J, Zhang Z, et al. Population Bottlenecks and Intra-host Evolution During Human-to-Human Transmission of SARS-CoV-2. *Front Med* 2021;8:47. <https://doi.org/10.3389/FMED.2021.585358/BIBTEX>.
- [119] Armero A, Berthet N, Avarre J-C, Heraud J-M, Lavergne A, Njouom R. Intra-Host Diversity of SARS-Cov-2 Should Not Be Neglected: Case of the State of Victoria, Australia. *Viruses* 2021, Vol 13, Page 133 2021;13:133. <https://doi.org/10.3390/V13010133>.
- [120] Li J, Du P, Yang L, Zhang J, Song C, Chen D, et al. Two-step fitness selection for

- intra-host variations in SARS-CoV-2. *Cell Rep* 2022;38:110205.
<https://doi.org/10.1016/J.CELREP.2021.110205>.
- [121] van Kampen JJA, van de Vijver DAMC, Fraaij PLA, Haagmans BL, Lamers MM, Okba N, et al. Duration and key determinants of infectious virus shedding in hospitalized patients with coronavirus disease-2019 (COVID-19). *Nat Commun* 2021 121 2021;12:1–6. <https://doi.org/10.1038/s41467-020-20568-4>.
- [122] Ma MJ, Qiu SF, Cui XM, Ni M, Liu HJ, Ye RZ, et al. Persistent SARS-CoV-2 infection in asymptomatic young adults. *Signal Transduct Target Ther* 2022 71 2022;7:1–4. <https://doi.org/10.1038/s41392-022-00931-1>.
- [123] Wang Y, Wang D, Zhang L, Sun W, Zhang Z, Chen W, et al. Intra-host variation and evolutionary dynamics of SARS-CoV-2 populations in COVID-19 patients. *Genome Med* 2021;13:1–13. <https://doi.org/10.1186/S13073-021-00847-5/FIGURES/4>.
- [124] Li B, Deng A, Li K, Hu Y, Li Z, Shi Y, et al. Viral infection and transmission in a large, well-traced outbreak caused by the SARS-CoV-2 Delta variant. *Nat Commun* 2022 131 2022;13:1–9. <https://doi.org/10.1038/s41467-022-28089-y>.
- [125] Ratcliff J, Simmonds P. Potential APOBEC-mediated RNA editing of the genomes of SARS-CoV-2 and other coronaviruses and its impact on their longer term evolution. *Virology* 2021;556:62–72. <https://doi.org/10.1016/J.VIROL.2020.12.018>.
- [126] Warren CJ, Xu T, Guo K, Griffin LM, Westrich JA, Lee D, et al. APOBEC3A Functions as a Restriction Factor of Human Papillomavirus. *J Virol* 2015;89:688–702. <https://doi.org/10.1128/jvi.02383-14>.
- [127] Stenglein MD, Burns MB, Li M, Lengyel J, Harris RS. APOBEC3 proteins mediate the clearance of foreign DNA from human cells. *Nat Struct Mol Biol* 2010;17:222–9. <https://doi.org/10.1038/nsmb.1744>.
- [128] Warren CJ, Westrich JA, Van Doorslaer K, Pyeon D. Roles of APOBEC3A and APOBEC3B in human papillomavirus infection and disease progression. *Viruses* 2017;9. <https://doi.org/10.3390/v9080233>.
- [129] Stavrou S, Ross SR. APOBEC3 Proteins in Viral Immunity. *J Immunol* 2015;195:4565–70. <https://doi.org/10.4049/jimmunol.1501504>.
- [130] Áine O’Toole, Andrew Rambaut. Initial observations about putative APOBEC3

- deaminase editing driving short-term evolution of MPXV since 2017. *Virological* n.d. <https://virological.org/t/initial-observations-about-putative-apobec3-deaminase-editing-driving-short-term-evolution-of-mpxv-since-2017/830> (accessed June 23, 2022).
- [131] Pecori R, Di Giorgio S, Paulo Lorenzo J, Nina Papavasiliou F. Functions and consequences of AID/APOBEC-mediated DNA and RNA deamination. *Nat Rev Genet* 2022;2022:1–14. <https://doi.org/10.1038/s41576-022-00459-8>.
- [132] Suspène R, Aynaud M-M, Koch S, Padeloup D, Labetoulle M, Gaertner B, et al. Genetic Editing of Herpes Simplex Virus 1 and Epstein-Barr Herpesvirus Genomes by Human APOBEC3 Cytidine Deaminases in Culture and In Vivo . *J Virol* 2011;85:7594–602. https://doi.org/10.1128/JVI.00290-11/SUPPL_FILE/SUPFIG1_.ZIP.
- [133] Kanu N, Cerone MA, Goh G, Zalmas LP, Bartkova J, Dietzen M, et al. DNA replication stress mediates APOBEC3 family mutagenesis in breast cancer. *Genome Biol* 2016;17. <https://doi.org/10.1186/s13059-016-1042-9>.
- [134] Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat Genet* 2013;45:970–6. <https://doi.org/10.1038/ng.2702>.
- [135] Henderson S, Chakravarthy A, Su X, Boshoff C, Fenton TR. APOBEC-mediated cytosine deamination links PIK3CA helical domain mutations to human papillomavirus-driven tumor development. *Cell Rep* 2014;7:1833–41. <https://doi.org/10.1016/J.CELREP.2014.05.012>.
- [136] Doorbar J, Griffin H. Refining our understanding of cervical neoplasia and its cellular origins. *Papillomavirus Res* 2019;7:176. <https://doi.org/10.1016/J.PVR.2019.04.005>.
- [137] Aksoy P, Gottschalk EY, Meneses PI. HPV entry into cells. *Mutat Res - Rev Mutat Res* 2017;772:13–22. <https://doi.org/10.1016/j.mrrev.2016.09.004>.
- [138] Stanley MA. Epithelial Cell Responses to Infection with Human Papillomavirus. *Clin Microbiol Rev* 2012;25:215. <https://doi.org/10.1128/CMR.05028-11>.
- [139] Plummer M, Schiffman M, Castle PE, Maucort-Boulch D, Wheeler CM. A 2-year prospective study of human papillomavirus persistence among women with a

- cytological diagnosis of atypical squamous cells of undetermined significance or low-grade squamous intraepithelial lesion. *J Infect Dis* 2007;195:1582–9.
<https://doi.org/10.1086/516784>.
- [140] Schiffman M, Wentzensen N, Wacholder S, Kinney W, Gage JC, Castle PE. Human papillomavirus testing in the prevention of cervical cancer. *J Natl Cancer Inst* 2011;103:368–83. <https://doi.org/10.1093/JNCI/DJQ562>.
- [141] Tainio K, Athanasiou A, Tikkinen KAO, Aaltonen R, Cárdenas J, Hernández, et al. Clinical course of untreated cervical intraepithelial neoplasia grade 2 under active surveillance: systematic review and meta-analysis. *BMJ* 2018;360.
<https://doi.org/10.1136/BMJ.K499>.
- [142] Motamedi M, Böhmer G, Neumann HH, von Wasielewski R. CIN III lesions and regression: Retrospective analysis of 635 cases. *BMC Infect Dis* 2015;15:1–9.
<https://doi.org/10.1186/S12879-015-1277-1/TABLES/4>.
- [143] Wölfel R, Corman VM, Guggemos W, Seilmaier M, Zange S, Müller MA, et al. Virological assessment of hospitalized patients with COVID-2019. *Nat* 2020 5817809 2020;581:465–9. <https://doi.org/10.1038/s41586-020-2196-x>.
- [144] Lamers MM, Haagmans BL. SARS-CoV-2 pathogenesis. *Nat Rev Microbiol* 2022;20:270–84. <https://doi.org/10.1038/s41579-022-00713-0>.
- [145] Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* 2020;181:271–280.e8.
<https://doi.org/10.1016/J.CELL.2020.02.052>.
- [146] Shang J, Wan Y, Luo C, Ye G, Geng Q, Auerbach A, et al. Cell entry mechanisms of SARS-CoV-2. *Proc Natl Acad Sci U S A* 2020;117.
https://doi.org/10.1073/PNAS.2003138117/SUPPL_FILE/PNAS.2003138117.SD01.XLSX.
- [147] Xia S, Zhu Y, Liu M, Lan Q, Xu W, Wu Y, et al. Fusion mechanism of 2019-nCoV and fusion inhibitors targeting HR1 domain in spike protein. *Cell Mol Immunol* 2020;17:765–7. <https://doi.org/10.1038/s41423-020-0374-2>.
- [148] Malone B, Urakova N, Snijder EJ, Campbell EA. Structures and functions of

- coronavirus replication–transcription complexes and their relevance for SARS-CoV-2 drug design. *Nat Rev Mol Cell Biol* 2021 231 2021;23:21–39.
<https://doi.org/10.1038/s41580-021-00432-z>.
- [149] Alturki SO, Alturki SO, Connors J, Cusimano G, Kutzler MA, Izmirly AM, et al. The 2020 Pandemic: Current SARS-CoV-2 Vaccine Development. *Front Immunol* 2020;11:1880. <https://doi.org/10.3389/FIMMU.2020.01880/BIBTEX>.
- [150] V'kovski P, Kratzel A, Steiner S, Stalder H, Thiel V. Coronavirus biology and replication: implications for SARS-CoV-2. *Nat Rev Microbiol* 2020 193 2020;19:155–70. <https://doi.org/10.1038/s41579-020-00468-6>.
- [151] Pizzato M, Baraldi C, Boscato Sopotto G, Finozzi D, Gentile C, Gentile MD, et al. SARS-CoV-2 and the Host Cell: A Tale of Interactions. *Front Virol* 2022;0:46. <https://doi.org/10.3389/FVIRO.2021.815388>.
- [152] Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus–Infected Pneumonia in Wuhan, China. *JAMA* 2020;323:1061–9. <https://doi.org/10.1001/JAMA.2020.1585>.
- [153] Guan W, Ni Z, Hu Y, Liang W, Ou C, He J, et al. Clinical Characteristics of Coronavirus Disease 2019 in China. *N Engl J Med* 2020;382:1708–20. https://doi.org/10.1056/NEJMOA2002032/SUPPL_FILE/NEJMOA2002032_DISCLOSURES.PDF.
- [154] Berlin DA, Gulick RM, Martinez FJ. Severe Covid-19. *N Engl J Med* 2020;383:2451–60. https://doi.org/10.1056/NEJMCP2009575/SUPPL_FILE/NEJMCP2009575_DISCLOSURES.PDF.
- [155] Zaim S, Chong JH, Sankaranarayanan V, Harky A. COVID-19 and Multiorgan Response. *Curr Probl Cardiol* 2020;45:100618. <https://doi.org/10.1016/J.CPCARDIOL.2020.100618>.
- [156] Hojyo S, Uchida M, Tanaka K, Hasebe R, Tanaka Y, Murakami M, et al. How COVID-19 induces cytokine storm with high mortality. *Inflamm Regen* 2020;40. <https://doi.org/10.1186/S41232-020-00146-3>.
- [157] Plumb ID, Feldstein LR, Barkley E, Posner AB, Bregman HS, Hagen MB, et al.

- Effectiveness of COVID-19 mRNA Vaccination in Preventing COVID-19–Associated Hospitalization Among Adults with Previous SARS-CoV-2 Infection — United States, June 2021–February 2022. *MMWR Morb Mortal Wkly Rep* 2022;71:549–55. <https://doi.org/10.15585/MMWR.MM7115E2>.
- [158] Nyberg T, Ferguson NM, Nash SG, Webster HH, Flaxman S, Andrews N, et al. Comparative analysis of the risks of hospitalisation and death associated with SARS-CoV-2 omicron (B.1.1.529) and delta (B.1.617.2) variants in England: a cohort study. *Lancet* 2022;399:1303–12. [https://doi.org/10.1016/S0140-6736\(22\)00462-7/ATTACHMENT/F0F039E8-DCA2-4A04-A602-37E21F276C38/MMC1.PDF](https://doi.org/10.1016/S0140-6736(22)00462-7/ATTACHMENT/F0F039E8-DCA2-4A04-A602-37E21F276C38/MMC1.PDF).
- [159] World Health Organization. A clinical case definition of post COVID-19 condition by a Delphi consensus, 6 October 2021 n.d. https://www.who.int/publications/i/item/WHO-2019-nCoV-Post_COVID-19_condition-Clinical_case_definition-2021.1 (accessed July 2, 2022).
- [160] Lehtinen M, Dillner J. Clinical trials of human papillomavirus vaccines and beyond. *Nat Rev Clin Oncol* 2013;10:400–10. <https://doi.org/10.1038/NRCLINONC.2013.84>.
- [161] 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. *N Engl J Med* 2015;372:711–23. https://doi.org/10.1056/NEJMOA1405044/SUPPL_FILE/NEJMOA1405044_DISCLOSURES.PDF.
- [162] Brown DR, Joura EA, Yen GP, Kothari S, Luxembourg A, Saah A, et al. Systematic literature review of cross-protective effect of HPV vaccines based on data from randomized clinical trials and real-world evidence. *Vaccine* 2021;39:2224–36. <https://doi.org/10.1016/J.VACCINE.2020.11.076>.
- [163] Covert C, Ding L, Brown D, Franco EL, Bernstein DI, Kahn JA. Evidence for cross-protection but not type-replacement over the 11 years after human papillomavirus vaccine introduction. <https://doi.org/10.1080/2164551520181564438> 2019;15:1962–9. <https://doi.org/10.1080/21645515.2018.1564438>.
- [164] Tsang SH, Sampson JN, Schussler J, Porras C, Wagner S, Boland J, et al. Durability of Cross-Protection by Different Schedules of the Bivalent HPV Vaccine: The CVT Trial. *JNCI J Natl Cancer Inst* 2020;112:1030–7. <https://doi.org/10.1093/JNCI/DJAA010>.

- [165] World Health Organization. Human papillomavirus vaccines: WHO position paper, May 2017–Recommendations. *Vaccine* 2017;35:5753–5.
<https://doi.org/10.1016/j.vaccine.2017.05.069>.
- [166] Schiller JT, Castellsagué X, Garland SM. A review of clinical trials of human papillomavirus prophylactic vaccines. *Vaccine* 2012;30 Suppl 5.
<https://doi.org/10.1016/J.VACCINE.2012.04.108>.
- [167] Munoz N, Kjaer SK, Sigurdsson K, Iversen OE, Hernandez-Avila M, Wheeler CM, et al. Impact of human papillomavirus (HPV)-6/11/16/18 vaccine on all HPV-associated genital diseases in young women. *J Natl Cancer Inst* 2010;102:325–39.
<https://doi.org/10.1093/JNCI/DJP534>.
- [168] FHI. HPV-vaksine (Humant papillomavirus) - veileder for helsepersonell n.d.
<https://www.fhi.no/nettpub/vaksinasjonsveilederen-for-helsepersonell/vaksiner-mot-de-enkelte-sykdommene/hpv-vaksinasjon-humant-papillomavir/> (accessed June 14, 2022).
- [169] World Health Organization. Major milestone reached as 100 countries have introduced HPV vaccine into national schedule. *World Heal Organ* 2019.
<https://www.who.int/news/item/31-10-2019-major-milestone-reached-as-100-countries-have-introduced-hpv-vaccine-into-national-schedule> (accessed June 14, 2022).
- [170] Gallagher KE, LaMontagne DS, Watson-Jones D. Status of HPV vaccine introduction and barriers to country uptake. *Vaccine* 2018;36:4761–7.
<https://doi.org/10.1016/J.VACCINE.2018.02.003>.
- [171] Solomon D, Davey D, Kurman R, Moriarty A, O'Connor D, Prey M, et al. The 2001 Bethesda System: terminology for reporting results of cervical cytology. *JAMA* 2002;287:2114–9. <https://doi.org/10.1001/JAMA.287.16.2114>.
- [172] Koliopoulos G, Nyaga VN, Santesso N, Bryant A, Martin-Hirsch PPL, Mustafa RA, et al. Cytology versus HPV testing for cervical cancer screening in the general population. *Cochrane Database Syst Rev* 2017;2017.
<https://doi.org/10.1002/14651858.CD008587.pub2>.
- [173] Gage JC, Schiffman M, Katki HA, Castle PE, Fetterman B, Wentzensen N, et al. Reassurance Against Future Risk of Precancer and Cancer Conferred by a Negative Human Papillomavirus Test. *JNCI J Natl Cancer Inst* 2014;106:153.

- <https://doi.org/10.1093/JNCI/DJU153>.
- [174] Ronco G, Dillner J, Elfström KM, Tunesi S, Snijders PJF, Arbyn M, et al. Efficacy of HPV-based screening for prevention of invasive cervical cancer: Follow-up of four European randomised controlled trials. *Lancet* 2014;383:524–32. [https://doi.org/10.1016/S0140-6736\(13\)62218-7](https://doi.org/10.1016/S0140-6736(13)62218-7).
- [175] Krefregisteret. HPV i primærskanning n.d. <https://www.krefregisteret.no/screening/livmorhalsprogrammet/Helsepersonell/screeningstrategi-og-nasjonale-retningslinjer/HPV-i-primarscreening/> (accessed June 14, 2022).
- [176] Krefregisteret. Flytskjema for vurdering av væskebaserte livmorhalsprøver n.d. <https://www.krefregisteret.no/globalassets/masseundersokelsen-mot-livmorhalskreft/flytdiagram/202003-revidert-algoritme-hpv.pdf> (accessed June 14, 2022).
- [177] Buckley C, Butler E, Fox H. Cervical intraepithelial neoplasia. *J Clin Pathol* 1982;35:1–13. <https://doi.org/10.1136/JCP.35.1.1>.
- [178] Zaino RJ. Symposium part I: Adenocarcinoma in situ, glandular dysplasia, and early invasive adenocarcinoma of the uterine cervix. *Int J Gynecol Pathol* 2002;21:314–26. <https://doi.org/10.1097/00004347-200210000-00002>.
- [179] Krefregisteret. Quality assurance manual: Cervical Cancer Screening Programme 2014. <https://www.krefregisteret.no/globalassets/kvalitetsmanual-leasevnnlig-versjon-mai-2014.pdf> (accessed June 14, 2022).
- [180] Khan MJ, Smith-McCune KK. Treatment of Cervical Precancers: Back to Basics. *Obstet Gynecol* 2014;123:1339. <https://doi.org/10.1097/AOG.0000000000000287>.
- [181] Petry KU. Management options for cervical intraepithelial neoplasia. *Best Pract Res Clin Obstet Gynaecol* 2011;25:641–51. <https://doi.org/10.1016/J.BPOBGYN.2011.04.007>.
- [182] Du L, He Y, Zhou Y, Liu S, Zheng BJ, Jiang S. The spike protein of SARS-CoV — a target for vaccine and therapeutic development. *Nat Rev Microbiol* 2009 73 2009;7:226–36. <https://doi.org/10.1038/nrmicro2090>.
- [183] World Health Organisation. COVID-19 Vaccines Advice. *World Heal Organ* 2022;5.

- [184] Norwegian Institute of Public Health. Who Will Get the Coronavirus Vaccine First? - The New York Times. Nor Inst Public Heal 2020. <https://www.fhi.no/en/id/vaccines/coronavirus-immunisation-programme/who-will-get-coronavirus-vaccine-first/> (accessed July 2, 2022).
- [185] Norwegian Institute of Public Health. Coronavirus vaccine - NIPH. Nor Inst Public Heal 2021. <https://www.fhi.no/en/id/vaccines/coronavirus-immunisation-programme/coronavirus-vaccine/> (accessed July 2, 2022).
- [186] Awadasseid A, Wu Y, Tanaka Y, Zhang W. Current advances in the development of sars-cov-2 vaccines. *Int J Biol Sci* 2021;17:8–19. <https://doi.org/10.7150/ijbs.52569>.
- [187] Creech CB, Walker SC, Samuels RJ. SARS-CoV-2 Vaccines. *JAMA - J Am Med Assoc* 2021;325:1318–20. <https://doi.org/10.1001/jama.2021.3199>.
- [188] Golob JL, Lugogo N, Luring AS, Lok AS. SARS-CoV-2 vaccines: a triumph of science and collaboration. *JCI Insight* 2021;6:149187. <https://doi.org/10.1172/JCI.INSIGHT.149187>.
- [189] Coleman CM, Liu Y V., Mu H, Taylor JK, Massare M, Flyer DC, et al. Purified coronavirus spike protein nanoparticles induce coronavirus neutralizing antibodies in mice. *Vaccine* 2014;32:3169–74. <https://doi.org/10.1016/J.VACCINE.2014.04.016>.
- [190] Karim SSA, Karim QA. Omicron SARS-CoV-2 variant: a new chapter in the COVID-19 pandemic. *Lancet* 2021;398:2126–8. [https://doi.org/10.1016/S0140-6736\(21\)02758-6](https://doi.org/10.1016/S0140-6736(21)02758-6).
- [191] Talic S, Shah S, Wild H, Gasevic D, Maharaj A, Ademi Z, et al. Effectiveness of public health measures in reducing the incidence of covid-19, SARS-CoV-2 transmission, and covid-19 mortality: systematic review and meta-analysis. *BMJ* 2021;375. <https://doi.org/10.1136/BMJ-2021-068302>.
- [192] Duchene S, Featherstone L, Freiesleben De Blasio B, Holmes EC, Bohlin J, Pettersson JHO. Assessment of Coronavirus Disease 2019 Intervention Strategies in the Nordic Countries Using Genomic Epidemiology. *Open Forum Infect Dis* 2022;9. <https://doi.org/10.1093/OFID/OFAB665>.
- [193] Li D, Sempowski GD, Saunders KO, Acharya P, Haynes BF. SARS-CoV-2 Neutralizing Antibodies for COVID-19 Prevention and Treatment. *Annu Rev Med*

- 2022;73:1–16. <https://doi.org/10.1146/annurev-med-042420-113838>.
- [194] Bloch EM, Shoham S, Casadevall A, Sachais BS, Shaz B, Winters JL, et al. Deployment of convalescent plasma for the prevention and treatment of COVID-19. *J Clin Invest* 2020;130:2757–65. <https://doi.org/10.1172/JCI138745>.
- [195] Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell* 2011;144:646–74. <https://doi.org/10.1016/j.cell.2011.02.013>.
- [196] Mittal S, Banks L. Molecular mechanisms underlying human papillomavirus E6 and E7 oncoprotein-induced cell transformation. *Mutat Res - Rev Mutat Res* 2017;772:23–35. <https://doi.org/10.1016/j.mrrev.2016.08.001>.
- [197] Aubrey BJ, Kelly GL, Janic A, Herold MJ, Strasser A. How does p53 induce apoptosis and how does this relate to p53-mediated tumour suppression? *Cell Death Differ* 2018 251 2017;25:104–13. <https://doi.org/10.1038/cdd.2017.169>.
- [198] Giacinti C, Giordano A. RB and cell cycle progression. *Oncogene* 2006 2538 2006;25:5220–7. <https://doi.org/10.1038/sj.onc.1209615>.
- [199] McBride AA, Warburton A. The role of integration in oncogenic progression of HPV-associated cancers. *PLoS Pathog* 2017;13:e1006211. <https://doi.org/10.1371/journal.ppat.1006211>.
- [200] 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. *Nat Genet* 2015;47:158–63. <https://doi.org/10.1038/ng.3178>.
- [201] Li W, Tian S, Wang P, Zang Y, Chen X, Yao Y, et al. The characteristics of HPV integration in cervical intraepithelial cells. *J Cancer* 2019;10:2783. <https://doi.org/10.7150/JCA.31450>.
- [202] 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 Res* 2014;24:185–99. <https://doi.org/10.1101/gr.164806.113>.
- [203] Dürst M, Croce CM, Gissmann L, Schwarz E, Huebner K. Papillomavirus sequences integrate near cellular oncogenes in some cervical carcinomas (viral DNA integration/c-myc/genital cancer). vol. 84. 1987.

- [204] Kadaja M, Isok-Paas H, Laos T, Ustav E, Ustav M. Mechanism of Genomic Instability in Cells Infected with the High-Risk Human Papillomaviruses. *PLOS Pathog* 2009;5:e1000397. <https://doi.org/10.1371/JOURNAL.PPAT.1000397>.
- [205] Zhang R, Shen C, Zhao L, Wang J, McCrae M, Chen X, et al. Dysregulation of host cellular genes targeted by human papillomavirus (HPV) integration contributes to HPV-related cervical carcinogenesis. *Int J Cancer* 2016;138:1163–74. <https://doi.org/10.1002/ijc.29872>.
- [206] Warburton A, Markowitz TE, Katz JP, Pipas JM, McBride AA. Recurrent integration of human papillomavirus genomes at transcriptional regulatory hubs. *Npj Genomic Med* 2021 61 2021;6:1–15. <https://doi.org/10.1038/s41525-021-00264-y>.
- [207] 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. <https://doi.org/10.1371/journal.pone.0119566>.
- [208] Kraus I, Driesch C, Vinokurova S, Hovig E, Schneider A, Von Knebel Doeberitz M, et al. The Majority of Viral-Cellular Fusion Transcripts in Cervical Carcinomas Cotranscribe Cellular Sequences of Known or Predicted Genes. *Cancer Res* 2008;68:2514–36. <https://doi.org/10.1158/0008-5472.CAN-07-2776>.
- [209] Leeman JE, Li Y, Bell A, Hussain SS, Majumdar R, Rong-Mullins X, et al. Human papillomavirus 16 promotes microhomology-mediated end-joining. *Proc Natl Acad Sci U S A* 2019;116:21573–9. <https://doi.org/10.1073/PNAS.1906120116/-DCSUPPLEMENTAL>.
- [210] Burk RD, Chen Z, Saller C, Tarvin K, Carvalho AL, Scapulatempo-Neto C, et al. Integrated genomic and molecular characterization of cervical cancer. *Nature* 2017;543:378–84. <https://doi.org/10.1038/nature21386>.
- [211] 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 Res* 2008;68:307–13. <https://doi.org/10.1158/0008-5472.CAN-07-2754>.
- [212] Stoler MH, Wright TL. Don't Forget HPV-45 in Cervical Cancer Screening. *Am J Clin Pathol* 2012;137:161–3. <https://doi.org/10.1309/AJCPYB6C4HIMLZIX>.
- [213] Cheung JLK, Cheung TH, Tang JWT, Chan PKS. Increase of integration events and

- infection loads of human papillomavirus type 52 with lesion severity from low-grade cervical lesion to invasive cancer. *J Clin Microbiol* 2008;46:1356–62.
<https://doi.org/10.1128/JCM.01785-07>.
- [214] Ho CM, Chien TY, Huang SH, Lee BH, Chang SF. Integrated human papillomavirus types 52 and 58 are infrequently found in cervical cancer, and high viral loads predict risk of cervical cancer. *Gynecol Oncol* 2006;102:54–60.
<https://doi.org/10.1016/J.YGYNO.2005.11.035>.
- [215] Marongiu L, Godi A, Parry J V., Beddows S. Human Papillomavirus 16, 18, 31 and 45 viral load, integration and methylation status stratified by cervical disease stage. *BMC Cancer* 2014;14:1–10. <https://doi.org/10.1186/1471-2407-14-384/FIGURES/4>.
- [216] Woodman CBJ, Collins SI, Young LS. The natural history of cervical HPV infection: unresolved issues. *Nat Rev Cancer* 2007 71 2007;7:11–22.
<https://doi.org/10.1038/nrc2050>.
- [217] Jang MK, Shen K, McBride AA. Papillomavirus genomes associate with BRD4 to replicate at fragile sites in the host genome. *PLoS Pathog* 2014;10.
<https://doi.org/10.1371/JOURNAL.PPAT.1004117>.
- [218] Zhang L, Richards A, Inmaculada Barrasa M, Hughes SH, Young RA, Jaenisch R. Reverse-transcribed SARS-CoV-2 RNA can integrate into the genome of cultured human cells and can be expressed in patient-derived tissues. *Proc Natl Acad Sci U S A* 2021;118.
https://doi.org/10.1073/PNAS.2105968118/SUPPL_FILE/PNAS.2105968118.SD04.XLSX.
- [219] Parry R, Gifford RJ, Lytras S, Ray SC, Coin LJM. No evidence of SARS-CoV-2 reverse transcription and integration as the origin of chimeric transcripts in patient tissues. *Proc Natl Acad Sci U S A* 2021;118.
<https://doi.org/10.1073/PNAS.2109066118>.
- [220] Smits N, Rasmussen J, Bodea GO, Amarilla AA, Gerdes P, Sanchez-Luque FJ, et al. No evidence of human genome integration of SARS-CoV-2 found by long-read DNA sequencing. *Cell Rep* 2021;36:109530.
<https://doi.org/10.1016/J.CELREP.2021.109530>.
- [221] Yan B, Chakravorty S, Mirabelli C, Wang L, Trujillo-Ochoa JL, Chauss D, et al. Host-

- Virus Chimeric Events in SARS-CoV-2-Infected Cells Are Infrequent and Artifactual. *J Virol* 2021;95. <https://doi.org/10.1128/JVI.00294-21/ASSET/75A39BD0-82E4-486E-9F39-934870948F2D/ASSETS/IMAGES/LARGE/JVI.00294-21-F0003.JPG>.
- [222] Ziegert C, Wentzensen N, Vinokurova S, Kissel'ov F, Einkenel J, Hoeckel M, et al. A comprehensive analysis of HPV integration loci in anogenital lesions combining transcript and genome-based amplification techniques. *Oncogene* 2003;22:3977–84. <https://doi.org/10.1038/sj.onc.1206629>.
- [223] WHO. Genomic sequencing of SARS-CoV-2: A guide to implementation for maximum impact on public health. 2021.
- [224] GISAID Initiative n.d. <https://www.epicov.org/epi3/frontend#2f544> (accessed June 30, 2022).
- [225] Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* 2017;22:30494. <https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494>.
- [226] Hill V, Ruis C, Bajaj S, Pybus OG, Kraemer MUG. Progress and challenges in virus genomic epidemiology. *Trends Parasitol* 2021;37:1038–49. <https://doi.org/10.1016/J.PT.2021.08.007>.
- [227] Yang J, Li J, Lai S, Ruktanonchai CW, Xing W, Carioli A, et al. Uncovering two phases of early intercontinental COVID-19 transmission dynamics. *J Travel Med* 2020;27. <https://doi.org/10.1093/JTM/TAAA200>.
- [228] Nadeau SA, Vaughan TG, Scire J, Huisman JS, Stadler T. The origin and early spread of SARS-CoV-2 in Europe. *Proc Natl Acad Sci U S A* 2021;118. https://doi.org/10.1073/PNAS.2012008118/SUPPL_FILE/PNAS.2012008118.SAPP.PDF.
- [229] Fountain-Jones NM, Appaw RC, Carver S, Didelot X, Volz E, Charleston M. Emerging phylogenetic structure of the SARS-CoV-2 pandemic. *Virus Evol* 2020;6. <https://doi.org/10.1093/VE/VEAA082>.
- [230] Osnes MN, Alfsnes K, Bråte J, Garcia I, Riis RK, Instefjord KH, et al. The impact of global lineage dynamics, border restrictions, and emergence of the B.1.1.7 lineage on the SARS-CoV-2 epidemic in Norway. *Virus Evol* 2021;7:1–7.

<https://doi.org/10.1093/VE/VEAB086>.

- [231] Candido DS, Claro IM, de Jesus JG, Souza WM, Moreira FRR, Dellicour S, et al. Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science* (80-) 2020;369:1255–60. https://doi.org/10.1126/SCIENCE.ABD2161/SUPPL_FILE/ABD2161-CANDIDO-SM.PDF.
- [232] du Plessis L, McCrone JT, Zarebski AE, Hill V, Ruis C, Gutierrez B, et al. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* (80-) 2021;371:708–12. https://doi.org/10.1126/SCIENCE.ABF2946/SUPPL_FILE/DUPLESSIS_SM.PDF.
- [233] Ragonnet-Cronin M, Boyd O, Geidelberg L, Jorgensen D, Nascimento FF, Siveroni I, et al. Genetic evidence for the association between COVID-19 epidemic severity and timing of non-pharmaceutical interventions. *Nat Commun* 2021 121 2021;12:1–7. <https://doi.org/10.1038/s41467-021-22366-y>.
- [234] Duchene S, Featherstone L, De Blasio BF, Holmes EC, Bohlin J, Pettersson JHO. The impact of public health interventions in the Nordic countries during the first year of SARS-CoV-2 transmission and evolution. *Eurosurveillance* 2021;26:2001996. <https://doi.org/10.2807/1560-7917.ES.2021.26.44.2001996/CITE/PLAINTEXT>.
- [235] Miller D, Martin MA, Harel N, Tirosh O, Kustin T, Meir M, et al. Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel. *Nat Commun* 2020 111 2020;11:1–10. <https://doi.org/10.1038/s41467-020-19248-0>.
- [236] Geoghegan JL, Ren X, Storey M, Hadfield J, Jelley L, Jefferies S, et al. Genomic epidemiology reveals transmission patterns and dynamics of SARS-CoV-2 in Aotearoa New Zealand. *Nat Commun* 2020;11. <https://doi.org/10.1038/S41467-020-20235-8>.
- [237] Meredith LW, Hamilton WL, Warne B, Houldcroft CJ, Hosmillo M, Jahun AS, et al. Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *Lancet Infect Dis* 2020;0. [https://doi.org/10.1016/S1473-3099\(20\)30562-4](https://doi.org/10.1016/S1473-3099(20)30562-4).
- [238] Arons MM, Hatfield KM, Reddy SC, Kimball A, James A, Jacobs JR, et al. Presymptomatic SARS-CoV-2 Infections and Transmission in a Skilled Nursing Facility. *N Engl J Med* 2020;382:2081–90. <https://doi.org/10.1056/NEJMoa2008457>.

- [239] Kimball A, Hatfield KM, Arons M, James A, Taylor J, Spicer K, et al. Asymptomatic and Presymptomatic SARS-CoV-2 Infections in Residents of a Long-Term Care Skilled Nursing Facility — King County, Washington, March 2020. *MMWR Morb Mortal Wkly Rep* 2020;69:377–81. <https://doi.org/10.15585/mmwr.mm6913e1>.
- [240] McMichael TM, Currie DW, Clark S, Pogojans S, Kay M, Schwartz NG, et al. Epidemiology of Covid-19 in a Long-Term Care Facility in King County, Washington. *N Engl J Med* 2020;382:2005–11. <https://doi.org/10.1056/nejmoa2005412>.
- [241] Liu M, Cheng SZ, Xu KW, Yang Y, Zhu QT, Zhang H, et al. Use of personal protective equipment against coronavirus disease 2019 by healthcare professionals in Wuhan, China: Cross sectional study. *BMJ* 2020;369. <https://doi.org/10.1136/bmj.m2195>.
- [242] Folgueira MD, Munoz-Ruiperez C, Alonso-Lopez MA, Delgado R. SARS-CoV-2 infection in Health Care Workers in a large public hospital in Madrid, Spain, during March 2020 2020. <https://doi.org/10.1101/2020.04.07.20055723>.
- [243] ECDPC. Risk Assessment: Risk related to the spread of new SARS-CoV-2 variants of concern in the EU/EEA – first update. *Eur Cent Dis Prev Control* 2021. <https://www.ecdc.europa.eu/en/publications-data/covid-19-risk-assessment-spread-new-variants-concern-eueea-first-update> (accessed June 30, 2022).
- [244] Wu SL, Mertens AN, Crider YS, Nguyen A, Pokpongkiat NN, Djajadi S, et al. Substantial underestimation of SARS-CoV-2 infection in the United States. *Nat Commun* 2020 111 2020;11:1–10. <https://doi.org/10.1038/s41467-020-18272-4>.
- [245] Mohanan M, Malani A, Krishnan K, Acharya A. Prevalence of SARS-CoV-2 in Karnataka, India. *JAMA* 2021;325:1001–3. <https://doi.org/10.1001/JAMA.2021.0332>.
- [246] Watson OJ, Barnsley G, Toor J, Hogan AB, Winskill P, Ghani AC. Global impact of the first year of COVID-19 vaccination: a mathematical modelling study. *Lancet Infect Dis* 2022;0:1–10. [https://doi.org/10.1016/S1473-3099\(22\)00320-6](https://doi.org/10.1016/S1473-3099(22)00320-6).
- [247] United Nations Development Program. Global Dashboard for Vaccine Equity - UNDP Data Futures Platform. United Nations Dev Progr 2021. <https://data.undp.org/vaccine-equity/> (accessed July 5, 2022).
- [248] Lawal Y. Africa’s low COVID-19 mortality rate: A paradox? *Int J Infect Dis*

- 2021;102:118–22. <https://doi.org/10.1016/J.IJID.2020.10.038>.
- [249] Cohen C, Kleynhans J, von Gottberg A, McMorrow ML, Wolter N, Bhiman JN, et al. SARS-CoV-2 incidence, transmission, and reinfection in a rural and an urban setting: results of the PHIRST-C cohort study, South Africa, 2020–21. *Lancet Infect Dis* 2022;22:821–34. [https://doi.org/10.1016/S1473-3099\(22\)00069-X](https://doi.org/10.1016/S1473-3099(22)00069-X).
- [250] Kleynhans J, Tempia S, Wolter N, von Gottberg A, Bhiman JN, Buys A, et al. SARS-CoV-2 Seroprevalence in a Rural and Urban Household Cohort during First and Second Waves of Infections, South Africa, July 2020–March 2021 - Volume 27, Number 12—December 2021 - *Emerging Infectious Diseases journal* - CDC. *Emerg Infect Dis* 2021;27:3020–9. <https://doi.org/10.3201/EID2712.211465>.
- [251] Gill CJ, Mwananyanda L, MacLeod W, Kwenda G, Pieciak R, Etter L, et al. Sustained high prevalence of COVID-19 deaths from a systematic post-mortem study in Lusaka, Zambia: one year later. *MedRxiv* 2022:2022.03.08.22272087. <https://doi.org/10.1101/2022.03.08.22272087>.
- [252] World Health Organization. Global strategy to accelerate the elimination of cervical cancer as a public health problem and its associated goals and targets for the period 2020 – 2030. vol. 2. 2021.
- [253] Amponsah-Dacosta E, Blose N, Nkwinka VV, Chepkurui V. Human Papillomavirus Vaccination in South Africa: Programmatic Challenges and Opportunities for Integration With Other Adolescent Health Services? *Front Public Heal* 2022;10:59. <https://doi.org/10.3389/FPUBH.2022.799984/BIBTEX>.
- [254] Mbulawa ZZA, Phohlo K, Garcia-Jardon M, Williamson AL, Businge CB. High human papillomavirus (HPV)-35 prevalence among South African women with cervical intraepithelial neoplasia warrants attention. *PLoS One* 2022;17:e0264498. <https://doi.org/10.1371/JOURNAL.PONE.0264498>.
- [255] Tropé A, Sjøborg K, Eskild A, Cuschieri K, Eriksen T, Thoresen S, et al. Performance of human papillomavirus DNA and mRNA testing strategies for women with and without cervical neoplasia. *J Clin Microbiol* 2009;47:2458–64. <https://doi.org/10.1128/JCM.01863-08>.
- [256] Tropé A, Sjøborg KD, Nygård M, Røysland K, Campbell S, Alfsen GC, et al. Cytology and human papillomavirus testing 6 to 12 months after ASCUS or LSIL cytology in



- organized screening to predict high-grade cervical neoplasia between screening rounds. *J Clin Microbiol* 2012;50:1927–35. <https://doi.org/10.1128/JCM.00265-12>.
- [257] Corman VM, Landt O, Kaiser M, Molenkamp R, Meijer A, Chu DKW, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Eurosurveillance* 2020;25. <https://doi.org/10.2807/1560-7917.ES.2020.25.3.2000045>.
- [258] 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. *Sci Rep* 2019;9:524. <https://doi.org/10.1038/s41598-018-36669-6>.
- [259] Lagström S. Characterisation of human papillomavirus genomic variation and chromosomal integration in cervical samples by. Dr Thesis 2020.
- [260] Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat Protoc* 2017;12:1261–6. <https://doi.org/10.1038/nprot.2017.066>.
- [261] Köster J, Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, et al. Sustainable data analysis with Snakemake. *F1000Research* 2021;10:33. <https://doi.org/10.12688/f1000research.29032.2>.
- [262] Hodcroft EB, Hadfield J, Neher RA, Bedford T. Year-letter genetic clade naming for SARS-CoV-2 on nextstrain.org. *Nextstrain* 2020:2020.
- [263] COG-UK n.d. <https://pangolin.cog-uk.io/> (accessed November 4, 2020).
- [264] Hothorn T, Bretz F, Westfall P. Simultaneous inference in general parametric models. *Biom J* 2008;50:346–63. <https://doi.org/10.1002/BIMJ.200810425>.
- [265] Anchordoquy TJ, Molina MC. Preservation of DNA. *Cell Preserv Technol* 2007;5:180–8. <https://doi.org/10.1089/cpt.2007.0511>.
- [266] Teka B, Gizaw M, Firdawoke E, Addissie A, Sisay TA, Schreckenberger C, et al. A Technical Comparison of Human Papillomavirus Genotyping Assays from a Population-Based Cervical Cancer Screening in South Central Ethiopia. *Cancer Manag Res* 2022;14:2253–63. <https://doi.org/10.2147/CMAR.S360712>.

- [267] Lamble S, Batty E, Attar M, Buck D, Bowden R, Lunter G, et al. Improved workflows for high throughput library preparation using the transposome-based nextera system. *BMC Biotechnol* 2013;13:1–10. <https://doi.org/10.1186/1472-6750-13-104/FIGURES/4>.
- [268] Lu X, Wang T, Zhang Y, Liu Y. Analysis of influencing factors of viral load in patients with high-risk human papillomavirus. *Virology* 2021;18:1–8. <https://doi.org/10.1186/S12985-020-01474-Z/TABLES/4>.
- [269] Kebschull JM, Zador AM. Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res* 2015;43. <https://doi.org/10.1093/NAR/GKV717>.
- [270] Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing and measuring bias in sequence data. *Genome Biol* 2013;14:1–20. <https://doi.org/10.1186/GB-2013-14-5-R51/FIGURES/6>.
- [271] Eckert KA, Kunkel TA. DNA polymerase fidelity and the polymerase chain reaction. *PCR Methods Appl* 1991;1:17–24. <https://doi.org/10.1101/GR.1.1.17>.
- [272] Levy SE, Myers RM. Advancements in Next-Generation Sequencing. *Annu Rev Genomics Hum Genet* 2016;17:95–115. <https://doi.org/10.1146/ANNUREV-GENOM-083115-022413>.
- [273] Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016;17:333–51. <https://doi.org/10.1038/nrg.2016.49>.
- [274] Bull RA, Adikari TN, Ferguson JM, Hammond JM, Stevanovski I, Beukers AG, et al. Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. *Nat Commun* 2020;11:1–8. <https://doi.org/10.1038/s41467-020-20075-6>.
- [275] Petrackova A, Vasinek M, Sedlarikova L, Dyskova T, Schneiderova P, Novosad T, et al. Standardization of Sequencing Coverage Depth in NGS: Recommendation for Detection of Clonal and Subclonal Mutations in Cancer Diagnostics. *Front Oncol* 2019;9:851. <https://doi.org/10.3389/fonc.2019.00851>.
- [276] Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW, Schupp JM, et al. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Front*

- Genet 2015;6. <https://doi.org/10.3389/FGENE.2015.00235>.
- [277] Jennings LJ, Arcila ME, Corless C, Kamel-Reid S, Lubin IM, Pfeifer J, et al. Guidelines for Validation of Next-Generation Sequencing-Based Oncology Panels: A Joint Consensus Recommendation of the Association for Molecular Pathology and College of American Pathologists. *J Mol Diagn* 2017;19:341–65. <https://doi.org/10.1016/J.JMOLDX.2017.01.011>.
- [278] Peter M, Stransky N, Couturier J, Hupé P, Barillot E, De Cremoux P, et al. Frequent genomic structural alterations at HPV insertion sites in cervical carcinoma. *J Pathol* 2010;221:320–30. <https://doi.org/10.1002/path.2713>.
- [279] Rusan M, Li YY, Hammerman PS. Genomic landscape of human papillomavirus-associated cancers. *Clin Cancer Res* 2015;21:2009–19. <https://doi.org/10.1158/1078-0432.CCR-14-1101>.
- [280] McHugh ML. The Chi-square test of independence. *Biochem Medica* 2013;23:143. <https://doi.org/10.11613/BM.2013.018>.
- [281] Guan P, Howell-Jones R, Li N, Bruni L, De Sanjosé S, Franceschi S, et al. Human papillomavirus types in 115,789 HPV-positive women: A meta-analysis from cervical infection to cancer. *Int J Cancer* 2012;131:2349–59. <https://doi.org/10.1002/IJC.27485>.
- [282] Pett M, Coleman N. Integration of high-risk human papillomavirus: A key event in cervical carcinogenesis? *J Pathol* 2007;212:356–67. <https://doi.org/10.1002/path.2192>.
- [283] Jeon S, Allen-Hoffmann BL, Lambert PF. Integration of human papillomavirus type 16 into the human genome correlates with a selective growth advantage of cells. *J Virol* 1995;69:2989–97. <https://doi.org/10.1128/jvi.69.5.2989-2997.1995>.
- [284] Ojesina AI, Lichtenstein L, Freeman SS, Pedamallu CS, Imaz-Rosshandler I, Pugh TJ, et al. Landscape of genomic alterations in cervical carcinomas. *Nature* 2014;506:371–5. <https://doi.org/10.1038/nature12881>.
- [285] Siegel EM, Eschrich S, Winter K, Riggs B, Berglund A, Ajidahun A, et al. Epigenomic Characterization of Locally Advanced Anal Cancer: An RTOG 98-11 Specimen Study. *Dis Colon Rectum* 2014;57:941. <https://doi.org/10.1097/DCR.000000000000160>.
- [286] Niitsu H, Hinoi T, Kawaguchi Y, Sentani K, Yuge R, Kitadai Y, et al. KRAS mutation leads to decreased expression of regulator of calcineurin 2, resulting in tumor

- proliferation in colorectal cancer. *Oncog* 2016 58 2016;5:e253–e253.
<https://doi.org/10.1038/oncis.2016.47>.
- [287] Li Y, Wang H, Huang H. Long non-coding RNA MIR205HG function as a ceRNA to accelerate tumor growth and progression via sponging miR-122–5p in cervical cancer. *Biochem Biophys Res Commun* 2019;514:78–85.
<https://doi.org/10.1016/J.BBRC.2019.04.102>.
- [288] Yin L, Zhang Y, Zheng L. Analysis of differentially expressed long non-coding RNAs revealed a pro-tumor role of MIR205HG in cervical cancer. *Mol Med Rep* 2022;25:1–8. <https://doi.org/10.3892/MMR.2021.12558/HTML>.
- [289] Dong M, Dong Z, Zhu X, Zhang Y, Song L. Long non-coding RNA MIR205HG regulates KRT17 and tumor processes in cervical cancer via interaction with SRSF1. *Exp Mol Pathol* 2019;111. <https://doi.org/10.1016/J.YEXMP.2019.104322>.
- [290] Chakravarthy A, Reddin I, Henderson S, Dong C, Kirkwood N, Jeyakumar M, et al. Integrated analysis of cervical squamous cell carcinoma cohorts from three continents reveals conserved subtypes of prognostic significance. *BioRxiv* 2021:2020.04.02.019711. <https://doi.org/10.1101/2020.04.02.019711>.
- [291] Mesplède T, Gagnon D, Bergeron-Labrecque F, Azar I, Sénéchal H, Coutlée F, et al. p53 Degradation Activity, Expression, and Subcellular Localization of E6 Proteins from 29 Human Papillomavirus Genotypes. *J Virol* 2012;86:94–107.
<https://doi.org/10.1128/jvi.00751-11>.
- [292] Zheng Y, Li X, Jiao Y, Wu C. High-Risk Human Papillomavirus Oncogenic E6/E7 mRNAs Splicing Regulation. *Front Cell Infect Microbiol* 2022;0:790.
<https://doi.org/10.3389/FCIMB.2022.929666>.
- [293] Khan HA, Baig FK, Mehboob R. Nosocomial infections: Epidemiology, prevention, control and surveillance. *Asian Pac J Trop Biomed* 2017;7:478–82.
<https://doi.org/10.1016/J.APJTb.2017.01.019>.
- [294] Abbas M, Robalo Nunes T, Cori A, Cordey S, Laubscher F, Baggio S, et al. Explosive nosocomial outbreak of SARS-CoV-2 in a rehabilitation clinic: the limits of genomics for outbreak reconstruction. *J Hosp Infect* 2021;117:124–34.
<https://doi.org/10.1016/j.jhin.2021.07.013>.

- [295] Snell LB, Fisher CL, Taj U, Stirrup O, Merrick B, Alcolea-Medina A, et al. Combined epidemiological and genomic analysis of nosocomial SARS-CoV-2 infection early in the pandemic and the role of unidentified cases in transmission. *Clin Microbiol Infect* 2022;28:93–100. <https://doi.org/10.1016/J.CMI.2021.07.040>.
- [296] Lumley SF, Constantinides B, Sanderson N, Rodger G, Street TL, Swann J, et al. Epidemiological data and genome sequencing reveals that nosocomial transmission of SARS-CoV-2 is underestimated and mostly mediated by a small number of highly infectious individuals. *J Infect* 2021;83:473–82. <https://doi.org/10.1016/J.JINF.2021.07.034>.
- [297] Stirrup OT, Hughes J, Parker M, Partridge DG, Shepherd JG, Blackstone J, et al. Rapid feedback on hospital onset sars-cov-2 infections combining epidemiological and sequencing data. *Elife* 2021;10. <https://doi.org/10.7554/ELIFE.65828>.
- [298] San JE, Ngcapu S, Kanzi AM, Tegally H, Fonseca V, Giandhari J, et al. Transmission dynamics of SARS-CoV-2 within-host diversity in two major hospital outbreaks in South Africa. *Virus Evol* 2021;7:41. <https://doi.org/10.1093/VE/VEAB041>.
- [299] Sikkema RS, Pas SD, Nieuwenhuijse DF, O’Toole Á, Verweij J, van der Linden A, et al. COVID-19 in health-care workers in three hospitals in the south of the Netherlands: a cross-sectional study. *Lancet Infect Dis* 2020;0. [https://doi.org/10.1016/S1473-3099\(20\)30527-2](https://doi.org/10.1016/S1473-3099(20)30527-2).
- [300] Sender R, Bar-On YM, Gleizer S, Bernshtein B, Flamholz A, Phillips R, et al. The total number and mass of SARS-CoV-2 virions. *Proc Natl Acad Sci U S A* 2021;118. https://doi.org/10.1073/PNAS.2024815118/SUPPL_FILE/PNAS.2024815118.SD01.XLSX.
- [301] Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, et al. SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol* 2021 197 2021;19:409–24. <https://doi.org/10.1038/s41579-021-00573-0>.
- [302] Gutiérrez S, Michalakis Y, Blanc S. Virus population bottlenecks during within-host progression and host-to-host transmission. *Curr Opin Virol* 2012;2:546–55. <https://doi.org/10.1016/j.coviro.2012.08.001>.
- [303] Zhao LH, Liu X, Yan HX, Li WY, Zeng X, Yang Y, et al. Genomic and oncogenic preference of HBV integration in hepatocellular carcinoma. *Nat Commun* 2016 71

- 2016;7:1–10. <https://doi.org/10.1038/ncomms12992>.
- [304] Kenan DJ, Mieczkowski PA, Burger-Calderon R, Singh HK, Nickeleit V. The oncogenic potential of BK-polyomavirus is linked to viral integration into the human genome. *J Pathol* 2015;237:379–89. <https://doi.org/10.1002/PATH.4584>.
- [305] Martinez MA, Franco S. Therapy Implications of Hepatitis C Virus Genetic Diversity. *Viruses* 2021, Vol 13, Page 41 2020;13:41. <https://doi.org/10.3390/V13010041>.
- [306] Ni M, Chen C, Qian J, Xiao HX, Shi WF, Luo Y, et al. Intra-host dynamics of Ebola virus during 2014. *Nat Microbiol* 2016 111 2016;1:1–9. <https://doi.org/10.1038/nmicrobiol.2016.151>.
- [307] IARC. Cervix cancer screening. IARC handbooks of cancer prevention, International Agency for Research on Cancer. World Heal Organ IARC Press 2005.
- [308] Braun KM, 1  ID, Moreno GK, 2  ID, Wagner Id C, Id MAA, et al. Acute SARS-CoV-2 infections harbor limited within-host diversity and transmit via tight transmission bottlenecks 2021. <https://doi.org/10.1371/journal.ppat.1009849>.
- [309] Attwood SW, Hill SC, Aanensen DM, Connor TR, Pybus OG. Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic. *Nat Rev Genet* 2022 2022:1–16. <https://doi.org/10.1038/s41576-022-00483-8>.
- [310] Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis* 2020;20:533–4. [https://doi.org/10.1016/S1473-3099\(20\)30120-1](https://doi.org/10.1016/S1473-3099(20)30120-1).
- [311] Thakur S, Sasi S, Pillai SG, Nag A, Shukla D, Singhal R, et al. SARS-CoV-2 Mutations and Their Impact on Diagnostics, Therapeutics and Vaccines. *Front Med* 2022;9:815389. <https://doi.org/10.3389/FMED.2022.815389/FULL>.

8. PAPERS I-III

Paper I

Lagström S, **Hesselberg Løvestad A**, Umu SU, Ambur OH, Nygård M, Rounge TB, Christiansen IK. HPV16 and HPV18 type-specific APOBEC3 and integration profiles in different diagnostic categories of cervical samples. *Tumour Virus Research* 2021;12:200221.

DOI: <https://doi.org/10.1016/j.tvr.2021.200221>



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

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

^a Department of Microbiology and Infection Control, Akershus University Hospital, Lørenskog, Norway

^b Department of Research, Cancer Registry of Norway, Oslo, Norway

^c Institute of Clinical Medicine, University of Oslo, Oslo, Norway

^d Faculty of Health Sciences, OsloMet, Oslo Metropolitan University, Oslo, Norway

^e Department of Informatics, University of Oslo, Oslo, Norway

^f Department of Clinical Molecular Biology (EpiGen), Division of Medicine, Akershus University Hospital and University of Oslo, Lørenskog, Norway

ARTICLE INFO

Keywords:

Human papillomavirus
Minor nucleotide variation
APOBEC3
Chromosomal integration
Viral genomic deletion

ABSTRACT

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

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

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

1. Introduction

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

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

* Corresponding author. Department of Microbiology and Infection Control, Akershus University Hospital, Lørenskog, Norway.

** Corresponding author. Department of Research, Cancer Registry of Norway, Oslo, Norway.

E-mail addresses: trine.rounge@krefregisteret.no (T.B. Rounge), irene.kraus.christiansen@ahus.no (I.K. Christiansen).

¹ Equal contribution.

Abbreviations

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

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

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

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

2. Material and methods

2.1. Sample selection

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

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

2.2. Library preparation and sequencing

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

2.3. Sequence alignment

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

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

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

^a By cytology.

^b By cytology; no cell abnormalities within 4-year follow-up.

^c Cytology taken at the time of consiation, with the histological diagnosis presented.

^d Includes cases of SCC (n = 4) and adenocarcinoma (n = 1).

^e Non-progressive category, samples combined for analysis.

2.4. Sequence variation analysis

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

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

2.5. Detection of chromosomal integration

Integration site detection was performed as described previously [34]. In brief, a two-step analysis strategy was employed to identify read pairs spanning integration sites. First, read pairs with one read mapped to HPV and the other to the human chromosome were identified using HISAT2. Second, unmapped reads were re-mapped using the LAST (v876) aligner (options -M -C2) [40] to increase detections of the above mentioned read pairs. Reads sharing the same start and end coordinates

were considered as potential PCR duplicates and were excluded. Selected integration sites were confirmed by PCR amplification and Sanger sequencing on the ABI® 3130xl/3100 Genetic Analyzer 16-Capillary Array (Thermo Fisher Scientific Inc., Waltham, MA) using BigDye™ Terminator v1.1 cycle sequencing kit (Thermo Fisher Scientific Inc., Waltham, MA). Samples with a mean depth of $> 1000\times$ and $< 85\%$ of the genome covered by minimum 100× were manually inspected using IGV (v2.3.90) to detect HPV genomic deletions.

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

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

2.7. Statistical analysis

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

2.8. Ethical considerations

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

3. Results

3.1. Characteristics and sequencing statistics

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

3.2. Minor nucleotide variation profiles similar for HPV16 and HPV18

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

3.3. Different level of variation in HPV16 and HPV18 genes

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

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

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

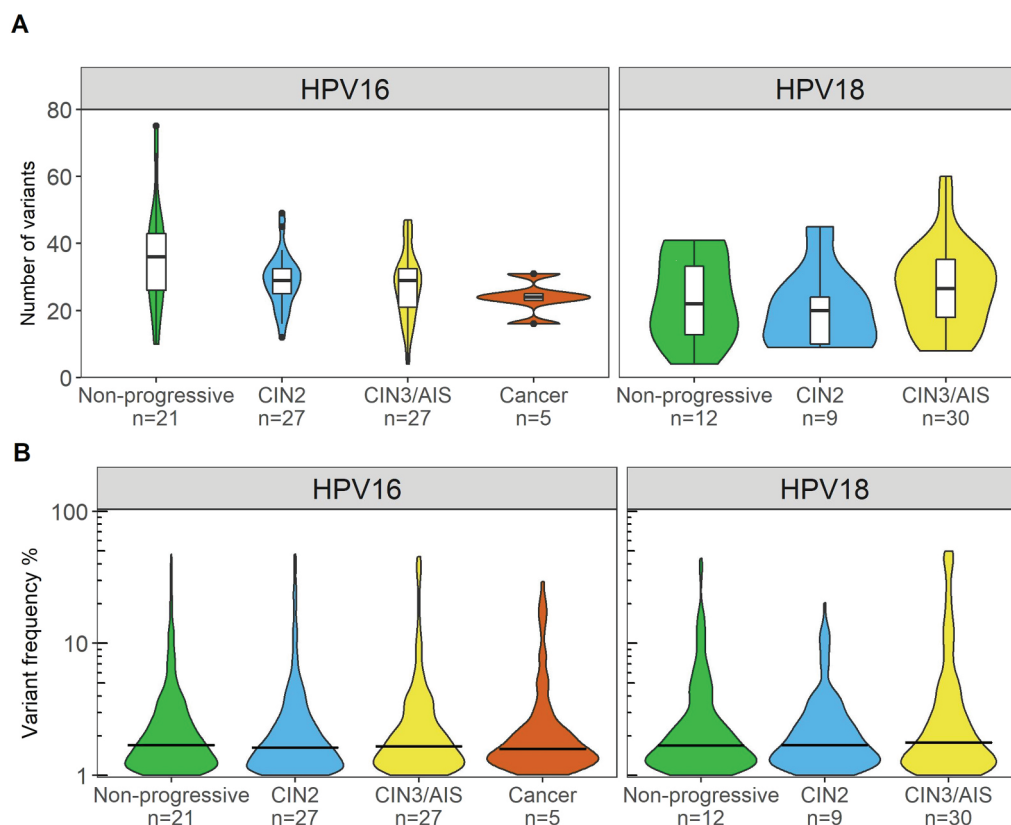


Fig. 1. Number of variants and variant frequencies in HPV16 and HPV18 positive samples. A) Number of variants presented as violin plots across the different diagnostic categories shown on x-axis. Violin plot shows the probability density of the data, using kernel density estimation. Box-and-whisker plots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

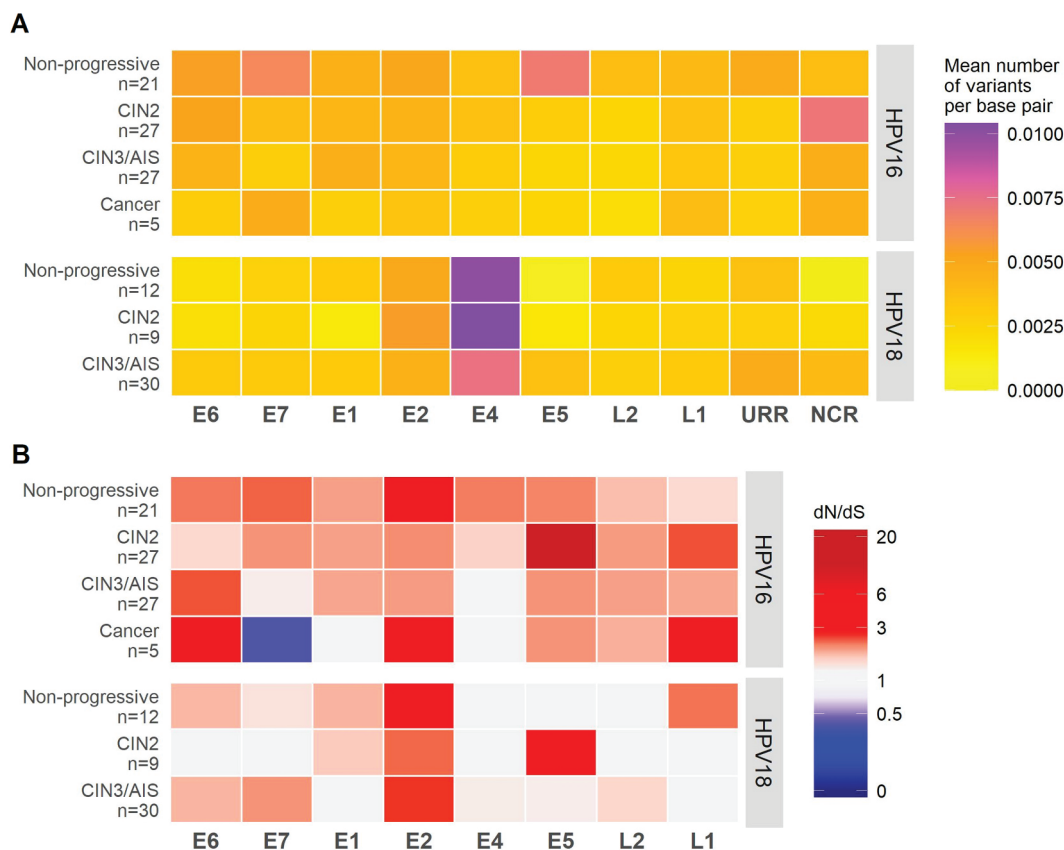


Fig. 2. Number of variants, nonsynonymous and synonymous variations in the different HPV genes. A) Heat map with yellow-orange-purple gradient colour-coding representing mean number of variants per sample in HPV16 and HPV18 genomic regions. Number of variants is normalised by the gene length and stratified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

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

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

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

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

(Supplementary Table S1).

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

3.6. Break points and deletions in the HPV genome

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

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

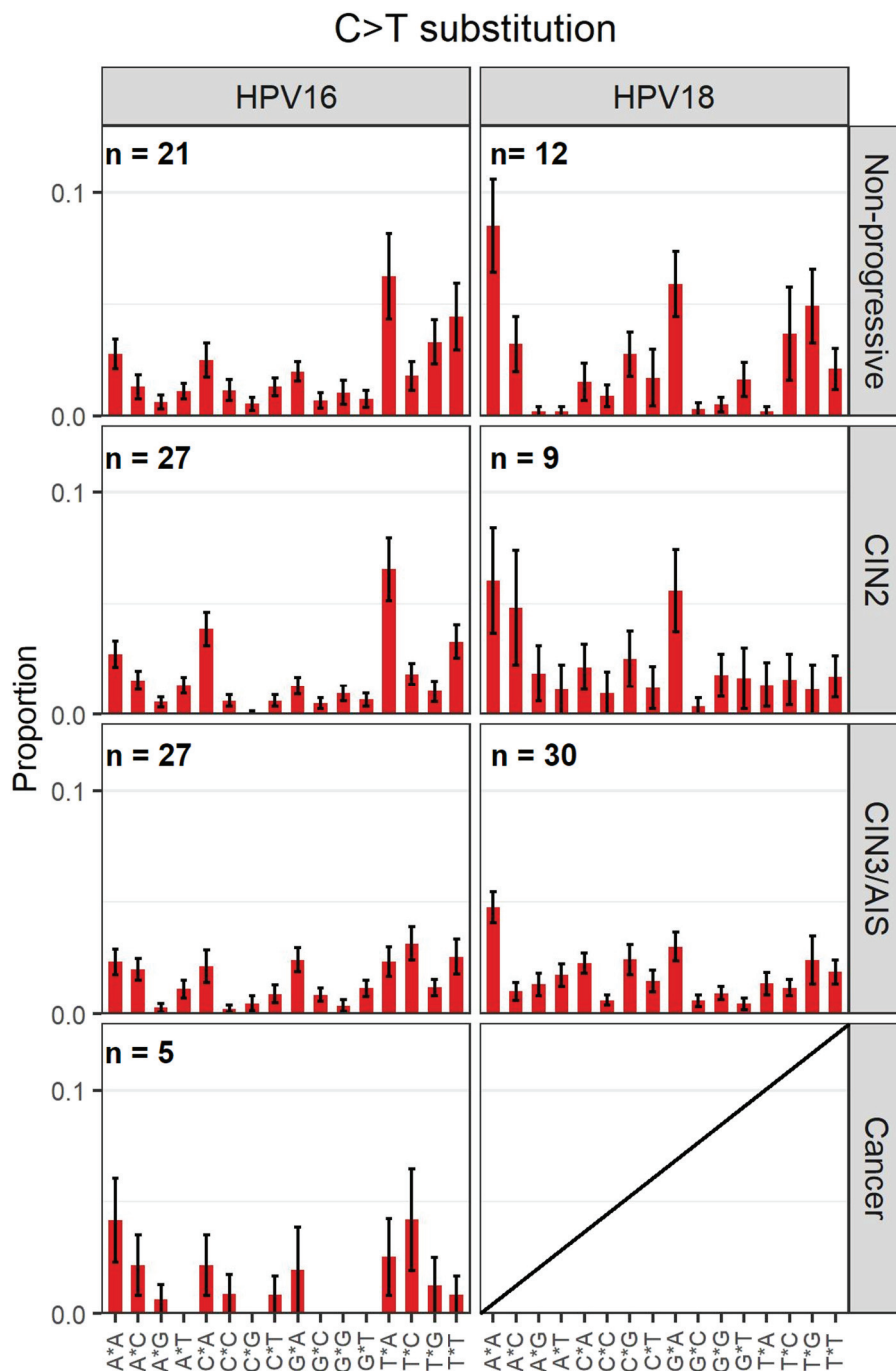


Fig. 3. C > T mutational signatures in HPV16 and HPV18 positive samples. The mean proportion of 16 trinucleotide substitution types is shown below the plots across the different diagnostic categories. Error bars represent the standard error of the mean.

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

3.7. Integration sites in the human genome

In HPV16 positive samples, integration sites (n = 17) were distributed on 10 chromosomes; for the cancer samples, all integration sites (n = 7) were located on chromosomes 1, 8 or 10 (Fig. 4C). Interestingly, the

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

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

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

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

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

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

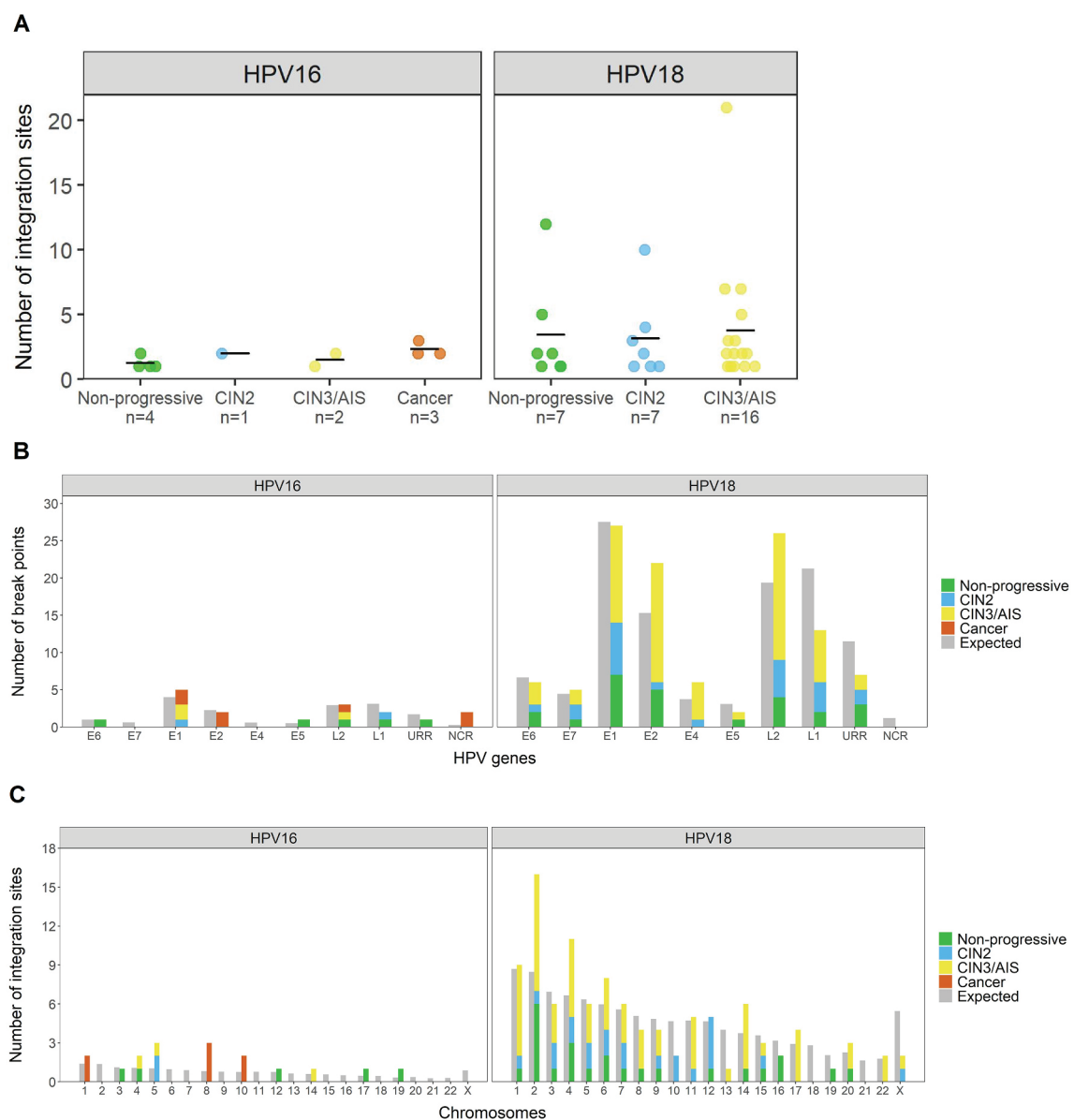


Fig. 4. Chromosomal integration sites and HPV break points in HPV16 and HPV18 positive samples. A) Number of integration sites in samples with observed integration. Each spot in the plot indicates one sample. Total number of samples with integration is specified for each diagnostic category on x-axis.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

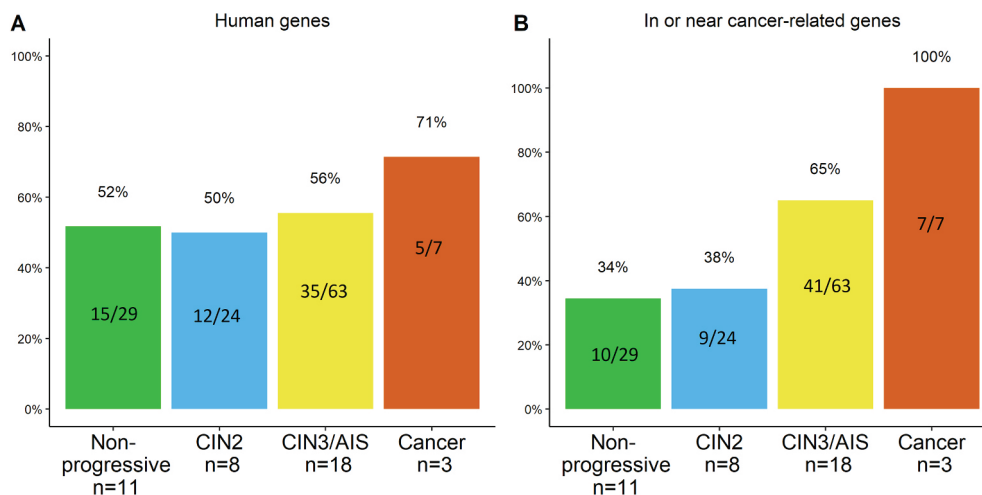


Fig. 5. The frequency of integration sites combined for HPV16 and HPV18 A) in human genes, and B) in or near human cancer-related genes. Number of integration sites is indicated inside the bars and total number of samples with integration (n) for each diagnostic category is specified on x-axis.

4. Discussion

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

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

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

as a novel mechanism to sustain L2 expression, and thereby justify the potential role of L2 in HPV-induced carcinogenesis [52]. However, an opposite finding has also been reported, showing more variation in NCR in clearing than in persistent HPV16 infections [46].

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

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

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

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

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

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

5. Conclusions

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

CRedit authorship contribution statement

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

Acknowledgements

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

Appendix A. Supplementary data

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

Funding

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

Data statement

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

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

Authors' contributions

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

Declaration of competing interest

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

References

- [1] F.X. Bosch, et al., The causal relation between human papillomavirus and cervical cancer, *J. Clin. Pathol.* 55 (2002) 244–265.
- [2] IARC working group on the evaluation of carcinogenic risks to humans, biological agents. Volume 100 B. A review of human carcinogens, IARC Monogr. Eval. Carcinog. Risks Hum. 100 (2012) 1–441.
- [3] S. de Sanjose, et al., Human papillomavirus genotype attribution in invasive cervical cancer: a retrospective cross-sectional worldwide study, *Lancet Oncol.* 11 (2010) 1048–1056, [https://doi.org/10.1016/s1470-2045\(10\)70230-8](https://doi.org/10.1016/s1470-2045(10)70230-8).
- [4] W.A. Tjalma, et al., Differences in human papillomavirus type distribution in high-grade cervical intraepithelial neoplasia and invasive cervical cancer in Europe, *Int. J. Canc.* 132 (2013) 854–867, <https://doi.org/10.1002/ijc.27713>.
- [5] H. zur Hausen, Papillomaviruses and cancer: from basic studies to clinical application, *Nat. Rev. Canc.* 2 (2002) 342–350, <https://doi.org/10.1038/nrc798>.
- [6] H.U. Bernard, Taxonomy and phylogeny of papillomaviruses: an overview and recent developments, *Infect. Genet. Evol.* 18 (2013) 357–361, <https://doi.org/10.1016/j.meegid.2013.03.011>.
- [7] B. Smith, et al., Sequence imputation of HPV16 genomes for genetic association studies, *PLoS One* 6 (2011), e21375, <https://doi.org/10.1371/journal.pone.0021375>.
- [8] D. Bzhalava, C. Eklund, J. Dillner, International standardization and classification of human papillomavirus types, *Virology* 476 (2015) 341–344, <https://doi.org/10.1016/j.virol.2014.12.028>.
- [9] R.D. Burk, A. Harari, Z. Chen, Human papillomavirus genome variants, *Virology* 445 (2013) 232–243, <https://doi.org/10.1016/j.virol.2013.07.018>.
- [10] K. Van Doorslaer, Evolution of the papillomaviridae, *Virology* 445 (2013) 11–20, <https://doi.org/10.1016/j.virol.2013.05.012>.
- [11] Y. Hirose, et al., Within-host variations of human papillomavirus reveal APOBEC-signature mutagenesis in the viral genome, *J. Virol.* (2018), <https://doi.org/10.1128/jvi.00017-18>.
- [12] R.S. Dube Mandishora, et al., Intra-host sequence variability in human papillomavirus, *Papillomavirus Res* (2018), <https://doi.org/10.1016/j.pvr.2018.04.006>.
- [13] S. Lagström, et al., HPV16 whole genome minority variants in persistent infections from young Dutch women, *J. Clin. Virol.* 119 (2019) 24–30, <https://doi.org/10.1016/j.jcv.2019.08.003>.
- [14] E. Domingo, J. Sheldon, C. Perales, Viral quasispecies evolution, *Microbiol. Mol. Biol. Rev.* 76 (2012) 159–216, <https://doi.org/10.1128/MMBR.05023-11>.
- [15] C.J. Warren, et al., APOBEC3A functions as a restriction factor of human papillomavirus, *J. Virol.* 89 (2015) 688–702, <https://doi.org/10.1128/JVI.02383-14>.
- [16] R.S. Harris, J.P. Dudley, APOBECs and virus restriction, *Virology* 479–480 (2015) 131–145, <https://doi.org/10.1016/j.virol.2015.03.012>.
- [17] L.B. Alexandrov, et al., Signatures of mutational processes in human cancer, *Nature* 500 (2013) 415–421, <https://doi.org/10.1038/nature12477>.
- [18] J.P. Vartanian, et al., Evidence for editing of human papillomavirus DNA by APOBEC3 in benign and precancerous lesions, *Science* 320 (2008) 230–233, <https://doi.org/10.1126/science.1153201>.
- [19] A.A. Mariaggi, et al., Presence of human papillomavirus (HPV) apolipoprotein B messenger RNA editing, catalytic polypeptide-like 3 (APOBEC)-Related minority variants in HPV-16 genomes from anal and cervical samples but not in HPV-52 and HPV-58, *J. Infect. Dis.* 218 (2018) 1027–1036, <https://doi.org/10.1093/infdis/jiy287>.
- [20] B. Zhu, et al., Mutations in the HPV16 genome induced by APOBEC3 are associated with viral clearance, *Nat. Commun.* 11 (2020) 886, <https://doi.org/10.1038/s41467-020-14730-1>.
- [21] V.C. Vieira, et al., Human papillomavirus E6 triggers upregulation of the antiviral and cancer genomic DNA deaminase APOBEC3B, *mBio* 5 (2014), <https://doi.org/10.1128/mBio.02234-14>.
- [22] K. Chan, et al., An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers, *Nat. Genet.* 47 (2015) 1067–1072, <https://doi.org/10.1038/ng.3378>.
- [23] L.S. Arroyo-Muhr, et al., Human papillomavirus type 16 genomic variation in women with subsequent in situ or invasive cervical cancer: prospective population-based study, *Br. J. Canc.* 119 (2018) 1163–1168, <https://doi.org/10.1038/s41416-018-0311-7>.
- [24] L. Mirabello, et al., HPV16 E7 genetic conservation is critical to carcinogenesis, *Cell* 170 (2017) 1164–1174, <https://doi.org/10.1016/j.cell.2017.08.001>, e6.
- [25] Cancer Genome Atlas Research Network, Integrated genomic and molecular characterization of cervical cancer, *Nature* 543 (2017) 378–384, <https://doi.org/10.1038/nature21386>.
- [26] J. Doorbar, et al., Human papillomavirus molecular biology and disease association, *Rev. Med. Virol.* 25 (Suppl 1) (2015) 2–23, <https://doi.org/10.1002/rmv.1822>.
- [27] A.A. McBride, A. Warburton, The role of integration in oncogenic progression of HPV-associated cancers, *PLoS Pathog.* 13 (2017), e1006211, <https://doi.org/10.1371/journal.ppat.1006211>.
- [28] K. Akagi, et al., Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability, *Genome Res.* 24 (2014) 185–199, <https://doi.org/10.1101/gr.164806.113>.
- [29] C. Bodelon, et al., Genomic characterization of viral integration sites in HPV-related cancers, *Int. J. Canc.* 139 (2016) 2001–2011, <https://doi.org/10.1002/ijc.30243>.
- [30] M. Peter, et al., Frequent genomic structural alterations at HPV insertion sites in cervical carcinoma, *J. Pathol.* 221 (2010) 320–330, <https://doi.org/10.1002/path.2713>.
- [31] I. Kraus, et al., The majority of viral-cellular fusion transcripts in cervical carcinomas cotranscribe cellular sequences of known or predicted genes, *Canc. Res.* 68 (2008) 2514–2522, <https://doi.org/10.1158/0008-5472.CAN-07-2776>.
- [32] Y. Liu, et al., Genome-wide profiling of the human papillomavirus DNA integration in cervical intraepithelial neoplasia and normal cervical epithelium by HPV capture technology, *Sci. Rep.* 6 (2016) 35427, <https://doi.org/10.1038/srep35427>.
- [33] J. Huang, et al., Comprehensive genomic variation profiling of cervical intraepithelial neoplasia and cervical cancer identifies potential targets for cervical cancer early warning, *J. Med. Genet.* 56 (2019) 186–194, <https://doi.org/10.1136/jmedgenet-2018-105745>.
- [34] S. Lagström, et al., TaME-seq: an efficient sequencing approach for characterisation of HPV genomic variability and chromosomal integration, *Sci. Rep.* 9 (2019) 524, <https://doi.org/10.1038/s41598-018-36699-6>.
- [35] A. Trope, et al., Performance of human papillomavirus DNA and mRNA testing strategies for women with and without cervical neoplasia, *J. Clin. Microbiol.* 47 (2009) 2458–2464, <https://doi.org/10.1128/JCM.01863-08>.
- [36] A. Trope, et al., Cytology and human papillomavirus testing 6 to 12 months after ASCUS or LSIL cytology in organized screening to predict high-grade cervical neoplasia between screening rounds, *J. Clin. Microbiol.* 50 (2012) 1927–1935, <https://doi.org/10.1128/JCM.00265-12>.
- [37] J.J. Kozich, et al., Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform, *Appl. Environ. Microbiol.* 79 (2013) 5112–5120, <https://doi.org/10.1128/AEM.01043-13>.
- [38] D. Kim, B. Langmead, S.L. Salzberg, HISAT: a fast spliced aligner with low memory requirements, *Nat. Methods* 12 (2015) 357–360, <https://doi.org/10.1038/nmeth.3317>.
- [39] H. Li, et al., The sequence alignment/map format and SAMtools, *Bioinformatics* 25 (2009) 2078–2079, <https://doi.org/10.1093/bioinformatics/btp352>.
- [40] S.M. Kielbasa, et al., Adaptive seeds tame genomic sequence comparison, *Genome Res.* 21 (2011) 487–493, <https://doi.org/10.1101/gr.113985.110>.
- [41] D. Welter, et al., The NHGRI GWAS Catalog, a curated resource of SNP-trait associations, *Nucleic Acids Res.* 42 (2014) D1001–D1006, <https://doi.org/10.1093/nar/gkt1229>.
- [42] D.R. Zerbino, et al., The ensembl regulatory build, *Genome Biol.* 16 (2015) 56, <https://doi.org/10.1186/s13059-015-0621-5>.
- [43] S.A. Roberts, et al., An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers, *Nat. Genet.* 45 (2013) 970–976, <https://doi.org/10.1038/ng.2702>.
- [44] M.B. Burns, et al., APOBEC3B is an enzymatic source of mutation in breast cancer, *Nature* 494 (2013) 366–370, <https://doi.org/10.1038/nature11881>.
- [45] S. Vinokurova, et al., Type-dependent integration frequency of human papillomavirus genomes in cervical lesions, *Canc. Res.* 68 (2008) 307–313, <https://doi.org/10.1158/0008-5472.CAN-07-2754>.
- [46] P. van der Wee, C. Meijer, A.J. King, Whole-genome sequencing and variant analysis of human papillomavirus 16 infections, *J. Virol.* 91 (2017), <https://doi.org/10.1128/jvi.00844-17>.
- [47] P. van der Wee, C. Meijer, A.J. King, High whole-genome sequence diversity of human papillomavirus type 18 isolates, *Viruses* 10 (2018), <https://doi.org/10.3390/v10020068>.
- [48] L.S. Arroyo-Muhr, et al., The HPV16 genome is stable in women who progress to in situ or invasive cervical cancer: a prospective population-based study, *Canc. Res.* 79 (2019) 4532–4538, <https://doi.org/10.1158/0008-5472.CAN-18-3933>.

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

Paper II

Hesselberg Løvestad A, Repesa A, Costanzi JM, Lagström S, Christiansen IK, Rounge TB, Ambur OH. Differences in integration frequencies and APOBEC3 profiles of five high-risk HPV types adheres to phylogeny. *Tumour Virus Research* 2022;14:200247.

DOI: <https://doi.org/10.1016/j.tvr.2022.200247>



Differences in integration frequencies and APOBEC3 profiles of five high-risk HPV types adheres to phylogeny

Alexander Hesselberg Løvestad^a, Adina Repesa^a, Jean-Marc Costanzi^b, Sonja Lagström^{b,c,d}, Irene Kraus Christiansen^{b,f}, Trine B. Rounge^{c,e,**}, Ole Herman Ambur^{a,*}

^a Department of Life Sciences and Health, Faculty of Health Sciences, OsloMet - Oslo Metropolitan University, Oslo, Norway

^b Department of Microbiology and Infection Control, Akershus University Hospital, Lørenskog, Norway

^c Department of Research, Cancer Registry of Norway, Oslo, Norway

^d Institute of Clinical Medicine, University of Oslo, Oslo, Norway

^e Centre for Bioinformatics, Department of Pharmacy, University of Oslo, Oslo, Norway

^f Department of Clinical Molecular Biology (EpiGen), Division of Medicine, Akershus University Hospital and University of Oslo, Lørenskog, Norway

ARTICLE INFO

Keywords:

Human papillomavirus
HPV16
HPV18
HPV31
HPV33
HPV45
Minor nucleotide variation
Chromosomal integration
APOBEC3
Alpha-7
Alpha-9
Cancer-related genes

ABSTRACT

Persistent infection with Human Papillomavirus (HPV) is responsible for almost all cases of cervical cancers, and HPV16 and HPV18 associated with the majority of these. These types differ in the proportion of viral minor nucleotide variants (MNVs) caused by APOBEC3 mutagenesis as well as integration frequencies. Whether these traits extend to other types remains uncertain. This study aimed to investigate and compare genomic variability and chromosomal integration in the two phylogenetically distinct Alpha-7 and Alpha-9 clades of carcinogenic HPV types. The TaME-seq protocol was employed to sequence cervical cell samples positive for HPV31, HPV33 or HPV45 and combine these with data from a previous study on HPV16 and HPV18. APOBEC3 mutation signatures were found in Alpha-9 (HPV16/31/33) but not in Alpha-7 (HPV18/45). HPV45 had significantly more MNVs compared to the other types. Alpha-7 had higher integration frequency compared to Alpha-9. An increase in integration frequency with increased diagnostic severity was found for Alpha-7. The results highlight important differences and broaden our understanding of the molecular mechanisms behind cervical cancer induced by high-risk HPV types from the Alpha-7 and Alpha-9 clades.

1. Introduction

Human papillomaviruses (HPVs) are a group of small, double-stranded DNA viruses with a genome size of ~7.9 kb that contains eight genes. The genome consists of early region genes (E1, E2, E4-E7), late region genes (L1, L2), and two non-coding regions, the upstream regulatory region (URR) and the non-coding region (NCR) [1]. Of the early region genes, E5, E6, and E7 encode oncoproteins that promote the transformation of the host cell through induction of cell proliferation and inactivation of cell cycle regulatory and tumour-suppressor mechanisms [2,3]. To date there are over 200 characterised HPVs [4], commonly distinguished by at least 10% nucleotide differences in the L1 gene [5,6] and further divided into lineages ($1 > 10\%$ whole-genome nucleotide differences) and sublineages ($0.5 > 1\%$ difference) [7–9]. There are at least 12 HPV types that are carcinogenic (16, 18, 31, 33, 35,

39, 45, 51, 52, 56, 58, and 59) [10]. Persistent infection with one of these is considered a necessary cause for cervical cancer development [11]. Still, only a minority of persistent infections progress to cancer [12], indicating that additional factors are necessary for cancer progression. All the oncogenic HPV types belong to the genus *Alphapapillomavirus* (Alpha-PV) where they cluster within the species-level clades Alpha-5, Alpha-6, Alpha-7, Alpha-9 and Alpha-11 [13]. The carcinogenic HPVs of the different clades exhibit differences in carcinogenicity and tissue tropism, among other characteristics, suggesting different evolutionary histories that have given rise to their carcinogenic potential as well as differences in the molecular mechanisms behind HPV-induced cancers.

HPVs are considered slowly evolving viruses [14]. Recent studies have uncovered nucleotide variation below the consensus level in HPV genomes present within an infected person [15–20]. HPV genomic

* Corresponding authors.

** Corresponding author. Department of Research, Cancer Registry of Norway, Oslo, Norway.

E-mail address: olam@oslomet.no (O.H. Ambur).

<https://doi.org/10.1016/j.tvr.2022.200247>

Received 1 July 2022; Received in revised form 5 September 2022; Accepted 6 September 2022

Available online 11 September 2022

2666-6790/© 2022 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

variation can have consequences for the infection outcome, and HPV16 sublineages together with the host ethnicity/genetic background has been shown to be associated with different risks of developing cervical cancer [21]. Intra-host HPV nucleotide variation is not uniformly distributed across the genome, as has been revealed in HPV16 cervical cancer cases where the E7 gene has been shown to have few non-synonymous mutations compared to the rest of the genome [16]. Additionally, persistent infections that progress to high-grade lesions or cancer are associated with less intra-host variation relative to infections that are cleared by the immune system [17,20,22].

The mutagenic processes behind HPV intra-host nucleotide variation are currently not fully understood, although it is clear that members of the gene family anti-viral host-defence enzyme apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3 (APOBEC3) contributes to the variation [17,19,20]. APOBEC3 are cytidine deaminases activated in response to viral infections and induces C > T substitutions in the trinucleotide context TCN, where N is any nucleotide (with the exception of APOBEC3G which has a preferred CCN target motif) [23]. APOBEC3-induced mutations can inhibit viral replication and generally lower viral fitness [24]. Accordingly, the APOBEC3 mutation signature is more often observed in HPV genomes from transient infections and pre-cancerous lesions compared to cancer samples and is associated with viral clearance [17,20]. Studies have also reported frequent T > C substitutions [17,19,20]; however, the mutagenic process behind this transition and its role in infection outcome is currently not understood. Another possible source for HPV intra-host nucleotide variation might be the recruitment of low-fidelity polymerases during the replication stage of the viral life cycle [25–27].

A well-studied molecular event related to HPV-induced carcinogenesis is the full or partial integration of the HPV genome into human chromosomes [28]. This is a common genomic event observed in more than 80% of HPV positive tumours and is considered a driving event in cervical carcinogenesis [29,30]. Integrations involving a deletion or disruption of E1 or E2 will lead to overexpression of oncogenes E6 and E7 [31,32], which in turn can lead to an accumulation of mutations and unregulated clonal cell division with a selective growth advantage [28]. Furthermore, integrations can also promote genomic instability in an E6/E7-independent manner by integrating within, or in close proximity to, host oncogenes or tumour-suppressor genes to functionally knock them out or affect their expression levels [33,34]. HPV integrations are associated with local altered genomic landscapes and changes in host gene expression in their vicinity, which might promote genomic instability and carcinogenesis depending on the integration site [35–39]. Integration “hot spots” have been observed repeatedly in high-grade lesions and tumours, indicating that integration in certain chromosomal loci might confer selective growth advantages and increase the risk of developing HPV-induced cancers [33,40,41].

Most studies on molecular mechanisms behind HPV-induced cancers have been conducted on HPV16 and HPV18 genotypes due to their high prevalence and carcinogenic potential [42]. By comparison, the remaining carcinogenic HPV-types are understudied. HPV16 and HPV18 have shown differences in integration frequencies and APOBEC3 interaction, suggesting that HPV-induced cancer development follow dissimilar type-dependent routes [43–45]. Within the Alpha-PVs, HPV16 sorts under Alpha-9 together with HPV31 and HPV33, while HPV18 and HPV45 sort under Alpha-7. It has been shown in previous studies that HPV45 has a high integration frequency (IF) like HPV18, reflecting similarities between these evolutionary closely related HPV types [45].

In this study, we aim to investigate genomic variability and chromosomal integration in cervical cell samples with different morphologies positive for HPV31, HPV33, and HPV45 utilizing the TaME-seq protocol [46]. Additionally, this study will include comparisons to reanalysed HPV16 and HPV18 data from a previous study [43] to gain a more comprehensive understanding of specific characteristics of the distinct clades Alpha-7 (HPV18 and HPV45) and Alpha-9 (HPV16, HPV31 and HPV33). A study going deeper into the nature of genomic

events in these lesser studied carcinogenic HPV types allows for a phylogenetic approach to better understand the molecular mechanisms of host-responses to infections and those responsible for HPV-induced carcinogenesis.

2. Materials and methods

2.1. Sample selection

Cervical cell samples were collected from women attending the Norwegian cervical cancer screening program between January 2005 and April 2008. Recruitment criteria, HPV detection, and genotyping have been described previously [47,48]. In total, 156 HPV16, 75 HPV18, 117 HPV31, 104 HPV33, and 66 HPV45 samples were categorized based on the HPV type and diagnostic category. The diagnostic categories were defined as “non-progressive”, cervical interepithelial neoplasia grade 2 (CIN2) and CIN3+ (Table 1). The non-progressive category consisted of samples with normal cytology (normal cytology the preceding two years and with no previous history of treatment for cervical neoplasia) or samples with atypical squamous cells of undetermined significance (ASC-US) or low-grade squamous intraepithelial lesions (LSIL) with no follow-up diagnosis within four years. The CIN3+ category consisted of samples with CIN3/adenocarcinoma in situ (AIS) and cancer.

2.2. Sample preparation and DNA extraction

Cervical cell samples had previously been collected in ThinPrep PreservCyt solution (Hologic, Marlborough, MA) and pelleted before storage -80°C to retain DNA quality and integrity. Collected samples were stored as both cell material and extracted DNA in a research biobank at Akershus University Hospital. DNA from some samples had to be re-extracted from cell material for this study, and an easyMAG® (Bio-mérieux, USA) was used for the extraction and the eluate stored in a biobank at -80°C . The DNA concentration was measured on Qubit® 3.0 Fluorometer (Life Technologies, USA) to ensure optimal DNA quantity in every sample before the PCR reaction.

2.3. Library preparation and sequencing

Library preparation was done according to the TaME-seq protocol as described previously [46]. To summarise, the samples were tagged using Nextera DNA library prep kit (Illumina, Inc., San Diego, CA). Tagmented DNA underwent target enrichment by multiplex PCR using respective HPV31, 33 and 45 type-specific HPV primers and a combination of i7 index primers and i5 index primers [49] from the Nextera index kit (Illumina, Inc., San Diego, CA). Each sample underwent separate PCR amplifications for the forward and reverse reactions. Sequencing was performed on the Illumina HiSeq2500 platform using 125 bp paired-end reads.

2.4. Sequence alignment

Data were analysed by an in-house bioinformatics pipeline as described previously [46], with some slight changes to the reference genomes and variant calling. The pipeline can be accessed here: <https://github.com/jean-marc-costanzi/TaME-seq/>. Briefly, reads were mapped to the human genome (GRCh38/hg38) using HISAT2 (v2.1.0)[50]. Reference genomes for HPV16, HPV18, HPV31, HPV33, and HPV45 were obtained from the PaVe database [51] and 1 kb overhangs were added to account for the circular HPV genome. BCFtools was used to calculate mapping statistics and coverage. Samples with a mean sequencing depth of $<300\times$ were excluded from the analysis.

Table 1

Number of samples sequenced and analysed and mean mapping statistics for each diagnostic category of HPV16, HPV18, HPV31, HPV33 and HPV45 infections.

| Diagnostic category | Sequenced samples | Analysed samples | Mean numbers in the analysed samples | | | |
|------------------------|-------------------|------------------|--------------------------------------|----------------------------|--------------------------|---|
| | | | Raw reads | Reads mapped to target HPV | Mean HPV genome coverage | Fraction of HPV genome covered by min. 100x |
| HPV16 | | | | | | |
| Non-progressive | 55 | 21 | 1.3 M | 1.3 M | 11148 | 0.77 |
| CIN2 | 46 | 25 | 0.6 M | 0.5 M | 4462 | 0.70 |
| CIN3+ | 55 | 31 | 1.3 M | 1.2 M | 9483 | 0.74 |
| HPV18 | | | | | | |
| Non-progressive | 16 | 12 | 39 M | 26 M | 48129 | 0.86 |
| CIN2 | 13 | 9 | 77 M | 40 M | 55097 | 0.86 |
| CIN3+ | 46 | 28 | 24 M | 13 M | 29138 | 0.82 |
| HPV31 | | | | | | |
| Non-progressive | 18 | 10 | 10 M | 6 M | 26508 | 0.87 |
| CIN2 | 22 | 20 | 14 M | 5 M | 23695 | 0.89 |
| CIN3+ | 77 | 54 | 9.3 M | 5.7 M | 26657 | 0.88 |
| HPV33 | | | | | | |
| Non-progressive | 12 | 9 | 16 M | 10 M | 24350 | 0.90 |
| CIN2 | 15 | 9 | 20 M | 10 M | 30699 | 0.95 |
| CIN3+ | 77 | 70 | 11 M | 7 M | 38731 | 0.97 |
| HPV45 | | | | | | |
| Non-progressive | 25 | 21 | 22 M | 8 M | 21593 | 0.83 |
| CIN2 | 14 | 12 | 39 M | 19 M | 27608 | 0.92 |
| CIN3+ | 27 | 23 | 32 M | 13 M | 32066 | 0.85 |

2.5. Sequence variation analysis

Nucleotide counts mapped to the HPV reference genomes were retrieved from the HISAT sequence alignment and average nucleotide mapping quality values were retrieved from the BCFtools mpileup output using an in-house R (v3.5.1) script as described in [46]. Briefly, for a variant to be called it had to be present in more than two reads in a position with $\geq 100x$ depth, have a Phred quality score ≥ 30 and a frequency $\geq 1\%$. In addition, the variant calling of minor nucleotide variants (MNVs) was done in a reference-independent manner where the most frequent base in each position was termed the major variant followed by the second most frequent as the MNV. The MNVs had to be present in both the independently amplified F and R reactions, unless where there was discordance between the F and R reactions – then the MNV with the highest coverage was called. HPV NCR have regions of homopolymeric T tracts (HPV16:4156–4173 and 4183–4212, HPV18:4198–4234, HPV31:4072–4077 and 4145–4167, HPV33:4149–4167 and 4186–4195, HPV45:4184–4219), which can cause polymerase or sequencing errors at high frequencies and were therefore filtered out during the variation analysis.

The ratio of non-synonymous to synonymous substitutions (dN/dS) was calculated to indicate whether genes in the different diagnostic categories were more or less prone to amino acid changes.

For mutational signature analysis, all nucleotide substitutions were classified into six base substitutions, C > A, C > G, C > T, T > A, T > C, and T > G, and further into 96 trinucleotide substitution types that include information on the bases immediately 5' and 3' of the substituted base. An extended mutation signature analysis was also done to investigate mutations in the APOBEC3A-favoured genomic context YTCA and APOBEC3B-favoured genomic context RTCA.

To investigate if the number of APOBEC3 target sequences differed between HPV types, FUZZNUC from the EMBOSS package (<http://emboss.toulouse.inra.fr/cgi-bin/emboss/fuzznuc>) was employed using reference genomes for HPV16, HPV18, HPV31, HPV33, and HPV45 obtained from the PaVE database. Both strands were investigated, and the search patterns used were TCA, YTCA, RTCA and NCN.

To calculate the proportion of TCA motifs out of all available NCN motifs, the number of TCA motifs were divided by the number of NCN motifs for each HPV type, and this proportion was treated as the expected proportion of C > T substitutions in the TCA motifs. To calculate the difference of observed vs expected proportions, the following formula was used (Observed proportions/Expected proportions)²/

Expected proportions. These values were used in a Wilcoxon rank sum test to investigate whether the difference in observed/expected proportions between the clades differed significantly. Variation and dN/dS analyses were performed using an in-house R (v3.5.1) script.

2.6. Detection of chromosomal integration sites and validation by sanger sequencing

Integration site detection was performed as described previously [46]. Briefly, a two-step analysis was employed. First, read pairs with one of the reads mapping to the human genome and the other to HPV were identified using HISAT2. Second, unmapped reads were re-mapped using LAST (v876) aligner (options -M -C2) [52] to increase detection of human-HPV read pairs. Reads sharing identical start and end coordinates were considered likely PCR duplicates and excluded from the analysis.

Validation of integration sites for HPV16 and HPV18 is previously described [43]. The Illumina reads from the respective HPV31, HPV33 and HPV45 sequencing reactions were used to make *in silico* DNA templates for design of integration-targeting primers suited for PCR and Sanger sequencing. Hybrid sequences containing human and HPV-specific sequences spanning the reported integration breakpoint, were used as templates. Primer3 [53] was used to create optimal primer pairs that included a human-specific forward primer and an HPV-specific reverse primer. Phusion™ Master Mix (Thermo Scientific, USA) was used to prepare the PCR reaction mix. The PCR conditions were as follows: initial denaturation at 98 °C for 30 s; 30 cycles at 98 °C for 10 s, at 60 °C for 30 s and 72 °C for 15 s; final extension at 72 °C for 10 min.

Samples were sequenced on the ABI® 3130xl/3100 Genetic Analyzer 16-Capillary Array (Thermo Fisher Scientific Inc., Waltham, MA) using BigDye™ Terminator v1.1 cycle sequencing kit (Thermo Fisher Scientific Inc., Waltham, MA). Sanger sequencing data was further processed in Geneious Prime (v2020.2.2) and if the sequence was homologous to the same chromosomal locus and HPV type as reported, the HPV integration was considered confirmed. Samples with inconclusive Sanger sequencing results that showed several unspecific bands on the agarose gel were re-run using a touchdown PCR with an additional 6 extra cycles. If the samples still had unspecific bands, individual bands were cut out from the agarose gel and extracted using Wizard® SV Gel and PCR Clean-Up System kit (Promega, USA) following the manufacturer's instructions before Sanger sequencing.

2.7. Determining microhomology regions

BLASTn and/or BLAT were used to identify short homologous sequences at the integration breakpoint in the Sanger-confirmed HPV integrations. If > 3 nt overlapping sequences were present between the human and HPV genome, it was designated a microhomology sequence. The overlapping bases were identified using the Geneious Prime genome browser after the assembly of Sanger reads.

2.8. Functional annotation of genes within 10 kb of reported integration sites

All genes 10 kb upstream or downstream of the reported integration site were identified by visual inspection in Geneious Prime and their molecular function annotated using Genecards (<https://www.genecards.org>). Genes were classified as cancer-related genes (CRGs) if they were involved in cell cycle regulation, apoptosis, tumour suppressor mechanisms, cancer-related pathways, genes interacting with these pathways, or if a cancer-related SNP association was assigned.

2.9. Statistical methods

Non-parametric Chi-square of independence was used to determine whether there was a significantly ($p < 0.05$) higher number of breakpoints in the E1, E2, E4 and L2 genes for Alpha-7 and in E1, L2 and NCR in Alpha-9 than would be expected by chance.

To understand the relationship between the dependent variables (MNVs, samples with integrations, integration breakpoints) and the independent variables (HPV type or diagnostic category), a generalized linear model (glm) was used. The glm used a negative binomial distribution for the number of integrations and MNV model and binomial distribution for the other models. Following this, multiple comparisons of means using Tukey HSD was done using the R package multcomp [54] to test the differences between the categories. To test the differences in APOBEC3 signature mutations and clades, a Wilcoxon rank sum test were used. All statistical tests were done in R (v3.6.3). The output of the tests can be found in the [supplementary material D](#).

Ethical approval

This study was approved by the regional committee for medical and health research ethics, Oslo, Norway (REK 2017/447).

3. Results

3.1. Characteristics and sequencing statistics

In total, 518 HPV16, HPV18, HPV31, HPV33 and HPV45 positive cervical cell samples stratified into the diagnostic categories non-progressive, CIN2 and CIN3+ were sequenced. Six samples were removed from further analysis based on a MNV pattern suggestive of co-infection of different viral sublineages ([Supplementary figure A1](#)). After removing samples that did not pass the filtering criteria ($n = 164$), 77 HPV16, 49 HPV18, 84 HPV31, 88 HPV33 and 56 HPV45 samples underwent comparative MNV and integration analysis. The mean sequencing depth for samples within the different diagnostic categories ranged from 4462 for HPV16 CIN2 to 55097 for HPV18 CIN2. The proportion of the genome covered with a minimum depth of 100x within the categories ranged from 0.70 for the HPV16 CIN2 category to 0.97 for the HPV33 CIN3+ categories ([Table 1](#)).

3.2. Minor nucleotide variation profiles reveal a higher number of MNVs in HPV45 positive samples

A total of 10664 MNVs were identified in the 354 analysed samples. Most of the MNVs were low-frequency variants with 1716 MNVs having

a minor base frequency $\geq 5\%$ and 850 with a minor base frequency $\geq 10\%$. The number of MNVs were significantly higher in HPV45 samples compared to other HPV types ($p < 0.001$), the average being 47.9, 47.92 and 34.91 in the HPV45 non-progressive, CIN2 and CIN3+ categories, respectively ([Fig. 1](#), [Table 2](#)). Excluding HPV45, HPV16 non-progressive was the category with the highest average number of MNVs with 34.57, while HPV33 CIN2 category had the lowest number of MNVs with an average of 19. The standard deviation of the number of MNVs within categories was also found to be highest in HPV45 samples. The same were true when investigating the MNV frequencies, with HPV45 having the highest mean and SD MNV frequency in all categories ([Supplementary figure A2](#) and [Table 2](#)). No significant differences were found in comparisons of diagnostic categories.

HPV MNVs were found within genomic elements of all the different categories. Most variation was found in HPV45 positive samples, where the genomic elements E4, URR and NCR in non-progressive samples and the E4 and E2 in CIN2 samples had the highest number of mean variants ([Supplementary figure A3](#)).

In total, 5734 nonsynonymous and 244 nonsense mutations were observed in the dataset. Most genes, across all categories, had a dN/dS ratio >1 or close to 1 ([Supplementary figure A4](#)).

3.3. APOBEC-related mutational signatures

The two most common substitutions in the dataset, C > T and T > C, were observed across all the diagnostic categories for all HPV types ([Supplementary figure A5](#) and [A6](#)). To investigate APOBEC3-induced mutations, C > T mutations in the APOBEC3-preferred trinucleotide context TCN found within different diagnostic categories for the different HPV types were compared against each other. We observed that TCA was the trinucleotide context with the highest proportion across all diagnostic categories in Alpha-9 samples, while no such pattern was observed for Alpha-7 samples ([Fig. 2](#)). C > T mutations in the TCA context were found to differ significantly between the Alpha-7 and Alpha-9 clades ($p < 0.001$), and the results were found to be consistent when the analysis was extended to include the TCW trinucleotide context and the inclusion of C > G substitutions, but not when investigating C > G substitutions by themselves. When investigating each HPV-type separately, HPV16 showed a decrease in TCA proportion of C > T MNVs with increased lesion severity compared to HPV31 and HPV33 ([Supplementary figure A6](#)). The extended mutation signature analysis did not reveal any strong signal for either APOBEC3A (YTCA) or APOBEC3B (RTCA) preference in the dataset ([Supplementary figure A7](#)).

The number of TCA motifs found within each HPV types differed between Alpha-7 (164 HPV18, 159 HPV45) and Alpha-9 (186 HPV16, 190 HPV31, 208 HPV33) ([Supplementary table B1](#)), with most motifs being present on the minus strand ([Supplementary figure A8](#)). Values for RTCA and YTCA motifs can be seen in [supplementary figure A9](#) and [supplementary table B1](#). To further investigate if the number of C > T substitutions in the TCA context occurred more frequently than expected the proportion of TCA motifs out of all available NCN motifs were calculated. The proportions of TCA motifs were found to range from 0.05 (HPV45) to 0.07 (HPV33) and were interpreted as the expected proportion of C > T substitutions in that trinucleotide context, assuming that substitutions are equally likely in all NCN motifs ([Supplementary table B2](#)). The difference in observed/expected C > T substitutions in the TCA context were found to be significantly larger for Alpha-9 samples ($p < 0.001$), thus there were found to be significantly more C > T mutations in the TCA context than would be expected in Alpha-9 relative to Alpha-7 ([Fig. 3](#)).

3.4. Higher integration frequencies (IFs) in Alpha-7s compared to Alpha-9s

The number of integrations in the Alpha-7 significantly outnumber those of the Alpha-9 clade although more than twice as many Alpha-9

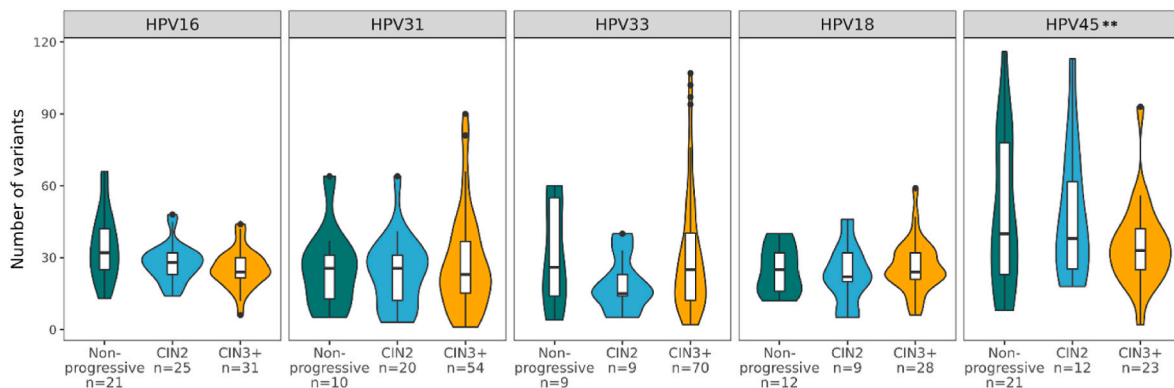


Fig. 1. Number of variants in HPV16, HPV18, HPV31, HPV33 and HPV45 positive samples. Violin plots representing the number of variants in the different diagnostic categories on the x-axis. Box-and-whisker plots are added to show the median number of MNVs (horizontal line), the box represents the 25% and 75% percentiles, and the whiskers represents the minimum and maximum number of MNVs found within one sample. The number of samples (n) is indicated below each category. Double asterisk (**) represents that HPV45 has significantly more MNVs compared to all other types ($p < 0.01$).

Table 2

Number of samples analysed and different statistics of MNVs for HPV16, HPV18, HPV31, HPV33 and HPV45 positive samples stratified across the diagnostic categories.

| Diagnostic category | Analysed samples | Mean number of variants | Minimum number of variants | Maximum number of variants | Standard deviation of number of variants | Mean MNV frequency | Standard deviation of MNV frequency |
|---------------------|------------------|-------------------------|----------------------------|----------------------------|--|--------------------|-------------------------------------|
| HPV16 | | | | | | | |
| Non-progressive | 21 | 34.57 | 13 | 66 | 14.25 | 2.74 | 3.65 |
| CIN2 | 25 | 27.80 | 14 | 48 | 7.88 | 2.80 | 4.82 |
| CIN3+ | 31 | 25.35 | 6 | 44 | 8.00 | 2.67 | 3.66 |
| HPV18 | | | | | | | |
| Non-progressive | 12 | 25.08 | 12 | 40 | 9.97 | 3.21 | 4.69 |
| CIN2 | 9 | 23.00 | 5 | 46 | 12.68 | 2.89 | 4.43 |
| CIN3+ | 28 | 26.36 | 6 | 59 | 11.05 | 4.35 | 6.79 |
| HPV31 | | | | | | | |
| Non-progressive | 10 | 25.70 | 5 | 64 | 17.12 | 4.53 | 6.72 |
| CIN2 | 20 | 23.90 | 3 | 64 | 14.84 | 3.30 | 5.40 |
| CIN3+ | 54 | 27.63 | 1 | 90 | 20.55 | 4.29 | 7.05 |
| HPV33 | | | | | | | |
| Non-progressive | 9 | 30.78 | 4 | 60 | 21.16 | 2.59 | 2.50 |
| CIN2 | 9 | 19.00 | 5 | 40 | 11.47 | 2.49 | 4.68 |
| CIN3+ | 70 | 30.74 | 2 | 107 | 24.77 | 3.36 | 4.77 |
| HPV45 | | | | | | | |
| Non-progressive | 21 | 47.90 | 8 | 116 | 31.28 | 6.79 | 9.07 |
| CIN2 | 12 | 47.92 | 18 | 113 | 28.52 | 4.16 | 6.64 |
| CIN3+ | 23 | 34.91 | 2 | 93 | 17.89 | 4.63 | 7.37 |

samples were sequenced (Fig. 4 and Table 3). In total, 42.8% of Alpha-7 samples had at least one integration site reported, significantly more than Alpha-9 with 6.4% ($p < 0.001$). Overall, 154 integration sites were observed in the whole dataset, of which 85% (131/154) were Alpha-7 and 15% (23/154) Alpha-9 (Table 3, Fig. 4). Alpha-7 also had the highest IF in all diagnostic categories with 21% of non-progressive, 33% of CIN2 and 61% of CIN3+ samples having integrations, thus higher IF correlated significantly with diagnostic severity ($p < 0.01$). Alpha-7 samples had significantly more integration sites in samples with integration compared to Alpha-9 ($p < 0.001$), with an average of 3.4, 3.14 and 2.74 integrations in the non-progressive, CIN2 and CIN3+ categories, respectively, compared to Alpha-9 with 1.25, 1.5 and 1.5.

Comparing the HPV-types within the two clades revealed some differences between the related types. Overall, HPV18 had significantly more samples with integrations than any other type, while HPV45 CIN3+ (the only diagnostic category with HPV45 integrations) had significantly more than all other Alpha-9 types. HPV18 also had significantly more integrations per sample than all other types, as well as the

sample with the most reported integrations with 21, compared to 4, 3, 1 and 1 for HPV45/16/31/33, respectively. Within Alpha-9, HPV16 reported higher IF than both HPV31 and HPV33, as well as higher average number of integrations per sample (Supplementary table B3). A complete list of annotated integration breakpoints can be found in supplementary table C1.

3.5. Deletions and breakpoints in the HPV genome

In Alpha-7 samples with integrations, breakpoints were found in all genetic elements of the HPV genome, except NCR, while Alpha-9 integrations lacked breakpoints in E7 and E4 (Supplementary figure A10). Any breakpoint location bias was investigated using the number of reported Alpha-7 and -9 integrations divided by the average gene lengths in each clade. It was observed that Alpha-7 samples had more breakpoints in the E1, E2, E4 and L2 genes than would be expected by chance, however, this difference was not statistically significant ($p = 0.24$). Breakpoints in Alpha-9 samples were observed in E1, L2 and NCR more

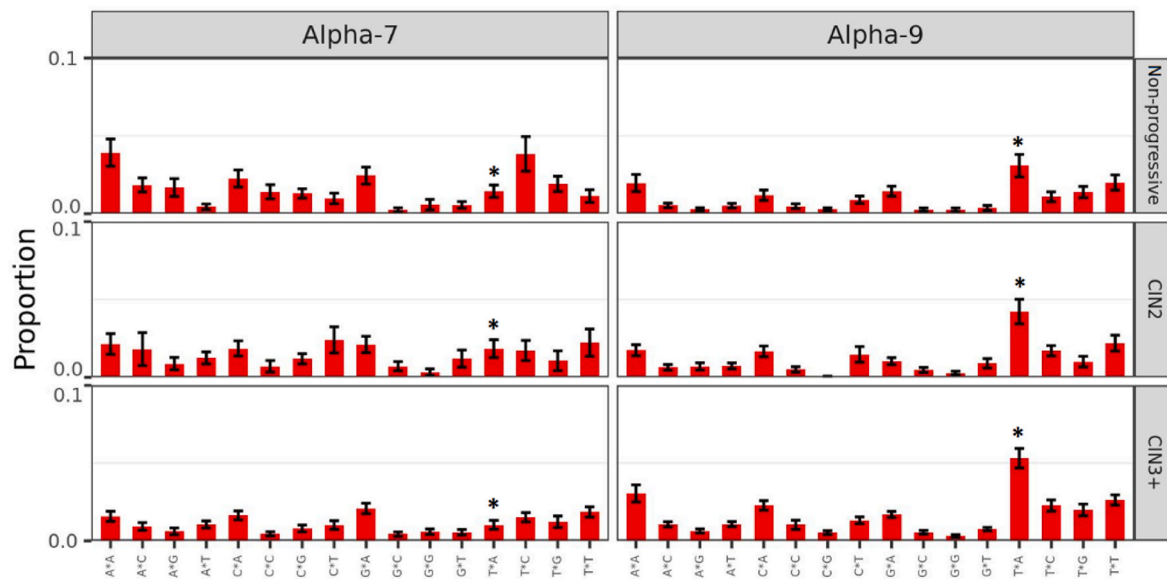


Fig. 2. C > T mutational signatures in Alpha-7 (HPV18 and HPV45) and Alpha-9 (HPV16, HPV31 and HPV33) positive samples across the different diagnostic categories. The mean proportion of C > T mutations is shown on the y-axis and the different trinucleotide contexts are shown on the x-axis. Error bars represent the standard error of the mean. Asterisk (*) denotes C > T substitutions found in the TCA context and was found to be overall significant between the clades.

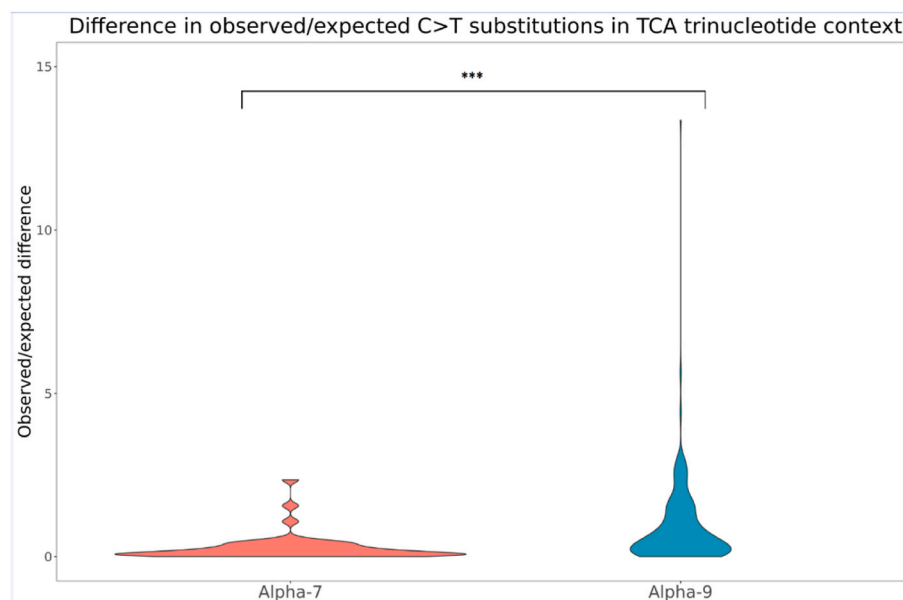


Fig. 3. Violin plots representing the difference in observed/expected number of C > T substitutions in the TCA context for individual samples in Alpha-7 and Alpha-9.

often than expected by chance, but was also not significant ($p = 0.20$). Combining and investigating the percentage of integrations with breakpoints in E1/E2 across the different diagnostic categories revealed that 38%, 36% and 51% of the integrations caused breakpoints in E1/E2 in the non-progressive, CIN2 and CIN3+ categories, respectively (Fig. 5a). A detailed figure showing sample integrations and number of breakpoints in E1/E2 can be seen in [Supplementary figure A11](#).

To investigate large HPV genomic deletions, coverage plots for each sample were inspected for extended regions without or with exceptionally low relative sequencing coverage. In addition to the six samples with deletions reported in the previous study (one HPV16 and five HPV18), the coverage plots revealed 11 additional samples with deletions or partial deletions (One HPV31 and 10 HPV45). All HPV45 positive samples with deletions were reported as having integrations

while the single HPV31 sample did not.

3.8 Presence of human cancer-related genes within ± 10 kb of integration sites.

Due to the uneven number of integrations found within Alpha-7 (131 integrations) and Alpha-9 (23 integrations), all integrations in the dataset were combined when investigating integrations in human genes within ± 10 kb of the integration sites. The results revealed that CRGs were present in 41% (12/29), 40% (10/25) and 59% (59/100) of non-progressive, CIN2 and CIN3+ samples, respectively (Fig. 5b, see [Supplementary figure A12](#) and [supplementary table C1](#) for more details). Of the integrations within ± 10 kb of CRGs, 58% (7/12), 80% (8/10) and 78% (46/59) were integrated inside the ORF of the reported CRG in the non-progressive, CIN2 and CIN3+ categories ([Supplementary figure A13](#)). Three CRGs were each found twice ± 10 kb of integrations sites in

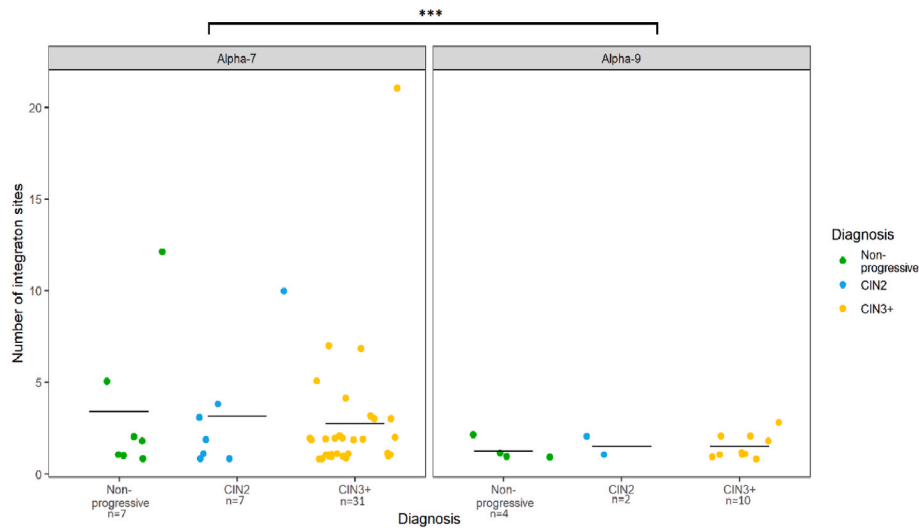


Fig. 4. Number of integrations in Alpha-7 and Alpha-9 samples with reported integration(s), stratified by diagnostic category. Horizontal line represents the mean number of integrations and n denotes number of samples in each category with reported integrations.

Table 3
Number of integrations in Alpha-7 and Alpha-9 positive samples, stratified by diagnostic category.

| Diagnostic categories (samples) | Number of samples with integrations (Frequency %) | Total number of integration sites | Mean number of integrations per sample |
|---------------------------------|---|-----------------------------------|--|
| Alpha-7 | | | |
| Non-progressive (n = 33) | 7 (21.2%) | 24 | 3.43 |
| CIN2 (n = 21) | 7 (33.3%) | 22 | 3.14 |
| CIN3+ (n = 51) | 31 (60.8%) | 85 | 2.74 |
| Total (n = 105) | 45 (42.8%) | 131 | 2.91 |
| Alpha-9 | | | |
| Non-progressive (n = 40) | 4 (10%) | 5 | 1.25 |
| CIN2 (n = 54) | 2 (3.7%) | 3 | 1.5 |
| CIN3+ (n = 155) | 10 (6.4%) | 15 | 1.5 |
| Total (n = 249) | 16 (6.4%) | 23 | 1.44 |

different samples, being RCAN2 (HPV18 CIN2 and HPV18 CIN3+), MIR205 (HPV16 CIN3+ and HPV45 CIN3+) and KLHL29 (HPV18 CIN2 and HPV31 CIN3+).

The percentages of integrations in the human genome with annotated genes present within ± 10 kb of the integration site were 76% (22/29), 68% (17/25) and 75% (75/100) in the non-progressive, CIN2 and CIN3+ categories, respectively (Supplementary figure A14). The percentages of integrations inside the ORF of human genes were 45%, 44% and 56% in the non-progressive, CIN2 and CIN3+ categories (Supplementary figure A15).

Alpha-7 samples had integrations in all human chromosomes except chromosome 18 and 21 (Supplementary figure A16). 28/131 Alpha-7 integrations were in chromosome 1 and 2. Alpha-9 samples had integrations in 15 different chromosomes, with chromosome 5 and 8 having most.

3.6. Validation of reported integration sites by sanger sequencing

In total, 31 reported integration sites in 21 patient samples qualified for validation by Sanger sequencing after QC filtering based on read mapping and sequencing artefacts [55]. Of the integration sites, 5 integrations were detected in HPV31 samples (84 analysed samples), 1 integration in HPV33 samples (88 analysed samples) and 25 integrations in HPV45 samples (56 analysed samples). In total, 21 of the 31 reported

integrations sites were validated by Sanger sequencing which confirmed correct chromosomal coordinates and HPV type. The remaining 10 did not yield high-quality chromatograms, possibly due to low DNA concentrations, suboptimal PCR amplification, unspecific primer hybridization or genomic structural rearrangements often associated with HPV integrations [36,56]. Microhomology regions were identified in 19% (5/21) of the confirmed HPV integrations, the length ranging from 3 bp to 12 bp and are presented in supplementary table C2.

4. Discussion

This study aimed to investigate and compare type-specific intra-host variation and integration characteristics of five high-risk HPV types belonging to Alpha-7 and Alpha-9 across diagnostic categories of increasing severity. We observed differences between the diagnostic categories, as well as between Alpha-7 and Alpha-9 clades. The differences adhere to their phylogenetic assortment, where there is a statistically significant signal of APOBEC3-induced C > T mutations in Alpha-9 samples that is not found in Alpha-7. IF is also significantly higher in samples positive for Alpha-7 HPV-types relative to Alpha-9.

Minor nucleotide variants (MNVs) are variants found below the consensus level that might play a role in the development of cancer [16]. In our dataset, we observe that HPV45 samples have more MNVs compared to the other four investigated HPV types, and that Alpha-7s have more variation in E4 compared to the Alpha-9s. The biological significance of these results is currently unclear. The MNVs are called in a reference-independent manner, and we rule out that these results are artefacts that could arise from mapping to divergent reference genomes. Co-infections with two or more variants of the same HPV type could be a likely alternative source of diversity [57] and six apparent co-infections were excluded from the analysis based on indicative MNV patterns (Supplementary figure A1). When comparing the amount of MNVs, we observed no differences between the diagnostic categories, suggesting that the total quantity of MNVs in a sample is not directly associated with carcinogenic risk. Rather than quantity, the quality of MNVs in their HPV genomic context may be of significance, as has been shown for the HPV16 E7 gene in cervical cancer cases and for certain positions in HPV16 URR where MNVs have been shown associated with developing CIN3+ [16].

C > T substitutions in the trinucleotide context TCA and TCT is correlated with APOBEC3A/B activity [20]. While they are part of the innate immune system in response to viral infections, their mutation signatures are also commonly found in host genomes of HPV-positive

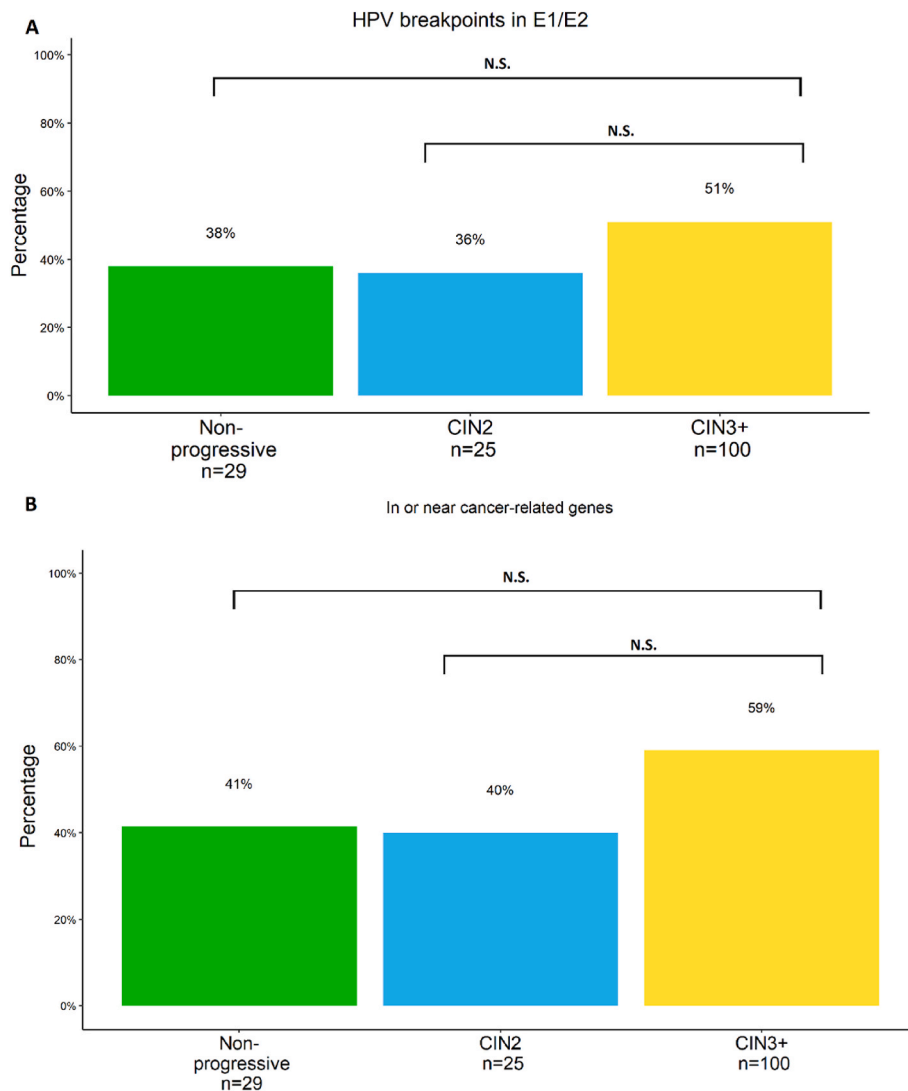


Fig. 5. A) Percentage of integrations with breakpoints in E1/E2, stratified by diagnostic category. B) Percentage of integrations with CRGs within ± 10 kb of integration site or inside the ORF, stratified by diagnostic category. n is the number of integrations in each category.

cancer cells as well as in viral HPV genomes [58]. Most research investigating APOBEC3-HPV interactions have had a focus on HPV16, while less research has been done on the other high risk types [17,20,59,60]. In our study, APOBEC3A/B induced C > T mutations were found to be most common in the trinucleotide context TCA for Alpha-9s, while the Alpha-7s did not have this pattern. This is to our knowledge the first study to show that differences in APOBEC3-induced mutation profiles between any HPV types are highly significant and that the difference is in concordance with phylogeny. The number of all available TCA motifs in the HPV genomes studied here differs between the types, with Alpha-7s (average 162) having less than Alpha-9s (average 195). To investigate whether this difference in the number of available motifs affected the observed mutational C > T patterns, the proportion of TCA motifs out of all available NCN motifs were compared against the observed proportion of C > T mutations in TCA motifs found within samples. The difference between observed and expected proportions of C > T mutations in TCA motifs were found to be significantly higher for Alpha-9s than Alpha-7s. Thus, Alpha-9 samples were found to have relatively more C > T mutations in the TCA-motifs, even when correcting for the higher abundance of TCA motifs in the genomes relative to Alpha-7. This finding suggests that the Alpha-9 infections trigger a detectable APOBEC3-response not found in Alpha-7 infections; a finding that warrants further investigation considering their differences in

clinical epidemiology [10], human molecular and genomic cancer characteristics [44,61], evolutionary histories [62] and impact of viral life-cycle factors and expected tropisms [63]. Our previous study looked at HPV16 and HPV18 and by including one additional Alpha-7 and two more Alpha-9 types in the comparative analysis, the phylogenetic dichotomy in mutational signature was established more broadly and emphasises the biological significance of the results. Using different whole genome sequencing protocols Hirose et al., 2018 identified the same APOBEC signature (C > T in the TCA context) across the three Alpha-9s HPV16/52/58, and the exact same signature was described in HPV16 (signature A) in Zhu et al., 2020 [17,20]. The combined presence of the APOBEC signature in the Alpha-9s now therefore encompass HPV16/31/33/52/58. Yet, contrary to current understanding it cannot be established that APOBEC3-induced mutagenesis in viral genomes is a general detectable feature in high-risk HPV infections. In HPV16 positive samples, the number of APOBEC3-related nucleotide substitutions decreases with lesion severity, which also has been shown in previous studies [17,20,43]. However, this decrease is not observed in HPV31 and HPV33 positive samples, suggesting this is not an Alpha-9 specific tendency, but rather a feature of HPV16 carcinogenesis. Our results are in corroboration with previous studies that have observed that the number of APOBEC3-related nucleotide substitutions decreases with lesion severity in HPV16 positive samples [17,20,43].

The lack of APOBEC3 activity found in Alpha-7 genomes might be driven by differences in host-response to viral infections between the different cell types and/or by genomic variation between the HPV types. Alpha-7 types have been found to be significantly more common in cases of adenocarcinoma (ADC) than squamous cell carcinoma (SCC), while Alpha-9 is the predominant type found in SCC [64,65]. These differences are likely reflecting type-specific tropisms and cells of cancer origin, where Alpha-9 is found to associate with lesions in squamous cells while Alpha-7 predominantly cause lesions in glandular cells [66]. Additionally, expression profiles of SCC tumours caused by Alpha-7 and Alpha-9 types have been shown to differ in expression levels of keratin gene family members [44]. APOBEC3 substitutions are associated with viral clearance in HPV16 infections [20], and it is interesting to note that while HPV31 and HPV33 have been shown to have a high risk to progress to CIN3, their risk to progress from CIN3 to invasive cervical cancer is relatively low compared to HPV16/18/45 [66]. We observe a decrease in APOBEC3-related nucleotide substitutions in HPV16 positive CIN3+ samples, however, this pattern is not present in HPV31/33. Thus, one can speculate that the different HPV types possess different abilities to trigger APOBEC3-activity, and that this differs both between Alpha-7 and Alpha-9 types as well as between HPV16 and HPV31/33. In a genome evolution perspective, we note that the preferred APOBEC TCN-motif is underrepresented in Alpha-HPV genomes generally and more so in the URR of HPV16 than HPV18/31 which includes the origin or replication and promoter of the E6/7 oncogenes [67]. Further comparative studies into APOBEC-HPV interaction mechanisms and evolutionary dynamics are warranted in order to better understand the trade-off between immune exposure and oncogenicity in individual HPV types.

Integration of HPV-genomes in the human genome is a suggested driver event during HPV-induced carcinogenesis and previous studies have shown that the IF differs between genotypes in cancers [30]. Our results are in line with these studies showing that Alpha-7s HPV18/45 have a higher IF than Alpha-9s HPV16/31/33, and that HPV16 have a higher IF than HPV31 and HPV33 [45]. Women with invasive cervical cancer caused by HPV16, HPV18 and HPV45 are typically younger than women with cervical cancer caused by other HPV types, suggesting that infection with these three genotypes progress to invasive cervical cancer faster than other types [64]. In this study, Alpha-7-positive samples show a significant increase in IF with increase in diagnostic severity, and HPV45-positive samples only had reported integrations in the CIN3+ category. Alpha-7 positive samples with integrations also had significantly more integrations per sample compared to Alpha-9 positive samples. These findings might reflect conserved differences in the biology of Alpha-7 and Alpha-9 types that affects IFs, for example in that Alpha-7s more often than Alpha-9s can warrant integration(s) as a contributing factor to drive oncogenic transformation.

Integrations are associated with increased genomic instability, mainly through overexpression of viral oncogenes E6/E7, but also by triggering host oncogenes and disrupting of tumour suppressor genes [28,68,69]. One of the mechanisms by which integrations can cause overexpression of E6/E7 is the disruption of the E2 gene upon linearization of the circular HPV genome. When we looked at all HPV breakpoints combined, 51% of samples with integrations in the CIN3+ category had breakpoints in E1/E2, compared to 38% and 36% in the non-progressive and CIN2 categories, respectively. In addition, when investigating the coverage plots, nine of out fifteen HPV45 samples with reported integrations showed either full deletions or partial deletions in regions encompassing E1/E2.

Another mechanism by which HPV integrations can step up carcinogenic transformation is by integrating inside the ORF or in genomic proximity of host oncogenes and tumour suppressor genes, and thereby disrupting their function or altering their expression levels, respectively [30,68–70]. We found that the presence of at least one human CRG within ± 10 kb of integration sites increased with almost 20% in the most severe category (CIN3+ at 59%). Chromosomal integrations have

been shown to cause genomic instability in the vicinity of the integration sites by causing structural rearrangements and affecting host gene expression [37,56]. Several observed CRGs in our dataset have previously been reported in studies investigating HPV integrations and are associated with cervical cancer, including TP63, MIR205HG, MMP12 and ENO1 [30,71–73]. Additionally, RCAN2, KLHL29 and MIR205HG were observed close to integrations twice in independent samples. Decrease in RCAN2 expression has been associated with tumour proliferation in colorectal cancers, while differential methylation patterns of KLHL29 have been observed in small and large anal cancer tumours [74,75]. MIR205HG, on the other hand, has been implicated in playing a role in the development in cervical cancer by targeting and regulating genes involved in proliferation, migration and apoptosis of cervical cancer cells [71,76,77]. While the role of MIR205HG in HPV-induced carcinogenesis is established, the presence of RCAN2 and KLHL29 close to integration sites in more than one sample may also suggest their involvement in HPV-induced carcinogenesis. These findings warrant further investigation.

Taken together, the increased number of integrations with the increase in diagnostic severity observed for Alpha-7s and the general tendencies for having breakpoints in E1/E2 and integrating within ± 10 kb of CRGs, supports the notion that integrations are key molecular events in driving HPV-induced carcinogenesis. The differences between the diagnostic categories regarding integration breakpoints in E1/E2 and proximity to CRGs, were not statistically significant when a glm model was applied to the data. However, the number of observations in the three categories differ substantially, and studies including more samples in the “non-progressive” category combined with follow-up data, should be conducted to ascertain their role in HPV-induced carcinogenesis. Furthermore, the difference in IFs of the Alpha-7s and -9s suggests that IF is a consistent phenomenon within phylogenetically related HPV types. What drives this difference between the clades is currently poorly understood. Different high risk HPVs produce different splice isoforms of viral oncogenes E6 and E7 and do also have different capabilities of inducing p53-degradation among other differences [62,63,78,79]. It is possible that the discrepancy observed is due to Alpha-7 oncoproteins having weaker oncogenic potential relative to Alpha-9 oncoproteins and therefore Alpha-7 infections to a larger extent require integrations to drive carcinogenesis, reflecting that nearly all Alpha-7-induced tumours have integrated viral DNA and Alpha-9 can induce cancer in episomal form [44]. HPVs are also known to induce DNA damage and uses DNA damage response pathways for the amplification of the viral genome, which could consequently lead to integration of HPV DNA by nonhomologous end joining and/or microhomology-mediated recombination [32,80–84]. To our knowledge there is nothing in the existing literature that directly compares the ability of different HPV types to induce DNA damage. Since viral proteins have been shown to induce DNA damage an alternative explanation to these clade-specific differences in IFs beyond oncogenic potential of E6/7 remains an option. However, more research into the subject is needed to better understand the molecular mechanisms which drive different manifestation of IFs.

5. Conclusions

This study reveals differences in APOBEC3-mediated mutations in concordance with evolutionary related HPV types, where Alpha-9 positive samples have a clear APOBEC3 mutation signature not observed for Alpha-7. Additionally, Alpha-7 samples are shown to have significantly more integrations and an increase in number of integrations with increased diagnostic severity. This study expands our knowledge, beyond HPV 16 and HPV18, by including three additional high risk HPV types and shows that the type-specific patterns for these molecular events extends to more closely related carcinogenic HPV types within the Alpha-7 and Alpha-9 clades. The results broaden our understanding of the molecular mechanisms behind HPV-induced cancers while also

shedding light on some of the similarities and differences between the HPV types investigated.

Authors' contributions

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

Funding

This work was supported by a PhD grant to AHL and a Research track grant to AR from Faculty of Health Sciences, Oslo Metropolitan University, and by a post-doctoral research grant to JMC from the South-Eastern Norway Regional Health Authority, project number 2020010. The work was also supported by innovation grants from the South-Eastern Norway Regional Health Authority 2019 and the Research Council of Norway (FORNY2020, project number 296671). The funders had no role in study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

Data statement

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

Declaration of competing interest

We have no competing interests to declare.

Data availability

The data presented in this article are not readily available because of the principles and conditions set out in the GDPR regulations. Request to access the data should be directed to the corr.authors

Acknowledgments

We thank Karin Helmersen for her help with the Sanger sequencing and DNA extraction, Milan Stosic for helpful discussions and assistance with uncooperative Genbank files, and Sinan U. Umu for bioinformatic advise. The sequencing service was provided by the Norwegian Sequencing Centre (<https://www.sequencing.uio.no>), a national technology platform hosted by Oslo University Hospital and the University of Oslo supported by the Research Council of Norway and the South-Eastern Regional Health Authority.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.tvr.2022.200247>.

References

- [1] A. Harari, Z. Chen, R.D. Burk, Human papillomavirus genomics: past, present and future, *Curr. Probl. Dermatol. (Basel)* 45 (2014) 1–18, <https://doi.org/10.1159/000355952>.
- [2] D. DiMaio, L.M. Petti, The E5 proteins, *Virology* 445 (1–2) (2013) 99–114, <https://doi.org/10.1016/j.virol.2013.05.006>.
- [3] K. Hoppe-Seyler, F. Bossler, J.A. Braun, A.L. Herrmann, F. Hoppe-Seyler, The HPV E6/E7 oncogenes: key factors for viral carcinogenesis and therapeutic targets, *Trends Microbiol.* 26 (2) (2018), <https://doi.org/10.1016/j.tim.2017.07.007>.
- [4] Human Reference clones – hpvcenter. https://www.hpvcenter.se/human_reference_clones/. (Accessed 5 March 2020).
- [5] D. Bzhalava, C. Eklund, J. Dillner, International standardization and classification of human papillomavirus types, *Virology* 476 (2015) 341–344, <https://doi.org/10.1016/j.virol.2014.12.028>.
- [6] PaVe, Papilloma virus genome database. https://pave.niaid.nih.gov/#explore/reference_genomes/human_genomes. (Accessed 6 February 2020).
- [7] Z. Chen, M. Schiffman, R. Herrero, et al., Classification and evolution of human papillomavirus genome variants: alpha-5 (HPV26, 51, 69, 82), alpha-6 (HPV30, 53, 56, 66), alpha-11 (HPV34, 73), alpha-13 (HPV54) and alpha-3 (HPV61), *Virology* 516 (2018) 86–101, <https://doi.org/10.1016/j.virol.2018.01.002>.
- [8] R.D. Burk, A. Harari, Z. Chen, Human papillomavirus genome variants, *Virology* 445 (1–2) (2013) 232–243, <https://doi.org/10.1016/j.virol.2013.07.018>.
- [9] H.U. Bernard, R.D. Burk, Z. Chen, K. Van Doorslaer, H zur Hausen, E.M. de Villiers, Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments, *Virology* 401 (1) (2010) 70–79, <https://doi.org/10.1016/j.virol.2010.02.002>.
- [10] C. De Martel, M. Plummer, J. Vignat, S. Franceschi, Worldwide burden of cancer attributable to HPV by site, country and HPV type, *Int Agency Res Cancer (IARC/WHO)* 141 (2017) 664–670, <https://doi.org/10.1002/ijc.30716>.
- [11] F.X. Bosch, A. Lorincz, N. Muñoz, L.M. Meijer, The Causal Relation between Human Papillomavirus and Cervical Cancer, *vol.* 55, 2002. <http://www.ipvsoc.org>. (Accessed 2 February 2019).
- [12] H. Zur Hausen, Papillomaviruses and cancer: from basic studies to clinical application, *Nat. Rev. Cancer* 2 (5) (2002) 342–350, <https://doi.org/10.1038/nrc798>.
- [13] R.D. Burk, Z. Chen, K. Van Doorslaer, Human papillomaviruses: genetic basis of carcinogenicity, *Public Health Genomics* 12 (5–6) (2009) 281–290, <https://doi.org/10.1159/000214919>.
- [14] K. Van Doorslaer, Evolution of the papillomaviridae, *Virology* 445 (1–2) (2013) 11–20, <https://doi.org/10.1016/j.virol.2013.05.012>.
- [15] CM de Oliveira, I.G. Bravo, NCS.e. Souza, et al., High-level of viral genomic diversity in cervical cancers: a Brazilian study on human papillomavirus type 16, *Infect. Genet. Evol.* 34 (2015) 44–51, <https://doi.org/10.1016/j.meegid.2015.07.002>.
- [16] L. Mirabello, M. Yeager, K. Yu, et al., HPV16 E7 genetic conservation is critical to carcinogenesis, *Cell* 170 (6) (2017) 1164–1174, <https://doi.org/10.1016/j.cell.2017.08.001>, e6.
- [17] Y. Hirose, M. Onuki, Y. Tenjimbayashi, et al., Within-host variations of human papillomavirus reveal APOBEC signature mutagenesis in the viral genome, *J. Virol.* 92 (12) (2018) e00017–e00018, <https://doi.org/10.1128/jvi.00017-18>.
- [18] R.S. Dube Mandishora, K.S. Gjøtterud, S. Lagström, et al., Intra-host sequence variability in human papillomavirus, *Papillomavirus Res* 5 (2018) 180–191, <https://doi.org/10.1016/j.pvr.2018.04.006>.
- [19] S. Lagström, P. van der Weele, T.B. Rounge, I.K. Christiansen, A.J. King, O. H. Ambur, HPV16 whole genome minority variants in persistent infections from young Dutch women, *J. Clin. Virol.* (August 2019), <https://doi.org/10.1016/J.JCV.2019.08.003>.
- [20] B. Zhu, Y. Xiao, M. Yeager, et al., Mutations in the HPV16 genome induced by APOBEC3 are associated with viral clearance, *Nat. Commun.* 11 (1) (2020) 886, <https://doi.org/10.1038/s41467-020-14730-1>.
- [21] G.M. Clifford, V. Tenet, D. Georges, et al., Human papillomavirus 16 sub-lineage dispersal and cervical cancer risk worldwide: whole viral genome sequences from 7116 HPV16-positive women, *Papillomavirus Res* 7 (2019) 67–74, <https://doi.org/10.1016/j.pvr.2019.02.001>.
- [22] L.S. Arroyo-Mühr, C. Lagheden, E. Hultin, et al., The HPV16 genome is stable in women who progress to in situ or invasive cervical cancer: a prospective population-based study, *Cancer Res.* 79 (17) (2019) 4532–4538, <https://doi.org/10.1158/0008-5472.CAN-18-3933>.
- [23] N.A. Wallace, K. Mürger, The curious case of APOBEC3 activation by cancer-associated human papillomaviruses, *PLoS Pathog.* 14 (1) (2018), e1006717, <https://doi.org/10.1371/journal.ppat.1006717>.
- [24] C.J. Warren, T. Xu, K. Guo, et al., APOBEC3A functions as a restriction factor of human papillomavirus, *J. Virol.* 89 (1) (2015) 688–702, <https://doi.org/10.1128/jvi.02383-14>.
- [25] J.R. Chapman, M.R.G. Taylor, S.J. Boulton, Playing the end game: DNA double-strand break repair pathway choice, *Mol. Cell.* 47 (4) (2012) 497–510, <https://doi.org/10.1016/j.molcel.2012.07.029>.
- [26] S Do Kang, S. Chatterjee, S. Alam, et al., Effect of productive human papillomavirus 16 infection on global gene expression in cervical epithelium, *J. Virol.* 92 (20) (2018), <https://doi.org/10.1128/JVI.01261-18>.
- [27] C.C. Spriggs, L.A. Laimins, Human Papillomavirus and the DNA Damage Response: Exploiting Host Repair Pathways for Viral Replication, 2017, <https://doi.org/10.3390/v9080232>.

- [28] A.A. McBride, A. Warburton, The role of integration in oncogenic progression of HPV-associated cancers, *PLoS Pathog.* 13 (4) (2017), e1006211, <https://doi.org/10.1371/journal.ppat.1006211>.
- [29] M. Pett, N. Coleman, Integration of high-risk human papillomavirus: a key event in cervical carcinogenesis? *J. Pathol.* 212 (4) (2007) 356–367, <https://doi.org/10.1002/path.2192>.
- [30] R.D. Burk, Z. Chen, C. Saller, et al., Integrated genomic and molecular characterization of cervical cancer, *Nature* 543 (7645) (2017) 378–384, <https://doi.org/10.1038/nature21386>.
- [31] S. Jeon, B.L. Allen-Hoffmann, P.F. Lambert, Integration of human papillomavirus type 16 into the human genome correlates with a selective growth advantage of cells, *J. Virol.* 69 (5) (1995) 2989–2997, <https://doi.org/10.1128/jvi.69.5.2989-2997.1995>.
- [32] C. Ziegert, N. Wentzensen, S. Vinokurova, et al., A comprehensive analysis of HPV integration loci in anogenital lesions combining transcript and genome-based amplification techniques, *Oncogene* 22 (25) (2003) 3977–3984, <https://doi.org/10.1038/sj.onc.1206629>.
- [33] C. Bodelon, M.E. Untereiner, M.J. Machiela, S. Vinokurova, N. Wentzensen, Genomic characterization of viral integration sites in HPV-related cancers, *Int. J. Cancer* 139 (9) (2016) 2001–2011, <https://doi.org/10.1002/ijc.30243>.
- [34] M. Dürst, C.M. Croce, L. Gissmann, et al., Papillomavirus sequences integrate near cellular oncogenes in some cervical carcinomas, *Proc. Natl. Acad. Sci. U. S. A.* 84 (4) (1987) 1070–1074, <https://doi.org/10.1073/pnas.84.4.1070>.
- [35] A.I. Ojesina, L. Lichtenstein, S.S. Freeman, et al., Landscape of genomic alterations in cervical carcinomas, *Nature* 506 (7488) (2014) 371–375, <https://doi.org/10.1038/nature12881>.
- [36] M. Peter, N. Stransky, J. Couturier, et al., Frequent genomic structural alterations at HPV insertion sites in cervical carcinoma, *J. Pathol.* 221 (3) (2010) 320–330, <https://doi.org/10.1002/path.2713>.
- [37] K. Akagi, J. Li, T.R. Broutian, et al., Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability, *Genome Res.* 24 (2) (2014) 185–199, <https://doi.org/10.1101/gr.164806.113>.
- [38] A. Warburton, C.J. Redmond, K.E. Dooley, et al., HPV integration hijacks and multimerizes a cellular enhancer to generate a viral-cellular super-enhancer that drives high viral oncogene expression, in: M. Ott (Ed.), *PLoS Genet.*, vol. 14, 2018, e1007179, <https://doi.org/10.1371/journal.pgen.1007179>, 1.
- [39] G. Gao, J. Wang, J.L. Kasperbauer, et al., Whole genome sequencing reveals complexity in both HPV sequences present and HPV integrations in HPV-positive oropharyngeal squamous cell carcinomas, *BMC Cancer* 19 (1) (2019) 352, <https://doi.org/10.1186/s12885-019-5536-1>.
- [40] I. Kraus, C. Driesch, S. Vinokurova, et al., The majority of viral-cellular fusion transcripts in cervical carcinomas cotranscribe cellular sequences of known or predicted genes, *Cancer Res.* 68 (7) (2008) 2514–2536, <https://doi.org/10.1158/0008-5472.CAN-07-2776>.
- [41] A. Warburton, T.E. Markowitz, J.P. Katz, J.M. Pipas, A.A. McBride, Recurrent integration of human papillomavirus genomes at transcriptional regulatory hubs, *2021* 61, *npj Genomic Med* 6 (1) (2021) 1–15, <https://doi.org/10.1038/s41525-021-00264-y>.
- [42] E.J. Crosbie, M.H. Einstein, S. Franceschi, H.C. Kitchener, Human papillomavirus and cervical cancer, *Lancet* 382 (9895) (2013) 889–899, [https://doi.org/10.1016/S0140-6736\(13\)60022-7](https://doi.org/10.1016/S0140-6736(13)60022-7).
- [43] S. Lagström, A.H. Løvestad, S.U. Umu, et al., HPV16 and HPV18 type-specific APOBEC3 and integration profiles in different diagnostic categories of cervical samples, *Tumour Virus Res* (2021) 12, <https://doi.org/10.1016/J.TVR.2021.200221>.
- [44] R.D. Burk, Z. Chen, C. Saller, et al., Integrated genomic and molecular characterization of cervical cancer, *Nature* 543 (7645) (2017) 378–384, <https://doi.org/10.1038/nature21386>.
- [45] S. Vinokurova, N. Wentzensen, I. Kraus, et al., Type-dependent integration frequency of human papillomavirus genomes in cervical lesions, *Cancer Res.* 68 (1) (2008) 307–313, <https://doi.org/10.1158/0008-5472.CAN-07-2754>.
- [46] S. Lagström, S.U. Umu, M. Lepistö, et al., TaME-seq: an efficient sequencing approach for characterisation of HPV genomic variability and chromosomal integration, *Sci. Rep.* 9 (1) (2019) 524, <https://doi.org/10.1038/s41598-018-36669-6>.
- [47] A. Tropé, K. Sjøborg, A. Eskild, et al., Performance of human papillomavirus DNA and mRNA testing strategies for women with and without cervical neoplasia, *J. Clin. Microbiol.* 47 (8) (2009) 2458–2464, <https://doi.org/10.1128/JCM.01863-08>.
- [48] A. Tropé, K.D. Sjøborg, M. Nygård, et al., Cytology and human papillomavirus testing 6 to 12 months after ASCUS or LSIL cytology in organized screening to predict high-grade cervical neoplasia between screening rounds, *J. Clin. Microbiol.* 50 (6) (2012) 1927–1935, <https://doi.org/10.1128/JCM.00265-12>.
- [49] J.J. Kozich, S.L. Westcott, N.T. Baxter, S.K. Highlander, P.D. Schloss, Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform, *Appl. Environ. Microbiol.* 79 (17) (2013) 5112–5120, <https://doi.org/10.1128/AEM.01043-13>.
- [50] D. Kim, B. Langmead, S.L. Salzberg, HISAT: a fast spliced aligner with low memory requirements, *Nat. Methods* 12 (4) (2015) 357–360, <https://doi.org/10.1038/nmeth.3317>.
- [51] K. Van Doorslaer, Q. Tan, S. Xirasagar, et al., The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis, *Nucleic Acids Res.* 41 (Database issue) (2013) D571–D578, <https://doi.org/10.1093/nar/gks984>.
- [52] S.M. Kiehbasa, R. Wan, K. Sato, P. Horton, M.C. Frith, Adaptive seeds tame genomic sequence comparison, *Genome Res.* 21 (3) (2011) 487–493, <https://doi.org/10.1101/gr.113985.110>.
- [53] A. Untergasser, I. Cutcutache, T. Koressaar, et al., Primer3–new capabilities and interfaces, *Nucleic Acids Res.* 40 (15) (2012), <https://doi.org/10.1093/NAR/GKS596>.
- [54] T. Hothorn, F. Bretz, P. Westfall, Simultaneous inference in general parametric models, *Biom. J.* 50 (3) (2008) 346–363, <https://doi.org/10.1002/BIMJ.200810425>.
- [55] A. Repesa, HPV Chromosomal Integration as a Biomarker for Cancer Progression, Master's thesis, OsloMet-Storbyuniversitetet., 2021 (May), <https://hdl.handle.net/11250/2836505>.
- [56] M. Rusan, Y.Y. Li, P.S. Hammerman, Genomic landscape of human papillomavirus-associated cancers, *Clin. Cancer Res.* 21 (9) (2015) 2009–2019, <https://doi.org/10.1158/1078-0432.CCR-14-1101>.
- [57] M. Cullen, J.F. Boland, M. Schiffman, et al., Deep sequencing of HPV16 genomes: a new high-throughput tool for exploring the carcinogenicity and natural history of HPV16 infection, *Papillomavirus Res* (2015;1(June) 3–11, <https://doi.org/10.1016/j.pvr.2015.05.004>.
- [58] D.L. Faden, K.A.L. Kuhs, M. Lin, et al., APOBEC mutagenesis is concordant between tumor and viral genomes in HPV-positive head and neck squamous cell carcinoma, 2021, Vol 13, Page 1666, *Viruses* 13 (8) (2021) 1666, <https://doi.org/10.3390/V13081666>.
- [59] C.J. Warren, J.A. Westrich, K. Van Doorslaer, D. Pyeon, Roles of APOBEC3A and APOBEC3B in human papillomavirus infection and disease progression, *Viruses* 9 (8) (2017), <https://doi.org/10.3390/v9080233>.
- [60] K. Chatfield-Reed, S. Gui, W.Q. O'Neill, T.N. Teknos, Q. Pan, HPV33+ HNSCC is associated with poor prognosis and has unique genomic and immunologic landscapes, *Oral Oncol.* 100 (2020), 104488, <https://doi.org/10.1016/J.ORALONCOLOGY.2019.104488>.
- [61] A. Chakravarthy, I. Reddin, S. Henderson, et al., Integrated analysis of cervical squamous cell carcinoma cohorts from three continents reveals conserved subtypes of prognostic significance, 2020.04.02, *bioRxiv*. December (2021), 019711, <https://doi.org/10.1101/2020.04.02.019711>.
- [62] A. Willemsen, I.G. Bravo, Origin and evolution of papillomavirus (onco)genes and genomes, *Philos Trans R Soc B Biol Sci* 374 (1773) (2019), <https://doi.org/10.1098/rstb.2018.0303>.
- [63] N. Egawa, Q. Wang, H.M. Griffin, et al., HPV16 and 18 genome amplification show different E4-dependence, with 16E4 enhancing E1 nuclear accumulation and replicative efficiency via its cell cycle arrest and kinase activation functions, *PLoS Pathog.* 13 (3) (2017), e1006282, <https://doi.org/10.1371/journal.ppat.1006282>.
- [64] S. de Sanjose, W.G.V. Quint, L. Alemany, et al., Human papillomavirus genotype attribution in invasive cervical cancer: a retrospective cross-sectional worldwide study, *Lancet Oncol.* 11 (11) (2010) 1048–1056, [https://doi.org/10.1016/S1470-2045\(10\)70230-8](https://doi.org/10.1016/S1470-2045(10)70230-8).
- [65] M. Demarco, N. Hyun, O. Carter-Pokras, et al., A study of type-specific HPV natural history and implications for contemporary cervical cancer screening programs, *EClinicalMedicine* 22 (2020), 100293, <https://doi.org/10.1016/J.ECLINM.2020.100293/ATTACHMENT/A07C2470-AA58-47A1-9593-962D28F8F692/MMC1.DOCX>.
- [66] P. Guan, R. Howell-Jones, N. Li, et al., Human papillomavirus types in 115,789 HPV-positive women: a meta-analysis from cervical infection to cancer, *Int. J. Cancer* 131 (10) (2012) 2349–2359, <https://doi.org/10.1002/IJC.27485>.
- [67] F. Poulain, N. Lejeune, K. Willemart, N.A. Gillet, Footprint of the host restriction factors APOBEC3 on the genome of human viruses, *PLoS Pathog.* 16 (8) (2020), e1008718, <https://doi.org/10.1371/JOURNAL.PPAT.1008718>.
- [68] Z. Hu, D. Zhu, W. Wang, et al., Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism, *Nat. Genet.* 47 (2) (2015) 158–163, <https://doi.org/10.1038/ng.3178>.
- [69] R. Zhang, C. Shen, L. Zhao, et al., Dysregulation of host cellular genes targeted by human papillomavirus (HPV) integration contributes to HPV-related cervical carcinogenesis, *Int. J. Cancer* 138 (5) (2016) 1163–1174, <https://doi.org/10.1002/ijc.29872>.
- [70] M. Dürst, C.M. Croce, L. Gissmann, E. Schwarz, K. Huebner, *Papillomavirus Sequences Integrate Near Cellular Oncogenes in Some Cervical Carcinomas (Viral DNA Integration/c-Myc/Genital Cancer)*, vol. 84, 1987.
- [71] M. Dong, Z. Dong, X. Zhu, Y. Zhang, L. Song, Long non-coding RNA MIR205HG regulates KRT17 and tumor processes in cervical cancer via interaction with SRSF1, *Exp. Mol. Pathol.* 111 (2019), <https://doi.org/10.1016/J.YEXMP.2019.104322>.
- [72] W. Xu, W. Yang, C. Wu, X. Ma, H. Li, J. Zheng, Enolase 1 correlated with cancer progression and immune-infiltrating in multiple cancer types: a pan-cancer analysis, *Front. Oncol.* 10 (2021) 3391, <https://doi.org/10.3389/FONC.2020.593706/BIBTEX>.
- [73] C.L. Lin, T.H. Ying, S.F. Yang, et al., MTA2 silencing attenuates the metastatic potential of cervical cancer cells by inhibiting AP1-mediated MMP12 expression via the ASK1/MEK3/p38/YB1 axis, 2021 125, *Cell Death Dis.* 12 (5) (2021) 1–12, <https://doi.org/10.1038/s41419-021-03729-1>.
- [74] E.M. Siegel, S. Eschrich, K. Winter, et al., Epigenomic characterization of locally advanced anal cancer: an RTOG 98-11 specimen study, *Dis. Colon Rectum* 57 (8) (2014) 941, <https://doi.org/10.1097/DCR.000000000000160>.
- [75] H. Niitsu, T. Hinoi, Y. Kawaguchi, et al., KRAS mutation leads to decreased expression of regulator of calcineurin 2, resulting in tumor proliferation in colorectal cancer, 2016 58, *Oncogene* 5 (8) (2016) e253, <https://doi.org/10.1038/oncsis.2016.47>, e253.
- [76] Y. Li, H. Wang, H. Huang, Long non-coding RNA MIR205HG function as a ceRNA to accelerate tumor growth and progression via sponging miR-122-5p in cervical

- cancer, *Biochem. Biophys. Res. Commun.* 514 (1) (2019) 78–85, <https://doi.org/10.1016/J.BBRC.2019.04.102>.
- [77] L. Yin, Y. Zhang, L. Zheng, Analysis of differentially expressed long non-coding RNAs revealed a pro-tumor role of MIR205HG in cervical cancer, *Mol. Med. Rep.* 25 (2) (2022) 1–8, <https://doi.org/10.3892/MMR.2021.12558/HTML>.
- [78] Y. Zheng, X. Li, Y. Jiao, C. Wu, High-risk human papillomavirus oncogenic E6/E7 mRNAs splicing regulation, *Front. Cell. Infect. Microbiol.* (2022) 790, <https://doi.org/10.3389/FCIMB.2022.929666>, 0.
- [79] T. Mesplède, D. Gagnon, F. Bergeron-Labrecque, et al., p53 degradation activity, expression, and subcellular localization of E6 proteins from 29 human papillomavirus genotypes, *J. Virol.* 86 (1) (2012) 94–107, <https://doi.org/10.1128/jvi.00751-11>.
- [80] J.E. Leeman, Y. Li, A. Bell, et al., Human papillomavirus 16 promotes microhomology-mediated end-joining, *Proc. Natl. Acad. Sci. U. S. A.* 116 (43) (2019) 21573–21579, <https://doi.org/10.1073/PNAS.1906120116/-/DCSUPPLEMENTAL>.
- [81] N. Sakakibara, R. Mitra, A.A. McBride, The papillomavirus E1 helicase activates a cellular DNA damage response in viral replication foci, *J. Virol.* 85 (17) (2011) 8981–8995, <https://doi.org/10.1128/JVI.00541-11/ASSET/A7D40B6F-A2F2-49CA-A9B3-AAE6F2F9516B/ASSETS/GRAPHIC/ZJV9990949670009.JPEG>.
- [82] V.M. Williams, M. Filippova, V. Filippov, K.J. Payne, P. Duerksen-Hughes, Human papillomavirus type 16 E6* induces oxidative stress and DNA damage, *J. Virol.* 88 (12) (2014) 6751–6761, <https://doi.org/10.1128/JVI.03355-13>.
- [83] R. Senapati, N.N. Senapati, B. Dwibedi, Molecular mechanisms of HPV mediated neoplastic progression, *Infect. Agents Cancer* 11 (1) (2016), <https://doi.org/10.1186/S13027-016-0107-4>.
- [84] S. Duensing, K. Münger, The human papillomavirus type 16 E6 and E7 oncoproteins independently induce numerical and structural chromosome instability, *Cancer Res.* 62 (23) (2002) 7075–7082.

Paper III

Hesselberg Løvestad A, Jørgensen SB, Handal N, Ambur OH, Aamot HV. Investigation of intra-hospital SARS-CoV-2 transmission using nanopore wholegenome sequencing. *Journal of Hospital Infection* 2021;111:107–16. DOI: <https://doi.org/10.1016/J.JHIN.2021.02.022>



Investigation of intra-hospital SARS-CoV-2 transmission using nanopore whole-genome sequencing

A.H. Løvestad^{a,b}, S.B. Jørgensen^b, N. Handal^b, O.H. Ambur^{a,1},
H.V. Aamot^{b,c,*}

^a Faculty of Health Sciences, Oslo Metropolitan University, Oslo, Norway

^b Department of Microbiology and Infection Control, Akershus University Hospital, Lørenskog, Norway

^c Department of Clinical Molecular Biology (Epigen), Akershus University Hospital and University of Oslo, Lørenskog, Norway

ARTICLE INFO

Article history:

Received 6 January 2021

Accepted 19 February 2021

Available online 26 February 2021

Keywords:

SARS-CoV-2

COVID-19

Transmission

Healthcare workers

Hospital

Whole-genome sequencing



SUMMARY

Background: During the SARS-CoV-2 pandemic, healthcare workers (HCWs) are being exposed to infection both at work and in their communities. Determining where HCWs might have been infected is challenging based on epidemiological data alone. At Akershus University Hospital, Norway, several clusters of possible intra-hospital SARS-CoV-2 transmission were identified based on routine contact tracing.

Aim: To determine whether clusters of suspected intra-hospital SARS-CoV-2 transmission could be resolved by combining whole genome sequencing (WGS) of SARS-CoV-2 with contact tracing data.

Methods: Epidemiological data were collected during routine contact tracing of polymerase chain reaction-confirmed SARS-CoV-2-positive HCWs. Possible outbreaks were identified as wards with two or more infected HCWs defined as close contacts who tested positive for SARS-CoV-2 less than three weeks apart. Viral RNA from naso-/oropharyngeal samples underwent nanopore sequencing in direct compliance to the ARTIC Network protocol.

Findings: Five outbreaks were suspected from contact tracing. Viral consensus sequences from 24 HCWs, two patients, and seven anonymous samples were analysed. Two outbreaks were confirmed, one refuted, and two remained undetermined. One new potential outbreak was discovered.

Conclusion: Combined with epidemiological data, nanopore WGS was a useful tool for investigating intra-hospital SARS-CoV-2 transmission. WGS helped to resolve questions about possible outbreaks and to guide local infection prevention and control measures.

© 2021 The Healthcare Infection Society. Published by Elsevier Ltd. All rights reserved.

Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic has created a heavy strain on healthcare workers (HCWs) treating COVID-19 patients. In addition to the risk of burnout and psychological distress reported, HCWs may also be at risk of infection at work [1,2]. Furthermore, HCWs may also be a source of infection for

* Corresponding author. Address: Akershus University Hospital, Department of Microbiology and Infection Control, Box 1000, Lørenskog, 1478, Norway. Tel.: +47 93850682.

E-mail address: hege.vangstein.aamot@ahus.no (H.V. Aamot).

¹ These authors contributed equally.

patients and colleagues via asymptomatic carriage and transmissibility prior to the onset of symptoms. Due to the high mortality rate of COVID-19 among the elderly, a particularly difficult challenge has been to avoid virus entry to nursing homes and hospitals. Studies describing the risk and events of SARS-CoV-2 intra-hospital transmission are discrepant [3–8]. During an epidemic, when there is frequent viral transmission in the community, it is not always clear whether HCWs are infected at work or during their spare time. As the pandemic is constantly evolving, new awareness towards specific variants has soared from fear of strains more transmissible, pathogenic and likely to evade immunization efforts. As outbreak definitions vary and outbreak reports mainly depend on epidemiological data with SARS-CoV-2 test results, the true transmission patterns remain uncertain [9]. An aggregation of infected HCWs in a ward over some days or weeks does not necessarily imply intra-hospital transmission or a local outbreak.

High-throughput sequencing technology enables the investigation of microbial outbreaks and transmissions at high resolution, including those of SARS-CoV-2. With an aim to reduce time from sampling to interpretable epidemiological results in viral outbreaks, the ARTIC network was established in the UK and is now a global effort having partnered with the World Health Organization and other public health bodies worldwide (<https://artic.network/ncov-2019>). Through the employment of portable sequencing instruments and rigging an online integrative analysis platform, the protocols, primers, and bioinformatics tools devised by the ARTIC network allow for real-time epidemiology of the SARS-CoV-2 outbreak. Yet, only a few studies have been published in which whole-genome sequencing (WGS) has been combined with epidemiological data to trace possible transmission chains in healthcare settings [10–15].

In this cross-sectional study, the aim was to employ the ARTIC network protocol and to combine resulting SARS-CoV-2 whole-genome sequence data with contact tracing data to determine whether clusters of suspected intra-hospital SARS-CoV-2 transmission could be resolved.

Methods

Contact tracing and epidemiological data

Akershus University Hospital is a secondary emergency care hospital in Norway. It serves 640,000 people (12% of Norway's population) with approximately 1000 beds and 10,000 employees. Between March 5th, 2020 and July 1st, 2020, a total of 200 COVID-19 patients had been admitted to the hospital. The patients were treated in designated COVID-19 wards or in the intensive care unit in single or double rooms, including bathrooms. HCWs caring for COVID-19 patients used personal protective equipment (PPE) in the form of gloves, gowns, goggles and surgical face masks (respiratory masks if performing aerosol-generating procedures). Other infection prevention and control measures initiated in the hospital to contain the spread of SARS-CoV-2 included testing of patients and HCWs, isolation of SARS-CoV-2-infected patients, contact tracing around all SARS-CoV-2-infected patients and employees, quarantine of close contacts, visitors restrictions, and enhanced cleaning routines. Masks or other PPE were not worn

in contact with patients or colleagues without symptoms or suspected infection.

Patients were tested for SARS-CoV-2 upon admission to the hospital if they had any respiratory, gastrointestinal, or central nervous system symptoms of infection, fatigue, or myalgia. Patients who developed any of these symptoms during their stay were also tested. Strict testing criteria were applied for HCWs in March 2020 (fever, cough, or shortness of breath), but changed during April 2020 to include any symptoms of respiratory or gastrointestinal tract infections, headaches, myalgia or fatigue. Symptomatic HCWs were tested regardless of whether they had had any contact with known SARS-CoV-2-infected individuals, either at work or in the community. Close contacts of positive cases (whether patients or HCWs) were kept in quarantine, but not routinely tested unless they developed symptoms.

The hospital's infection control staff routinely recorded epidemiological data during concurrent contact tracing of each reverse transcription–polymerase chain reaction (RT–PCR)-confirmed SARS-CoV-2-infected HCW. A close contact was defined as a person who had had physical contact with the infected HCW without use of PPE, or who had been in close proximity (<2 m) without PPE for >15 min to the infected HCW, starting from 24 h (48 h from June 2020) before the onset of symptoms. All close contacts were quarantined for 14 days (10 days since May 2020).

For the period from March 10th, 2020 to July 1st, 2020, possible outbreaks were searched for by identifying wards with two or more infected HCWs who had had close contact as previously defined, and who tested positive for SARS-CoV-2 less than three weeks apart. If we had a suspected outbreak in a ward, all isolates from HCWs in those wards were included in the study, regardless of documentation of close contact between all the cases. All the suspected outbreaks in the somatic wards occurred in wards that were designated COVID-19-wards, and where the HCWs used PPE when caring for patients. Hence, the patients were not included as close contacts, unless there were reported or suspected breaches of infection control practices.

To assess the local diversity of SARS-CoV-2, we also included viral genomes from some HCWs who had no known connection to other cases in the hospital, and who worked in different units, and some viral genomes from anonymous patients in the hospital.

The numbers of eligible and included samples are presented in [Supplementary Figure S1](#).

RNA isolation

RNA was isolated using an easyMAG extractor following the manufacturer's instructions for extraction of total nucleic acids from airways samples (bioMérieux, Marcy-l'Etoile, France). The qualitative RT–PCR detects the SARS-CoV-2 virus E-gene based on a method published by Corman *et al.* [16]. The eluate and samples of all positive RT–PCR are routinely stored at –80°C.

Library preparation and sequencing

Eluted RNA from 46 samples were reverse-transcribed and PCR-amplified using information provided by ARTIC Network (<https://artic.network/ncov-2019>). Briefly, the method uses

random hexamers for RT and multiplex PCR amplification of cDNA using a tiling amplicon scheme and the ARTIC nCoV-2019 version 3 primer set [17]. The annealing temperature of the PCR reaction was lowered to 63°C to increase amplification efficiency of problematic primer pairs. The PCR products were sequenced on a GridION sequencer (Oxford Nanopore Technologies, Oxford, UK). Five of the samples were sequenced twice to assess the reproducibility of the method.

Bioinformatic analysis

The COVID-19 bioinformatics Medaka-pipeline developed by the ARTIC network (<https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html>) was used to generate consensus sequences and call variant nucleotides relative to the reference sequence. Called variants were visualized in Geneious Prime (v2020.0.4) for validation using the BAM-files generated from the Artic pipeline. For the re-sequenced samples, the sample with the highest coverage was used for further analysis after determining the reproducibility of the method.

Phylogenetic analysis, Nextstrain clade assortment, and pangolin lineage assignment

To compare the study samples in broader context, published SARS-CoV-2 genomes were downloaded from GISAID (Supplementary Table S1) as follows: all from Norway ($N = 73$); international strains from European countries where contact tracing early in the pandemic had identified cases of SARS-CoV-2 importation to Norway ($N = 250$); and samples from China ($N = 6$) with collection dates up to 1 July 2020 [18]. A multiple sequence alignment (MSA) of the sequenced samples and downloaded SARS-CoV-2 genomes from GISAID was generated using MAFFT (v7.450) with the 1PAM scoring matrix. The MSA was then manually inspected to remove low-quality sequences. FastTree (v2.1.11) was used to generate phylogenetic trees, using GTR substitution model. The phylogenetic tree was further visualized and annotated using an in-house R-script with the ggtree package (v2.2.1) [19].

Samples were assorted to clades according to the Nextstrain nomenclature [20]. Clade assortment was carried out using a combination of phylogenetic placement of the samples and the presence of clade-specific signature mutations. In cases where samples had no coverage in areas of the genome with signature mutations, variants could in some cases be extrapolated from the presence of co-mutations.

Pangolin lineage assignment was done using the Pangolin COVID-19 Lineage Assigner online tool [21].

Outbreak assessment

The data generated by the nanopore sequencing were used to confirm or refute whether cases of close contacts were part of the same transmission chain. Whereas many variants make up the different SARS-CoV-2 clades, study-unique variants were weighted when assessing whether cases were the result of a suspected hospital transmission chain. Study-unique variants were defined as SARS-CoV-2 variants that met the following two criteria: (i) variants that showed no local geographic distribution and (ii) with two or more co-occurring mutations not found together in any other genome in the GISAID database.

Ethical approval

The study was approved by Akershus University Hospital's Data Protection Official (2020_62). The data were recorded as part of the hospital's routine for outbreak investigations, as authorized by the institutional infection control programme and the Norwegian regulation of infection control in the healthcare service (FOR-2005-06-17-610).

Results

Identification of transmission clusters based on routine contact tracing

During the study period, 68 HCWs from 38 wards tested positive for SARS-CoV-2. Based on routine contact tracings it appeared that the majority of the HCWs had been infected abroad or had a household/close social contact with SARS-CoV-2 infection that preceded their own illness.

Data from 24 HCWs and two patients from 11 wards were analysed (Table 1), as well as seven anonymous patient samples from our hospital. Five of the wards had two or more HCWs who tested positive for SARS-CoV-2 less than three weeks apart. In one of the wards, there had been close contact between the positive HCWs. Hence, these cases had originally not been regarded as part of the same outbreak. In the four other wards, there were five different clusters of cases in which direct transmission was suspected among some of the HCWs due to close contact or work on the same shift (Table 1, outbreaks A, C, and D). In addition, there was a possible link between one HCW who reported a breach in infection control procedures during treatment of a SARS-CoV-2-infected patient (Table 1, outbreak B), and a probable link between five HCWs from two different wards who all displayed COVID-19 symptoms a few days after treating the same SARS-CoV-2-positive patient (Table 1, outbreak E). The remaining samples were singletons with no epidemiological links to other cases. In Table 1, we list the cases by date and ward, and illustrate which cases were linked by contact tracing information, and how WGS helped us refute or confirm some of the suspected outbreaks.

Sequencing results

In total, 46 samples were sequenced on the GridION. The average genome coverage for all the samples was 84.6%. However, by removing samples with coverage <80% ($N = 9$), the coverage of the analysed samples increased to 95.5%. Thirty-three samples were chosen for downstream analysis after filtering out samples with <80% coverage and replicates (Table 1 and Supplementary Figure S1).

Variants analysed

In total, 273 variants were called relative to the reference genome (MN908947.3) over 62 sites. The lowest number of variants in any sample was five (HCW19 and HCW22) and the highest was 13 (Anonymous 5). The average number of variants per sample was 8.3. The reported variants were identical for all the re-sequenced samples where they shared coverage.

Table 1
Assessment of intra-hospital SARS-CoV-2 transmission

| Case | Sample date | Ward | Outbreak as defined by contact tracing | Outbreak as defined by WGS | Nextstrain clade | Pangolin lineage | Study-unique variants | Interpretation |
|-----------|----------------------|------|--|----------------------------|------------------|------------------|--|--|
| HCW1 | Mar 29 th | 1 | A | Singleton | 20C | B.1 | C20762T | HCW1, HCW2: outbreak refuted, the two close contacts had virus from different clades. |
| HCW2 | Apr 6 th | 1 | A | Singleton | 20B | B.1.1.64 | G21724T | No close contacts included in the study, but several cases from the same ward. |
| HCW3 | Apr 13 th | 1 | No positive close contacts | Singleton | 20A/ 20268G | B.1.35 | G6419A G15438T C23481T T27384C G15380T ^a | No close contacts included in the study, but several cases from the same ward. Shares one study-unique variant with Anonymous 1. |
| HCW4 | Apr 14 th | 1 | No positive close contacts | Singleton | 20C | B.1 | | Patient 1, HCW5: outbreak confirmed, including two close contacts as hypothesized. Two study-unique variants are also shared with Anonymous samples 4. |
| Patient 1 | Apr 19 th | 1 | B | B | 20A | B.1 | G4300T ^a G7975A ^a | |
| HCW5 | Apr 27 th | 1 | B | B | 20A | B.1 | G4300T ^a G7975A ^a C23185T C29095T | |
| HCW6 | Apr 20 th | 1 | C | C‡ | 20C/24368T | B.1 | T24304C | HCW6, HCW7: outbreak cannot be refuted or confirmed. Same clade, but there are no shared study-unique variants. Two HCWs who worked together on the same shift, but with no close contact. |
| HCW7 | Apr 22 nd | 1 | C | C‡ | 20C/24368T | B.1 | G21624T | HCW8, HCW9: outbreak cannot be refuted or confirmed. |
| HCW8 | May 11 th | 1 | D | D‡ | 20A | B.1 | C21114T | Close contacts from the same ward, but with no shared study-unique variants. |
| HCW9 | May 12 th | 1 | D | D‡ | 20A | B.1 | A25442G | No close contacts included in the study. |
| HCW10 | Apr 10 th | 2 | No positive close contacts | Singleton | 20C/24368T | B.1 | T6178C | No close contacts included in the study. |
| HCW11 | Apr 14 th | 2 | No positive close contacts | Singleton | 20A | B.1 | G17347T C23895T | HCW12, HCW13: new outbreak detected by WGS in two HCWs from the same ward, but with no record of close contact. |
| HCW12 | Apr 14 th | 3 | No positive close contacts | F | 20A | B.1 | C6706T ^a | |
| HCW13 | Apr 13 th | 3 | No positive close contacts | F | 20A | B.1 | C6706T ^a | |
| Patient 2 | Jun 10 th | 4/5 | E | E | 20C/24368T | B.1 | G5036A ^a G6986A ^a | Patient 2, HCW14–18: outbreak including five HCWs and one patient confirmed as likely despite use of PPE. |
| HCW14 | Jun 19 th | 4 | E | E | 20C/24368T | B.1 | G5036A ^a G6986A ^a | |
| HCW15 | Jun 19 th | 4 | E | E | 20C/24368T | B.1 | G5036A ^a G6986A ^a | |
| HCW16 | Jun 22 th | 5 | E | E | 20C/24368T | B.1 | G5036A ^a G6986A ^a | |

Phylogenetic analysis, Nextstrain clade assortment, and pangolin lineage assignment

The results from the phylogenetic analysis and clade assortment showed that the samples mainly clustered into two large clades based on shared mutation profiles (Figure 1, Supplementary Table S1). Eleven of the samples were classified as clade 20A, with three samples clustering within the Nextstrain emerging clade 20A/20268G. Sixteen of the samples clustered within 20C and these samples clustered within two distinct groups. The largest group consisted of 12 samples that shared the G24368T mutation causing the amino acid change D936Y in the heptad repeat 1 (HR1) domain of the spike protein. The mutation profile shared between these samples has a high frequency in other Nordic countries [22]. Therefore, the name 20C/24368T is used when referring to this group to distinguish them from the rest. Furthermore, two samples were classified as clade 19A and one as 20B. No samples were classified as clade 19B.

For the pangolin lineage assignment, the samples were assigned to lineage B.1 ($N = 26$), B.1.1.64 ($N = 1$), B.1.114 ($N = 1$), B.1.35 ($N = 1$), B.1.5 ($N = 1$), B.1.5.6 ($N = 1$), and B.2 ($N = 2$) (Table I).

Resolving outbreaks by contact tracing, sampling times, and phylogenetic relationships

In total, five possible outbreaks were identified based on routine contact tracing and one additional outbreak was identified based on WGS data (Table I). Groups A, C, D, and F all consist of HCWs with close contact or simultaneous work on the same ward, while outbreaks B and E consist of samples from both HCWs and patients.

Group A

Virus from HCW1 was classified as clade 20B and virus from HCW2 as 20C. Thus, they were classified as two different genetic clades and direct transmission was ruled out.

Group B

Patient 1 and HCW5 both had viruses with two variants that were neither shared with any other virus nor found with high frequency in the GISAID database (G4300T, G7975A). However, the virus from HCW5 had two additional variants. HCW5 was tested nine days after the patient. Anonymous 4 was sampled three days after the patient and shared G4300T, G7975A

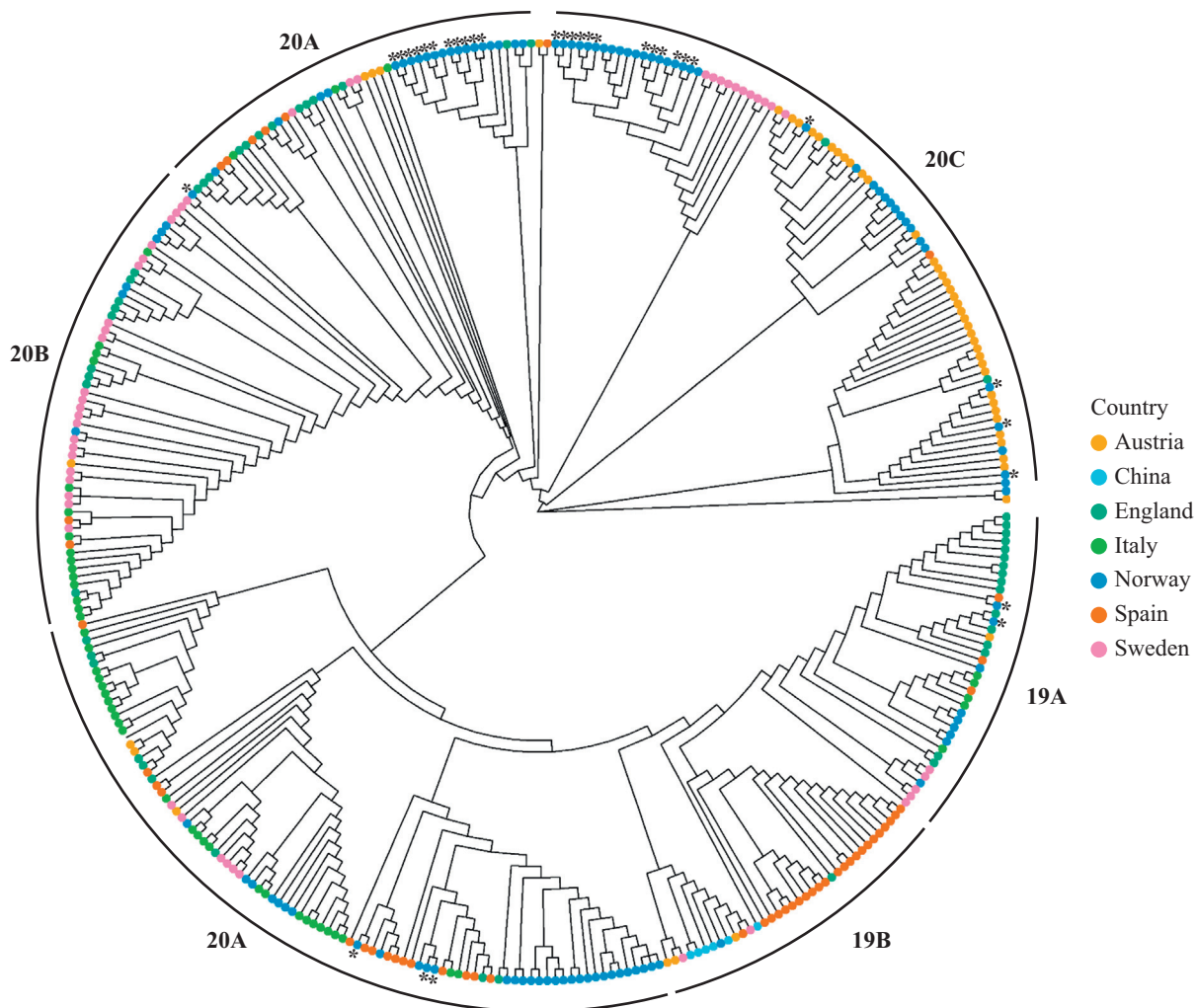


Figure 1. Phylogenetic tree of whole-genome-sequenced SARS-CoV-2 virus from Akershus University Hospital, Norway compared to all national and a selection of international viral genomes collected up until July 1st and published in the GISAID database. *Samples from this study.

without additional variants. HCW5 reported that the patient suffered from violent cough attacks, and that the PPE had felt insufficient during treatment of this patient.

Group C

Viruses from HCW6 and HCW7 were classified as 20C/24368T. However, as they both had one additional variant not shared by the other and did not share any study-unique variants, direct transmission between HCW6 and HCW7 was interpreted as uncertain.

Group D

The two viruses were classified as clade 20A. However, whereas virus from HCW8 had no additional variants, virus from HCW9 had two (C21114T, A25442G). These HCWs were close contacts and worked on the same ward for several shifts at a time when there was very low transmission activity in the community [23]. It is likely that they were linked in a transmission chain within the ward, but since the viruses did not share any unique variants this cannot be certain.

Group E

This group consists of primary case (Patient 2) and five samples from HCWs (HCW14–18) known to have interacted with them. Contact tracing indicated that HCW14–18 were all infected during the same shift. The viruses in this group shared the clade-defining G24368T variant and two study-unique variants (G5036A, G6986A). During this study a Norwegian sample was submitted to the GISAID database (Norway/2829/2020) harbouring the same three co-mutations (G5036A, G6986A, G24368T), leaving our set of variants in outbreak E not strictly study-unique according to the defined criteria. The Norway/2829/2020 sample was taken on June 29th, 2020, towards the end of this outbreak investigation, predating only the sample obtained from HCW18 (July 1st, 2020). The viruses from the patient and the two HCWs at ward 4 (HCW14, HCW15) were identical. HCW14 and HCW15 were tested on the same day and shortly after their only contact with the patient – nine days after the patient had been tested (June 10th, 2020). The viruses from the three HCWs from ward 5, where the patient was later transferred (HCW16–18), each had one or two additional non-shared variants. These samples were taken 12, 15, and 21 days after the patient's sample. The associations between contact tracing, individual sample timelines, and viral genotypes suggest a common source of infection in outbreak E. In addition, the low incidence of COVID-19 in the region at the time suggests that a common source of infection was to be found at the hospital and not in the community [23]. We elaborate on the appearance of non-shared variants in the Discussion.

Group F

HCW12 and HCW13 had no close contact according to definitions used in the contact tracing, but they worked in the same ward during the same week. Their viral samples had identical sequences and they shared the study-unique variant C6706T. Hence, they were most likely part of the same transmission cluster. This potential outbreak within the hospital would have gone undetected without the use of WGS data.

Discussion

By adding WGS of SARS-CoV-2 virus to routine contact tracing in investigations of hospital outbreaks, this study shows

both the potential power and challenges with high-resolution genotyping in local outbreak settings. Of the five suspected outbreaks, two were confirmed, two remain undetermined and one was refuted. In addition, one new possible transmission was detected, previously unidentified by routine contact tracing. Based on high-resolution genomic data, the timely implementation of SARS-CoV-2 WGS can guide local infection prevention and control measures. With the emergence of novel variants in the second and third waves of the COVID-19 pandemic with feared capabilities, the importance of swiftly obtaining high-resolution genomic SARS-CoV-2 data cannot be overstated. Different protective measures from PPE and personal behaviour recommendations to regional and national lockdowns and curfews are now guided by case-counts prioritized with data at the virus variant level. The rapid detection of new and potentially more transmissible strains in hospitals can raise the alert and devise even higher safety measures including HCW routines and staff rotation.

So far, data from the GISAID database has been useful for detecting potential structural changes in the virus, monitoring large-scale transmission dynamics, potential antigenic drift and SARS-CoV-2 evolution [24–27]. However, until now, there have been few attempts to use WGS in real-time outbreak investigations. In a retrospective cross-sectional Dutch study, genomes from three different hospitals were compared to genomes previously entered in GISAID, allowing the researchers to conclude that nosocomial transmission was probably not a common source of infection among the HCWs studied [11]. A British prospective surveillance study found possible transmission links involving patients and symptomatic HCWs, although it was not reported whether the HCWs were index cases [10].

The confirmed outbreaks in our study contained samples that all shared study-unique variants. By emphasizing the presence of study-unique variants instead of using a pre-determined cut-off of maximum allowed differences in variants to determine intra-hospital transmissions, we lean into a more stringent confirmation criterion than other studies. This approach was chosen because SARS-CoV-2 is a novel human virus with low genetic diversity, and there were few SARS-CoV-2 genomes from Norway available online for comparison at the time of analysis [26].

With limited data on the genetic background of virus circulating in the community, and few available genomes from hospital patients, we cannot confidently conclude that all our seemingly linked cases by contact tracing were in fact intra-hospital transmissions. In the one suspected outbreak that was refuted, the samples belonged to different Next-strain major clades and pangolin lineages, with several different variants reported. These cases are the easiest to resolve using WGS data, as the number of variants that distinguish them makes the probability for linked transmission during a short timeframe infinitesimally small. Hence, this method is, for the time being, a stronger tool for refuting outbreaks than for confirming outbreaks when used on its own. When suspected outbreak genomes fall into different clades, they are not from the same intra-hospital transmission chain.

However, the real challenge is that of cases that do not share any study-unique variants, but which belong to the same clade and are genetically very similar. It is difficult to determine whether the few variants they do not share are the result

of genetic variability in the viral genomes from a common source, de-novo mutations that have developed over time within the study participants, or due to infection from different sources. There is still little research into intra-host variation and the effects of transmission bottlenecks of SARS-CoV-2, but more knowledge in this field may help us interpret outbreaks at finer resolution [28–30]. In-hospital studies such as this may be helpful in studying these anticipated effects on transmission dynamics and genetic variability, since the environment and SARS-CoV-2 infections therein are tightly monitored and controlled.

Mutations found in one case but not in others from the same outbreak may be due to mutations that arise in the new host *de novo*. The association between sampling times and new variants supports the notion that new variants are generated in HCWs during outbreak B and E in the periods between suspected transmission events and sampling. Late acquired (>8 days) samples in both confirmed outbreaks carry individual variants ($N = 7$) in a total of four HCWs not found in the primary cases (Patients 1 and 2). Community acquisition of these unique variants is considered highly unlikely for HCWs who were under strict regimens to avoid SARS-CoV-2 infections at work and in their spare time as Norway was in lockdown (March 12th to July 15th, 2020). The SARS-CoV-2 mutation rate is estimated to result in about two mutations per month and it is possible that the consensus genome differs by one, or even two, nucleotides from one case to the next, especially if the date of sampling differs by about seven days, as is the situation in some of our cases [31]. Further studies are required to determine intra-person mutation rates and minor viral allelic diversities (i.e. signs of de-novo generation of mutations) in the context of COVID-19 disease severity and SARS-CoV-2 infectivity.

This study also discovered a potential outbreak using WGS data that would otherwise go unnoticed (outbreak F). This, again, shows the advantages of incorporating information from WGS technology to guide local infection prevention and control measures.

Oxford Nanopore sequencers have been used to investigate the global spread of SARS-CoV-2 from its origin in China and to follow transmissions between and within countries. This genomic information has been valuable in identifying local clusters of transmission and for evaluating the effect of preventive measures, as shown in studies from Iceland, China, and USA [32–34]. Several studies have used the Oxford Nanopore sequencing platform to generate whole genome sequences of SARS-CoV-2 and the technology produces highly accurate consensus-level results [35].

Regarding the reproducibility of the method, all re-sequenced samples in our set called the same variants relative to the reference genome. While nanopore sequencing has been shown to have a high per-read error rate, the strategy of generating consensus sequences from samples sequenced with enough depth overcomes this problem [36]. There were some differences in the coverage between the replicates; however, this is attributed to stochastic processes in the PCR reaction from primer performances and not the sequencing step. Using nanopore sequencing in real-time surveillance and outbreak investigation would help with better identification and demarcation of outbreaks and limit further spread by aiding the implementation of targeted measures. However, our results show that the analysis is dependent on samples with C_T

value <33 for consistent amplification efficiency and consequently high genome coverage (Supplementary Figure S2). Due to the low start-up cost, portability, in addition to the short time from sampling to interpretable and actionable results, nanopore sequencing is also well suited for 'lab-in-a-suitcase' initiatives where sequencing core facilities are missing.

Because we did not have the resources to sequence the viral genomes from all the patients who had been cared for by the infected HCWs, our study could not be used to investigate possible transmissions between HCWs with PPE and their patients in general. However, our sample includes one outbreak (outbreak E) in which the HCWs all wore PPE as recommended by the Norwegian Institute for Public Health and where no breach in infection control procedures was reported. The patient was transferred from a regular ward where the staff used surgical masks, eye protection, coats, and gloves to an intensive care unit where the staff wore the same equipment but with FFP2 or FFP3 respirators instead of surgical masks. All five HCWs who cared for the patient that one night were infected regardless of which mask was used. Hence, this is an example of a super-spreader event where a single person infected several other individuals within only a few hours. The patient had severe cough and respiratory failure and was treated with an oxygen mask with a flow of 12 L/min before transfer to the ICU.

In terms of patient safety and for the protection of HCWs, it is important to monitor and examine any possible SARS-CoV-2 outbreaks in healthcare settings. Our results show that nanopore WGS was a useful tool for investigating intra-hospital SARS-CoV-2 transmission in combination with epidemiological data. Epidemiological tracing alone falsely identified one hospital outbreak and overlooked one outbreak. WGS can provide a better understanding of nosocomial transmission pathways and allow for necessary and timely adaptations of local infection prevention and control routines.

Acknowledgements

The authors thank T. Senthakumaran and K. Helmersen for help with collecting samples and RNA extraction. Preliminary results were presented at the online ESCMID Conference on Coronavirus Disease (ECCVID) 2020 (Abstract #0400).

Conflict of interest statement

None declared.

Funding sources

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jhin.2021.02.022>.

References

- [1] Matsuo T, Kobayashi D, Taki F, Sakamoto F, Uehara Y, Mori N, et al. Prevalence of health care worker burnout during the coronavirus disease 2019 (COVID-19) pandemic in Japan. *JAMA Netw Open* 2020;3:e2017271. <https://doi.org/10.1001/jamanetworkopen.2020.17271>.

- [2] Shechter A, Diaz F, Moise N, Anstey DE, Ye S, Agarwal S, et al. Psychological distress, coping behaviors, and preferences for support among New York healthcare workers during the COVID-19 pandemic. *Gen Hosp Psychiatry* 2020;66:1–8. <https://doi.org/10.1016/j.genhosppsych.2020.06.007>.
- [3] McMichael TM, Currie DW, Clark S, Pogojans S, Kay M, Schwartz NG, et al. Epidemiology of covid-19 in a long-term care facility in King County, Washington. *N Engl J Med* 2020;382:2008–11. <https://doi.org/10.1056/NEJMoa2005412>.
- [4] Arons MM, Hatfield KM, Reddy SC, Kimball A, James A, Jacobs JR, et al. Presymptomatic SARS-CoV-2 infections and transmission in a skilled nursing facility. *N Engl J Med* 2020;382:2081–90. <https://doi.org/10.1056/NEJMoa2008457>.
- [5] Basso T, Nordbø SA, Sundqvist E, Martinsen TC, Witsø E, Wik TS. Transmission of infection from non-isolated patients with COVID-19 to health care workers. *J Hosp Infect* 2020;106:639–42. <https://doi.org/10.1016/j.jhin.2020.08.015>.
- [6] Folgueira MD, Munoz-Ruiperez C, Alonso-Lopez MA, Delgado R. SARS-CoV-2 infection in health care workers in a large public hospital in Madrid, Spain, during March 2020. *Nat Commun* 2020;11:3500. <https://doi.org/10.1038/s41467-020-17318-x>.
- [7] Liu M, Cheng SZ, Xu KW, Yang Y, Zhu QT, Zhang H, et al. Use of personal protective equipment against coronavirus disease 2019 by healthcare professionals in Wuhan, China: cross sectional study. *BMJ* 2020;369:m2195. <https://doi.org/10.1136/bmj.m2195>.
- [8] Lemieux J, Siddle KJ, Shaw BM, Loreth C, Schaffner S, Gladden-Young A, et al. Phylogenetic analysis of SARS-CoV-2 in the Boston area highlights the role of recurrent importation and super-spreading events. *MedRxiv* 2020;2020. <https://doi.org/10.1101/2020.08.23.20178236>. 08.23.20178236.
- [9] O'Neil EA, Naumova EN. Defining outbreak: breaking out of confusion. *J Public Health Policy* 2007;28:442–55. <https://doi.org/10.1057/palgrave.jphp.3200140>.
- [10] Meredith LW, Hamilton WL, Warne B, Houldcroft CJ, Hosmillo M, Jahun AS, et al. Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *Lancet Infect Dis* 2020;20:1263–71. [https://doi.org/10.1016/S1473-3099\(20\)30562-4](https://doi.org/10.1016/S1473-3099(20)30562-4).
- [11] Sikkema RS, Pas SD, Nieuwenhuijse DF, O'Toole Á, Verweij J, van der Linden A, et al. COVID-19 in health-care workers in three hospitals in the south of the Netherlands: a cross-sectional study. *Lancet Infect Dis* 2020;20:1273–80. [https://doi.org/10.1016/S1473-3099\(20\)30527-2](https://doi.org/10.1016/S1473-3099(20)30527-2).
- [12] Paltansing S, Sikkema RS, Man SJ de, Koopmans MPG, Munnink BBO, Man P de. Transmission of SARS-CoV-2 among healthcare workers and patients in a teaching hospital in the Netherlands confirmed by whole genome sequencing. *J Hosp Infect* 2021 Feb 8. <https://doi.org/10.1016/j.jhin.2021.02.005>. S0195-6701(21)00051-7 [online ahead of print].
- [13] Olmos C, Campaña G, Monreal V, Pidal P, Sanchez N, Airola C, et al. SARS-CoV-2 infection in asymptomatic healthcare workers at a clinic in Chile. *PLoS One* 2021;16:e0245913. <https://doi.org/10.1371/journal.pone.0245913>.
- [14] Safdar N, Moreno GK, Braun KM, Friedrich TC, O'Connor DH. Using virus sequencing to determine source of SARS-CoV-2 transmission for healthcare worker. *Emerg Infect Dis* 2020;26:2489–91. <https://doi.org/10.3201/eid2610.202322>.
- [15] Lucey M, Macori G, Mullane N, Sutton-Fitzpatrick U, Gonzalez G, Coughlan S, et al. Whole-genome sequencing to track severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) transmission in nosocomial outbreaks. *Clin Infect Dis* 2020. <https://doi.org/10.1093/cid/ciaa1433>. ciaa1433 [online ahead of print].
- [16] Corman VM, Landt O, Kaiser M, Molenkamp R, Meijer A, Chu DKW, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Eurosurveillance* 2020;25. <https://doi.org/10.2807/1560-7917.ES.2020.25.3.2000045>.
- [17] Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat Protoc* 2017;12:1261–6. <https://doi.org/10.1038/nprot.2017.066>.
- [18] Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* 2017;22:30494. <https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494>.
- [19] Yu G. Using ggtree to visualize data on tree-like structures. *Curr Protoc Bioinforma* 2020;69. <https://doi.org/10.1002/cpbi.96>.
- [20] Hodcroft EB, Hadfield J, Neher RA, Bedford T. Year-letter genetic clade naming for SARS-CoV-2 on nextstrain.org. *Nextstrain* 2020;2020.
- [21] COG-UK, n.d. <https://pangolin.cog-uk.io/> [last accessed November 2020].
- [22] Ling J, Hickman RA, Li J, Lu X, Lindahl JF, Lundkvist Å, et al. Spatio-temporal mutational profile appearances of Swedish SARS-CoV-2 during the early pandemic. *Viruses* 2020;12:1026. <https://doi.org/10.3390/v12091026>.
- [23] [ukesrapport FHI. Oppsummering uke 2020;20](https://www.fhi.no/aktuelt/ukesrapport).
- [24] Jaimes JA, André NM, Chappie JS, Millet JK, Whittaker GR. Phylogenetic analysis and structural modeling of SARS-CoV-2 spike protein reveals an evolutionary distinct and proteolytically sensitive activation loop. *J Mol Biol* 2020;432:3309–25. <https://doi.org/10.1016/j.jmb.2020.04.009>.
- [25] Benvenuto D, Angeletti S, Giovanetti M, Bianchi M, Pascarella S, Cauda R, et al. Evolutionary analysis of SARS-CoV-2: how mutation of Non-Structural Protein 6 (NSP6) could affect viral autophagy. *J Infect* 2020;81:e24–7. <https://doi.org/10.1016/j.jinf.2020.03.058>.
- [26] Dearlove B, Lewitus E, Bai H, Li Y, Reeves DB, Joyce MG, et al. A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. *Proc Natl Acad Sci USA* 2020;117:23652–62. <https://doi.org/10.1073/pnas.2008281117>.
- [27] Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell* 2020;182:812–27. <https://doi.org/10.1016/j.cell.2020.06.043>. e19.
- [28] Shen Z, Xiao Y, Kang L, Ma W, Shi L, Zhang L, et al. Genomic diversity of severe acute respiratory syndrome-coronavirus 2 in patients with coronavirus disease 2019. *Clin Infect Dis* 2020;71:713–20. <https://doi.org/10.1093/cid/ciaa203>.
- [29] Wang D, Wang Y, Sun W, Zhang L, Ji J, Zhang Z, et al. Population bottlenecks and intra-host evolution during human-to-human transmission of SARS-CoV-2. *BioRxiv* 2020. <https://doi.org/10.1101/2020.06.26.173203>. 2020.06.26.173203.
- [30] Pfeffelerle S, Guenther T, Kobbe R, Czech-Sioli M, Noerz D, Santer R, et al. Low and high infection dose transmission of SARS-CoV-2 in the first COVID-19 clusters in Northern Germany. *MedRxiv* 2020 Jun 16. <https://doi.org/10.1101/2020.06.11.20127332>.
- [31] Duchene S, Featherstone L, Haritopoulou-Sinanidou M, Rambaut A, Lemey P, Baele G. Temporal signal and the phylogenetic threshold of SARS-CoV-2. *Virus Evol* 2020;6(2):veaa061. <https://doi.org/10.1093/ve/veaa061>.
- [32] Lu J, du Plessis L, Liu Z, Hill V, Kang M, Lin H, et al. Genomic epidemiology of SARS-CoV-2 in Guangdong Province, China. *Cell* 2020;181:997–1003.e9. <https://doi.org/10.1016/j.cell.2020.04.023>.
- [33] Gonzalez-Reiche A, Hernandez M, Sullivan M, Ciferri B, Alshammery H, Obla A, et al. Introductions and early spread of SARS-CoV-2 in the New York City area 2020. *Science* 2020;369(6501):297–301. <https://doi.org/10.1126/science.abc1917>.
- [34] Gudbjartsson DF, Helgason A, Jonsson H, Magnusson OT, Melsted P, Norddahl GL, et al. Spread of SARS-CoV-2 in the Icelandic population. *N Engl J Med* 2020;382:2302–15. <https://doi.org/10.1056/NEJMoa2006100>.

- [35] Bull RA, Adikari TN, Ferguson JM, Hammond JM, Stevanovski I, Beukers AG, et al. Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. *Nat Commun* 2020;11:1–8. <https://doi.org/10.1038/s41467-020-20075-6>.
- [36] Bowden R, Davies RW, Heger A, Pagnamenta AT, de Cesare M, Oikkonen LE, et al. Sequencing of human genomes with nanopore technology. *Nat Commun* 2019;10. <https://doi.org/10.1038/s41467-019-09637-5>.