UNIVERSITY OF THE
FREE STATE
UNIVERSITEIT VAN DIE
VRYSTAAT
YUNIVESITHI YA
FREISTATA

UFS
UV

Characterisation of the human papillomavirus genome and *p53* mutations in head and neck squamous cell carcinomas.

Yuri Munsamy

Characterisation of the human papillomavirus genome and *p53* mutations in head and neck squamous cell carcinomas.

Yuri Munsamy

Submitted in fulfilment of the requirements in respect of the Doctoral degree Medical virology in the Division of Virology
In the Faculty of Health Sciences
At the University of the Free State

Promotor: Prof Felicity Burt
Division of Virology, Faculty of Health Sciences,
University of the Free State, Bloemfontein

February 2019

Declaration

I, Yuri Munsamy, declare that the Doctoral Degree research thesis that I herewith submit for the Doctoral Degree qualification Medical Virology at the University of the Free State is my independent work, and that I have not previously submitted it for a qualification at another institution of higher education.

Yuri Munsamy

Yuri Munsamy

"To be a scientist is to learn to live all one's life with questions that will never be answered, with the knowledge that one was too early or too late, with the anguish of not having been able to guess at the solution that, once presented, seems so obvious that one can only curse oneself for not seeing what one ought to have, if only one looked in a slightly different direction."

Hanya Yanagihara, *The People in the Trees* (2013)

Dedication

To my late grandmothers, Athilutchmiammal Naidoo and Velliamma Munsamy for recognising the value of education to better their families.

To my parents, Loganathan and Jeevarani Munsamy, for teaching me persistence through difficulties and for their unwavering belief in my abilities.

Abstract

High-risk human papillomaviruses (HR-HPV) are ubiquitous, sexually transmitted, aetiologic agents of head and neck cancer (HNC). To date, no large-scale South African studies report on HPV type distribution and prevalence associated with head and neck cancer. In a previous study from our research group, HR-HPV was detected in biopsy samples from histologically confirmed head and neck squamous cell carcinoma (HNSCC) patients in a South African cohort. This study went on to determine genetic changes that accumulate within HR-HPV genomes, other than the well-researched HPV 16, that confer differences in oncogenicity. Unlike cervical carcinomas, it is unknown whether HPV variant research in HNSCC translates into clinical application.

The first complete genomes of HPV 18 and HPV 31 from HNC were amplified and subjected to deep sequencing analysis. VBD 17/15, the South African HPV 18 isolate, clustered in lineage A1. Mutations were identified in the E2 and long control region (LCR) that might lead to differences in oncogenicity. Evidence of how papillomaviruses evolved is shown in this study, in a phenomenon known as linkage disequilibrium. A novel mutation of the South African isolate is described and further investigations in a larger cohort will determine whether this is a single nucleotide polymorphism unique to variants that preferentially infect the head and neck region. Although geographic and ethnic associations have been described for HPV 16 and 18, this study supports the use of alphanumeric nomenclature.

Having obtained the first complete genome of HPV 31, deep sequencing analysis showed that this laryngeal carcinoma was co-infected with a closely related viral variant. HPV quasispecies has recently been described in cervical carcinoma and this is the first evidence in the head and neck region. The quasispecies described belonged to HPV 31 lineage B2. A unique deletion within the E5 gene needs to be investigated further to determine what this deletion represents to viral fitness. Polymorphisms in the LCR were investigated with pBlue-Topo® vector, a reporter gene system. Increased β-galactosidase expression was observed in the mutant that possessed a single nucleotide change within the YY1 binding site. This study provides evidence of sequence variation within HPV 31 LCR having a functional effect on viral p97 promoter activity.

HPV-HNSCC is complicated by the synergistic interaction with the host. The human tumour suppressor gene, *p53* was investigated for mutations in a subset of HNSCC samples.

As the *p53* gene is frequently mutated in most cancers, it has been proposed as a biomarker to deintensify treatment of HPV-HNSCC patients. The involvement of high-risk HPV in HNSCC is an alternative mechanism to inactivate the p53 protein function. Evidence of *p53* mutations was shown, with a predominance of substitution patterns that are induced by a carcinogen from tobacco smoke. Immunostaining of p16 as a biomarker of HPV infection did not correlate with HPV detection by PCR. It is unknown whether the genomes of participants of African descent are too diverse from the reference

genome used in this study to accurately use frequency and functional data of known mutations. All of the mutations within this study were detected within intron 5. Whether there may be mutations lying outside of the area investigated or whether other cancer driver genes are involved in tumourigenesis of this cohort of HNSCC samples is still to be determined.

Sequence data for South African isolates from patients with HNSCC adds to the global understanding of this virus-related epidemic and contributes to elucidating the underlying molecular mechanisms of HPV infection in HNC in sub-Saharan Africa, especially in light of high HPV burden in the cervix.

Keywords: Human papillomavirus, head and neck squamous cell carcinoma, HPV genomics, HPV 18, HPV 31, next-generation sequencing, genetic diversity, HPV quasispecies, *p5*3 mutations

Acknowledgements

This thesis would not have been possible without the guidance from my supervisor, Prof Felicity Burt whose knowledge and critique over the last four years has been instrumental in my development as a scientist.

I would like to acknowledge the financial assistance of the National Research Foundation and the Poliomyelitis Research Foundation. Opinions expressed and conclusions arrived at are those of the author and are not necessarily attributed to these institutions.

My gratitude to Armand Bester for answering my questions, helping me find solutions and assistance with data analysis.

Thank you to Reyalan Munsamy for help with drawing figures.

I wish to thank my friends and my siblings Reyalan and Ruella. Thank you for bringing light in times of darkness, humouring my conversations of life and all its complexities and above all else, joining me in inappropriate laughter.

List of figures

Chapter	Legend	Page number
1	Figure 1.1. Genome organisation of a high-risk human papillomavirus type, HPV 16; E1-E7 early genes, L1-L2 late genes: capsid, LCR Long control region. Figure drawn with Geneious version 2019.0 (Biomatters). Available from https://www.geneious.com using HPV 16 reference isolate (GenBank accession number NC_001526).	2
	Figure 1.2. Role of E6 and E7, p16, RB and p53 in the cell cycle pathway leading to carcinogenesis. Adapted from Hayes *et al*., 2015.	6
	Figure 1.3. The human p53 protein is composed of 393 amino acids, numbered from the amino terminus (amino acid 1) to the carboxy terminus (amino acid 393). The DNA-binding domain (102-292) is a hotspot region for mutations in most cancers. Adapted from p53 KnowledgeBase Team (Available at: http://p53.bii.a-star.edu.sg/aboutp53/index.php).	7
2	Figure 2.1. Complete genome of VBD 17/15, 7857 bp in length with 40.42% GC content, indicating the location of open reading frames E1 to E7 coding for early proteins and L1 and L2 for late proteins, and the noncoding long control region, LCR, between L1 and E6 genes. Image constructed using Geneious version 2019 (Biomatters).	22
	Figure 2.2. The maximum likelihood tree was inferred from a global alignment of 126 complete sequences of HPV 18 isolates from cervical carcinoma and VBD 17/15 isolated from HNSCC. A bootstrap value of 1000 replicates was employed. Tree constructed with MEGA 7.0. Each isolate is represented by a GenBank accession number.	24
	Figure 2.3. HPV 18 variant tree topology using complete genomes. The evolutionary history was inferred by using the Maximum Likelihood method based on the Tamura-Nei model. The bootstrap consensus tree inferred from 1000 replicates is taken to represent the evolutionary history of the taxa analysed. Evolutionary analyses were conducted in MEGA7.	25
	Figure 2.4. Illustration of percentage variability within different genes/regions across HPV 18 lineages.	27
	Figure 2.5. Genomic plot of single-nucleotide polymorphisms (SNPs) across human papillomavirus (HPV) 18 variant lineage/sublineage genomes in comparison to the reference HPV 18 genome (NC_001357). Each dot denotes a variable site, with nucleotide change depicted in the colour key box. Black bars represent deletions.	30

3	Figure 3.1. Schematic representation of the full-length LCR variants cloned upstream of the β-galactosidase gene in the reporter vector pBlue-Topo®. YY1 binding site, Yin-Yang 1 binding site; Enhancer, keratinocyte-specific enhancer domain; Ori, origin of replication of the HPV 31 circular genome.	42
	Figure 3.2. The maximum likelihood tree was inferred from a global alignment of complete sequences of HPV 31 isolates from cervical carcinoma and VBD 13/14 isolated from HNSCC. A bootstrap value of 1000 replicates was employed. Tree constructed with MEGA 7.0. Each isolate is represented by a GenBank accession number.	45
	Figure 3.3. β-galactosidase staining for determination of transfection efficiency of transfected BHK cells. BHK cells transfected with pBlueVBD13_14, pBlue_SDM1 and pBlue_SDM3 and stained for β-galactosidase. Untransfected BHK cells with slight staining that shows endogeneous β-galactosidase activity. Black arrows indicate some of the stained transfected BHK cells per field.	47
	Figure 3.4. Transcriptional activity of HPV 31 full-length LCR variants. Relative β-galactosidase activities of BHK cells transfected with reporter constructs containing different HPV 31 LCR variants. Data shown represent the means of at least three independent transfection experiments.	49
4	Figure 4.1. Location of primers across p53 gene, NC_000017.10 (hg38) on chr17:7687599...7668459. Figure drawn with Geneious version 2019.0 (Biomatters). No sequence data was obtained for Intron 7.	61
	Figure 4.2. (A) Distribution of samples according to primary site and (B) HPV status according to primary site in this cohort (C) Percentage of samples with p53 mutations (D) p53 mutation distribution according to anatomical site, considering HPV status (E) Spectra of p53 mutations observed. (F) p53 mutation distribution according to HPV status.	65-67
	Figure 4.3. Correlation between HPV status and p16 staining.	68

All figures in this thesis are original or have been adapted and redrawn from various sources.

List of tables

Chapter	Legend	Page number
2	Table 2.1. Sequences of each primer pair used to amplify HPV 18 in two overlapping fragments with predicted amplicon size.	18
	Table 2.2. HPV 18 isolates, previously identified for discrimination between lineage/sublineage-specific SNPs.	20
	Table 2.3. Annotation of genes of isolate VBD 17/15 (Geneious V7.0 (Biomatters).	21
	Table 2.4. Estimates of evolutionary divergence between sequences.	26
	Table 2.5. Percentage of variance calculated from the number of SNPs identified within each region/or gene investigated and the size of each region/or gene.	27
	Table 2.6. Variations identified in VBD 17/15 when compared with the HPV 18 reference isolate.	29
3	Table 3.1. Properties of primers to amplify HPV 31 LCR and to conduct site-directed mutagenesis on the reporter plasmid.	41
	Table 3.2. Sequence differences in the noncoding LCR relative to the prototype HPV 31.	46
4	Table 4.1. Clinical characteristics of the head and neck cancer patients, including HPV status, p16 immunostaining and exposure to alcohol and smoking.	58-59
	Table 4.2. Primers used for amplification of p53 gene from exon 4-9 and primers used for sequencing PCR amplicons. Nucleotide position relative to p53 gene, NC_000017.10 (hg38).	60
	Table 4.3. Subset of HNSCC samples that contained p53 mutations. Position relative to p53 gene extracted from chromosome 17, GenBank (NC_000017.10) (hg38).	63
	Table 4.4. Characteristics of patients and association of p53 mutation.	64

List of abbreviations

ATCC	American Type Culture Collection
BHK	Baby hamster kidney 21 cells
BLAST	Basic Local Alignment Search Tool
bp	Base pairs
°C	degrees Celcius
CIN3	Cervical intraepithelial neoplasia 3
DNA	Deoxyribonucleic acid
dATP	Dinucleotide Adenosine triphosphate
DBD	DNA binding domain
dNTP	Dinucleotide triphosphate
E6AP	E6 associated protein
h	Hours
HNSCC	Head and neck squamous cell carcinoma
HNC	Head and neck cancer
HPV	Human papillomavirus
HPV- HNSCC	HPV-associated head and neck squamous cell carcinomas
HR-HPV	High-risk human papillomavirus
IARC	International Agency for Research on Cancer
Indels	Insertions or deletions
ISH	*In situ hybridisation*
Kb	Kilobases
kDa	Kilodalton
LCR	Long control region
LD	Linkage distribution
LR-HPV	Low-risk human papillomavirus
mg	Milligram
Min	minute
ml	Millilitre
mM	Millimolar
Ng	Nanogram
NGS	Next-generation sequencing
nmoles	nanomoles
nt	nucleotide
OD	Optical density
ONPG	Ortho-Nitrophenyl-β-galactoside
OPSCC	Oropharyngeal squamous cell carcinoma
OSCC	Oesophageal squamous cell carcinoma
ORFs	Open reading frames
PCR	Polymerase chain reaction
pg	Picogram
s	seconds
SA	South Africa
SCC	Squamous cell carcinoma
SDM	Site-directed mutagenesis
SNPs	Single nucleotide polymorphisms
Ts	Transition
Tv	Transversion
µl	microlitre
X-Gal	5-bromo-4-chloro-3-indolyl-b-D-galactopyranoside
YY1	Yin-Yang 1 transcription factor binding site

Contents

CHAPTER 1 ... i

Literature review and thesis outline ... i

Introduction ... 1

The virus ... 1

Infection and replication ... 3

The carcinogenic process and tumour-suppressor genes, p16 and *p53* ... 5

Epidemiology of HPV-HNSCC ... 7

Clinical features and detection of HPV-positive HNSCCs ... 8

Prevention and treatment of HPV-related HNSCC ... 8

HPV genetic variants ... 9

Rationale, aims and thesis outline ... 11

Structure of thesis ... 13

CHAPTER 2 ... 14

Complete genome sequence and comparative analysis of human papillomavirus type 18 isolated from a nasopharyngeal carcinoma ... 14

Abstract ... 15

Keywords ... 15

Introduction ... 15

Methods ... 17

Sample ... 17

DNA extraction and PCR for detection and genotyping of sample ... 17

MiSeq library preparation and sequencing ... 18

Next-generation sequencing data analysis ... 18

Phylogenetic relationship of HPV 18 ... 18

HPV 18 variant analysis ... 19

Results ... 21

Genotyping of sample ... 21

Next-generation sequencing data ... 21

Comparative phylogenetic analysis with 125 complete HPV 18 genomes ... 22

HPV 18 variant lineages SNP analysis ... 30

Discussion ... 30

Conclusion ... 33

Data Availability ... 34

Author Contributions ... 34

CHAPTER 3 ... 35

Complete genome sequence of HPV 31 and biological consequences for p97 promoter activity ... 35

Abstract 36
Keywords 36
Introduction 36
Methods 37
Sample preparation and next generation sequencing 37
Phylogenetic Analysis and Variant Lineage/Sublineage Identification 38
Amplification of HPV 31 LCR 39
Construction of reporter plasmid 39
Site-directed mutagenesis 39
Cell culture 43
Transfection of BHK cells 43
Transfection efficiency determine by β-galactosidase staining 43
Transcriptional activity measured by β-galactosidase assay 43
Results 44
Complete genome sequence and phylogenetic analysis of HPV 31 VBD 13/14 44
Transfection efficiency determined by β-galactosidase staining 46
Transcriptional activity of HPV 31 LCR mutants 48
Discussion 49
Conclusion 51
Funding 51
Competing interests 52
Author Contributions 52
CHAPTER 4 53
The role of *p53* mutations in HPV-associated HNSCC 53
Abstract 54
Keywords 54
Introduction 54
Materials and Methods 56
Samples 56
HPV detection and p16 staining 56
p53 sequencing 56
p53 mutation analysis 57
Results 62
Clinical characteristics of head neck cancer patients 62
Mutations in the *p53* gene in the head and neck region 62
Comparison of *p53* gene status with patients' clinical characteristics 63
Correlation of p16 staining and HPV status 68

Discussion 69
Conclusion 71
Author Contributions 72
CHAPTER 5 73
Conclusions and Future Directions 73
Conclusions and Future Directions 74
REFERENCES 79
APPENDICES 89
Appendix A 89
Appendix B 91
Appendix C 92
Appendix D 93
Appendix E 94
Appendix F 95

Ethics approval

Ethics approval for conducting this study was obtained from the Ethics Committee of the Faculty of Health Sciences, University of the Free State ECUFS NR 137/2013D (ECUFS NR 137/2013) [Appendix A].

Conference outputs

- "From Cuddles to Cancer: The HPV Epidemic" Author: Y Munsamy National 3-MT Competition, UFS, SA, 2018. (Presentation).
- "Complete genome sequence and comparative analysis of human papillomavirus type 18 isolated from a head and neck cancer biopsy" Authors: Y Munsamy, RY Seedat, PA Bester, FJ Burt, Faculty of Health Sciences 50th Research Forum, UFS, SA. 2018. (Presentation).
- "HPV and head and neck cancer" Author: Y Munsamy Provincial FameLab Competition, Central University of Technology, SA, 2018. (Presentation).
- "Site-directed mutagenesis to construct human papillomavirus type 31 long control region plasmid constructs" Authors: Y Munsamy, RY Seedat, PA Bester & FJ Burt, SASM conference, Muldersdrift, SA. 2018. (Poster).
- "From Cuddles to Cancer: The HPV Epidemic" Author: Y Munsamy Provincial 3-MT Competition, UFS, SA, 2017. (Presentation).
- "Characterisation of HPV 31 complete genome associated with head and neck cancer" Authors: Y. Munsamy, R. Seedat, P. Bester & F. Burt, 31st International Papillomavirus Conference, Cape Town, SA. 2017. (Poster).
- "Characterisation of HPV 31 complete genome associated with head and neck cancer" Authors: Y Munsamy, RY Seedat, PA Bester, FJ Burt, Faculty of Health Sciences 48th Research Forum, UFS, SA. 2016. (Presentation).

CHAPTER 1

Literature review and thesis outline

Introduction

Arising in the oral cavity, nasal cavity, larynx, hypopharynx, and oropharynx, head and neck squamous cell carcinoma (HNSCC) is the sixth most common cancer worldwide.[1,2] The two most common types of HNSCC, oral squamous cell carcinoma (OSCC) and oropharyngeal squamous cell carcinoma (OPSCC) accounted for 263 900 new cases and 128 000 deaths worldwide, respectively in 2008.[3] Human papillomaviruses (HPV) have coevolved alongside human populations and are well-known oncogenic agents for cervical cancer.[4] However it is only fairly recently that a link has been established for HPV-associated head and neck squamous cell carcinomas (HPV-HNSCC).[5] High-risk HPV (HR-HPV) is responsible for about 60% of OPSCC patients in the western world.[6] Tumours in the oral cavity, larynx, or hypopharynx are less likely to be HPV-positive than oropharyngeal tumours.[7]

By 2020, HPV-HNSCC is predicted to surpass cervical cancer incidences in the United States of America.[8] There are no Pap smear equivalents for diagnosing HPV in HNSCC and no therapeutics available that directly target the viral life cycle.[9] In addition, there is a lack of consensus on the accurate proportion of HPV-driven cases, the role of host genetic cofactors and the heterogeneity of HPV prevalence in anatomical sites of the head and neck and across geographical regions.[10]

The virus

HPV is a small double-stranded circular DNA virus with a genome of approximately 8 000 base pairs (bp) that contains between eight to nine ORFs, with dual promoters. The viral DNA is encapsidated by 72 capsomers.[11] The HPV genome is divided into three genetic regions based on the positioning in the genome and timing of expression. The non-structural or early (E) genes (E1, E2, E6, and E7) are expressed in the viral infectious cycle for regulation of transcription, plasmid replication, and transformation. The late region encodes viral structural proteins involved in packaging of the viral genome and virus release. The L1 is the major capsid protein whilst L2 is the minor capsid protein.[12] The long control region (LCR) comprises about 10% of the genome and contains the promoter, viral origin of replication (ori) and enhancer elements (Figure 1).[11] The E1 viral protein is an approximately 68 kDa protein that ranges in size, from 600 to 700 amino acids (aa). It is the largest, most highly conserved viral protein involved in replication of the HPV genome.[13]

E2, a 50 kDa protein plays a supporting role in viral replication and transcriptional regulation of the viral early genes. Expressed at both the early and late stages of the viral life cycle, the E2 negatively regulates viral gene expression as it binds to the promoters of E6 and E7.[12]

The E4 protein, 17 kDa in size, is expressed in the latest phase of the viral life cycle and is presumed to have a role in viral release and assembly. Other functions may include regulation of gene expression and interaction with and destruction of the keratin cytoskeleton and induction of G2 arrest.[14]

E5, together with E6 and E7, is one of the transforming proteins of HPV. E5 is comprised of approximately 40-85 hydrophobic amino acids that are grouped into three membrane-spanning domains.[15] It is expressed late in the viral life cycle and is considered to have a weaker transforming capacity than E6 and E7 and may not entirely be necessary for transformation.[16]

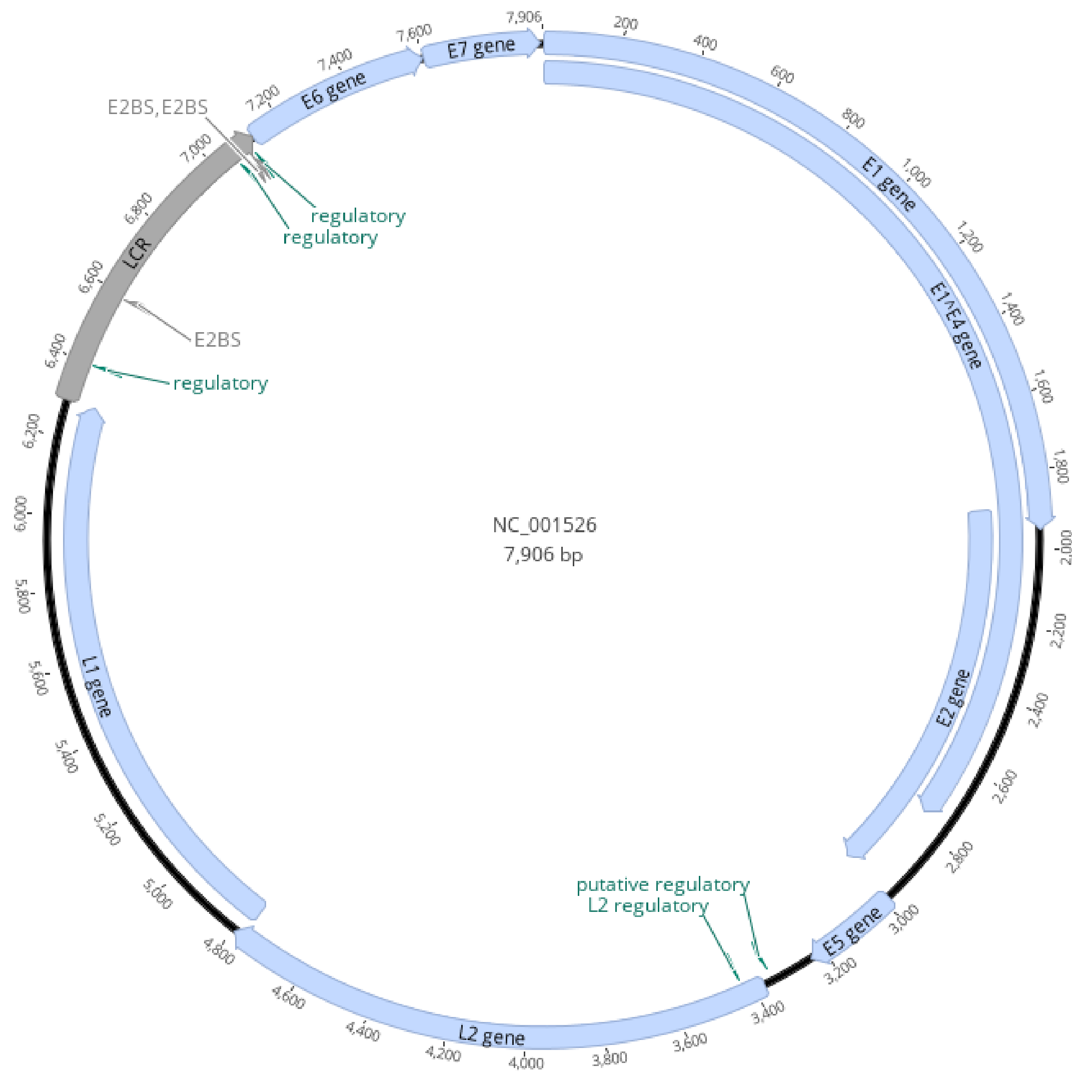


Figure 1.1. Genome organisation of a high-risk human papillomavirus type, HPV 16; E1-E7 early genes, L1-L2 late genes: capsid, LCR Long control region. Figure drawn with Geneious version 2019.0 (Biomatters). Available from https://www.geneious.com using HPV 16 reference isolate (GenBank accession number NC_001526).

HPVs are members of the family *Papillomaviridae*, genus *Papillomavirus*. There are five major HPV genera: alpha, beta, gamma, mu and nu-papillomaviruses.[17] HPV either show tropism for keratinised epithelia (cutaneotropic) or mucosal epithelia (mucosotropic). The alpha-papillomavirus genus comprises mucosal HPVs that can be further divided into HR-HPV and low risk HPV (LR-HPV),

depending on their association with cancer development.[17,18] LR-HPV types such as HPV 6 and HPV 11 infect mucosal epithelia but rarely cause cancer. However, evidence has linked these LR types with a minority of cancers, suggesting that these types are not entirely benign in the head and neck region.[19]

HR types differ in oncogenic potential and include types: 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 68.[20] Of these, HPV 16 is the most pathogenic and medically relevant type, associated with more than 80% of HPV-HNSCC.[10]

HPV are classified phylogenetically using the highly conserved L1 open reading frame (ORF) sequence.[17,21] Genera are separated by less than 60% nucleotide identity, species display 60-70% nucleotide identity, whereas types show 71-89% similarity. Currently there are over 200 established HPV types.[21,22] Within each of these types, there are variant lineages and sublineages that differ in nucleotide identity by 1-10% and 0.5-1.0%, respectively.[17,22] A large number of single-nucleotide polymorphisms within the viral genome contributes to less than 0.5% differences below the levels of lineages/sublineages.[23]

Infection and replication

HPV infection occurs through micro-injuries in the basal epithelial layer and is linked to the differentiation cycle of the epithelium.[18,24,25] The L1 protein facilitates binding to heparin sulphate proteoglycans (HSPG) in the basal membrane. Following this, the capsid undergoes numerous conformational changes, eventually exposing a binding site on L1 for binding to a cell surface receptor on keratinocytes that have migrated to the basal membrane to close the micro-injury.[26] The cell surface receptor is proposed to be α6-integrin, although cell entry may be achieved via other receptors.[27,28] Disruption of intracapsomeric disulphide bonds leads to uncoating of the virus in the late endosomes.[29] Internalisation of capsids occur via different pathways for various HPV types, either via a clathrin-dependent endocytic mechanism, through a caveolae-mediated pathway (HPV 31) or tetraspanin-enriched domains (HPV 16).[30,31] The viral entry process is lengthy, taking between 24-48 hours.[26]

Although multiple HPV infections in the head and neck region are rare, biopsies are most commonly co-infected by HPV 16.[32] Recent evidence shows that HPV 16 is able to block, or exclude HPV 18 on the cell surface during a co-infection. This phenomenon known as superinfection exclusion is in part due to differences in the HPV minor capsid protein, L2.[33]

Following viral entry, the genome enters the cell nucleus as it requires mitosis of the infected cell.[34,35] Host cellular factors interact with the LCR to activate transcription.[31] In this initial phase called establishment, there are between 20-50 copies of the viral genome per cell. In the second phase, differentiation and proliferation takes place but there is maintenance of 20-50 copies per cell.[9]

The main viral oncogenes HPV E6 and E7 genes work in conjunction to promote replication of the infected cell. The retinoblastoma suppressor RB is targeted for degradation by HR-HPV E7 proteins,

whilst HPV E6 proteins target the *p53* tumour suppressor pathway.[36,37] E1 and E2 coordinate viral replication and host proteins. Genome amplification occurs by E2, a DNA binding protein recruiting E1 DNA helicase to the viral origin of replication. The late promoter in E7 is upregulated, expressing viral replication proteins (E1, E2, E4 and E5) without directly affecting E6 and E7.[34] Finally, after genome amplification to around 1000 copies per cell, there is an accumulation of L1 and L2 to encapsulate the viral particles.[9]

The transforming activities of HR-HPVs reflect its viral replication strategy for replication in suprabasal, normally growth-arrested differentiated epithelial cells. The viral infection does not kill the target cell, completing the life cycle and can be maintained as a chronic, asymptomatic infection. Given the high turnover rate of epithelial cells, it is remarkable that the genome is maintained episomally as plasmids in the infected cell, sometimes for decades.[24] However, progression of infection to carcinogenesis is disadvantageous to HR-HPV as cancer is an abortive, terminal event.[38]

It is unknown what triggers HPV to integrate into the human genome or whether there are any viral factors that cause integration.[39] The current paradigm regarding viral genome status characterises tumours as follows:

- Category 1 tumours: Integrated with hybrid viral-human reads
- Category 2 tumours: Episomal with no viral-human reads
- Category 3 tumours: A mixture of episomal and integrated.

Morgan et al. (2017) propose that this third category has been mischaracterised as containing integrated HPV genomes. Contradicting the previous paradigm, virus–human hybrid episomes replicate from the HPV origin and were joined to a segment of human DNA.[9]

Integration has to occur in cervical carcinomas, although it is not understood how it serves to promote carcinogenesis. The expression of the viral early gene E2 is usually disrupted, leading to upregulation of transcription of E6 and E7 viral oncogenes. This is not the case with HPV-associated HNSCC, as integration does not occur as frequently as is the case in the cervix.[40] Sites of integration tend to occur in regions of genomic instability as a consequence of HPV E6/E7 induced damage.[41–43] Viral DNA integration may occur via microhomology-based DNA repair pathways and although it occurs initially at random, integration recurs at some loci known as hotspot genes.[44] HPV types do not integrate with the same frequency; the more prevalent high-risk types, HPV 16, 18 and 45 are more likely to integrate than HPV 31 and 33.[42,45] In addition, HPV 16-associated cancers are not always integrated whereas integration is present in almost all HPV 18 carcinomas.[46] Interestingly, HPV 18 integration events appear to be more common at a specific loci near the MYC oncogene compared with HPV 16-associated cancers.[42,47] Finer distinction can be made at the HPV 16 variant level, with differences in integration potential linked to changes within the E6 gene.[48] Integration signals a poorer clinical outcome in cervical carcinomas but its role as a prognostic indicator in head and neck cancer is unclear.[9]

The carcinogenic process and tumour-suppressor genes, p16 and *p53*

In understanding the complex role played by tumour suppressor genes and oncogenes in the DNA repair pathway and carcinogenesis, it is integral to understand the normal cell cycling process (Figure 1.2). Normal cell cycling starts at the quiescence phase or G0. There are three checkpoints in place to confirm that cells are ready to continue proliferation without error. The G1 checkpoint controls the passage of G1 into S phase, verifying that the size of the cell and the environment are correct and favourable to continue. Whilst, the G2 and M checkpoints mainly prevent the cell from entering mitosis (M phase) if the genome is damaged.[49] In HNSCC, cell cycle control is deregulated at the G1–S transition. [50]

In the early G1 phase, proteins that drive the cell cycle, cyclin-dependent kinases, CDK4/6 are produced. Rb prevents excessive cell growth by inhibiting cell cycle progression until a cell is ready to divide. When cyclin D binds to CDK4/6 a reaction occurs inside that cell that causes E2F to detach from the original Rb protein. When E2F is released, it acts like a transcription factor, allowing that particular cell to progress through to the S phase. When Rb is phosphorylated to pRb, the protein is inactivated, stopping cell cycle progression. [51–53]

The HPV E7 oncogene is implicated in the HNSCC causal pathways by acting on the Rb pathway. HPV E7 inactivates the Rb protein, control of E2F is inhibited and p16 is overexpressed (Figure 1.2).[52,53] This interaction disrupts cell cycle arrest and DNA repair pathways leading to the accumulation of genetic alterations.

p16 plays an important role in cell cycle regulation by decelerating the cell's progression from G1 phase to S phase, and therefore acts as a tumour suppressor. p16 encodes a CDK repressor, inhibiting the complex formation of cyclin D1 and CDK 4/6, thus, controlling abnormal cells progressing through the cell cycle.[51] However, p16 is affected both by the activity of the HPV E7 protein as well as a chromosome deletion (chromosome band 9p21-22) that occurs early on in the carcinogenic process. p16.[54]

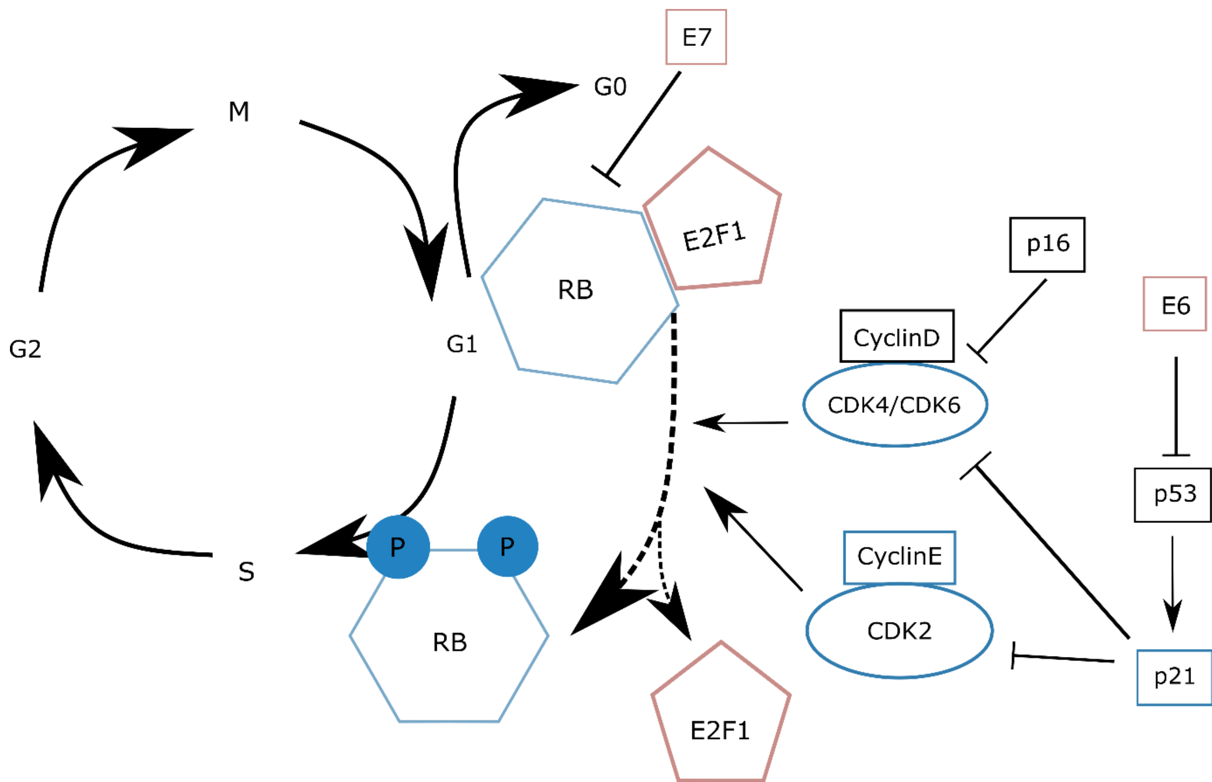


Figure 1.2. Role of E6 and E7, p16, RB and p53 in the cell cycle pathway leading to carcinogenesis. Adapted from Hayes *et al*., 2015.[52]

The *p53* gene is another tumour suppressor gene involved in restoring genomic instability. In HPV-associated head and neck cancer, wild-type *p53* is present and mutations occur at a rate of only 10% or less. Although, HPV interferes with the functioning of *p53* in other ways, E6 binds and forms a complex that leads to the degradation of p53.

Unlike HPV-driven tumours, tobacco-induced tumours frequently feature *p53* mutations, leading to impairment of protein function and genomic instability. p53 has been proposed as a biomarker to deintensify treatment of HPV-HNSCC patients. The *p53* gene is frequently mutated in most cancers, with 46-73% of HNSCC cases containing mutations.[55,56] p53 functions largely as a sequence-specific transcription factor with hundreds of targets in the human genome.[57]

The structure of the human p53 protein is shown in Figure 1.3. The amino terminus is known as the transactivation domain. The sequence-specific DNA-binding domain (amino acids 102–292) is frequently mutated in various cancers.[58] Mutations can be classed as: loss of function mutations or missense mutations.[59] Loss of function mutations (nonsense or frameshift mutations, deletions) don't produce a protein. Missense mutations result in production of a faulty protein. Transcription of p53 regulated genes occurs through the DNA binding domain (DBD), thus affecting this domain's ability to bind specifically to DNA sequence motifs (20 base pairs in length). The carboxy terminal domain, composed of amino acids 365 to 393, has strong regulatory effects upon p53 activity.[59]

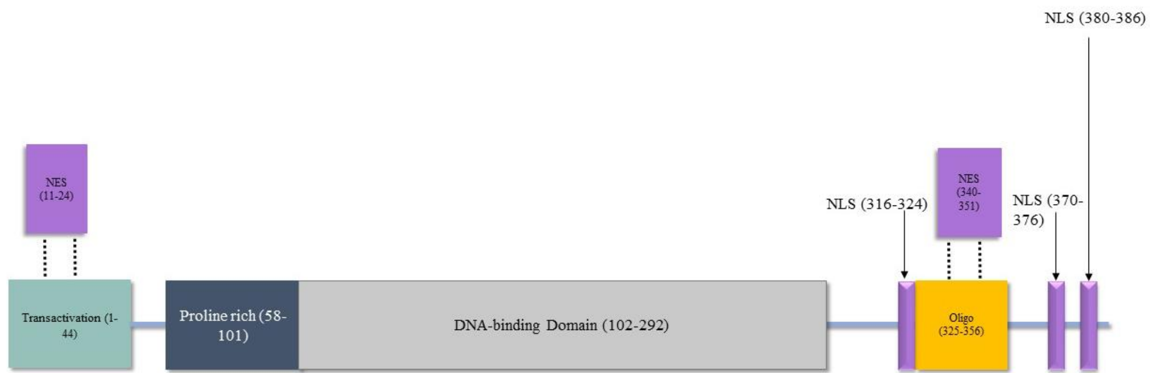


Figure 1.3. The human p53 protein is composed of 393 amino acids, numbered from the amino terminus (amino acid 1) to the carboxy terminus (amino acid 393). The DNA-binding domain (102-292) is a hotspot region for mutations in most cancers. Adapted from p53 KnowledgeBase Team (Available at: http://p53.bii.a-star.edu.sg/aboutp53/index.php).

Epidemiology of HPV-HNSCC

Globally, HNSCCs are thought to affect approximately 600 000 patients and more than 300 000 head and neck cancer deaths are attributed to HPV, annually.[32] Despite a decline in smoking, there is an increase in tonsillar and oropharyngeal cancers, linking HPV to these types of cancers.[60] There is still uncertainty, however, on the synergistic effect of tobacco/alcohol with HPV infection.[61]

HPV positive cancers differ from HPV-negative HNSCCs with regard to risk factor profiles, molecular genetic alterations and population level incidence trends over time, and prognosis.[62] HPV infection is causally associated with benign and malignant diseases of the upper airway, including respiratory papillomatosis and oropharyngeal cancer. Whether or not HPV vaccination has the potential to prevent oral HPV infections that lead to cancer or papillomatosis in the upper airway is currently unknown, as is the potential for secondary prevention with HPV detection.[61]

In terms of prevalence of HPV infection, evidence supports an increasing trend globally.[8] Since the 1970s, HPV positive tonsillar cases have risen from less than 25% to 93% of cases in 2007 in certain developed countries.[63] In the United States of America, approximately 40 to 80% of oropharyngeal cancers are caused by HPV, whilst in Europe that figure varies from 90% in Sweden to less than 20% in other communities with a high tobacco use.[6] The fact that economically developed countries have a higher incidence of oral HPV infection than developing countries, could reflect differences in sexual behaviours for oral HPV exposure including oral sex and multiple sex partners, sampling of different anatomical sites and differences in HPV detection methods.[64,65] In terms of population-based data, HPV-HNSCC patients are usually younger than HPV negative patients with a high proportion of males.[3,32,66] Despite distinct incidence trends by sex and race, the prevalence of HPV-related oropharyngeal squamous cell carcinoma (OPSCC) has significantly increased over time among women, as well as

men. In addition, among non-Caucasians, the prevalence of HPV in OPSCC also appeared to increase over time, although there was no statistical significance.[65] Much of the literature on HPV is based on studies conducted in Europe, North America and Southeast Asia, resulting in significant gaps on the reported global HPV prevalence rates, thus disparities should be interpreted with caution.[32]

High human immunodeficiency virus (HIV) prevalence in sub-Saharan Africa may contribute to increased acquisition and persistence of oncogenic HPV types at multiple anatomic sites. However, HPV prevalence data that is available for sub-Saharan Africa is limited and methods of detection vary. The oropharyngeal/oral cavity (90%) is the most commonly reported site of HNSCC with varying HPV prevalence rates obtained (1.8%-20%).[67–72] A recent publication from our research group details detection and genotyping of HPV in biopsies from patients with histologically confirmed HNSCCs. An overall total of 7/112 (6.3%) samples tested positive for HPV DNA using three PCR assays (MY09/11 and GP5+/6+ primers; PGMY09/11 and GP5+/6+ primers) as well as a multiplex heminested PCR targeting the E6 gene. Genotypes confirmed by sequencing included types 11, 16, 18, 31, 45.[73]

Not only is there a need for consensus on sampling methods to accurately determine the prevalence of HPV infection nationally, but also collectively these studies outline a role for exhaustive multi-continent research.

Clinical features and detection of HPV-positive HNSCCs

HPV-associated HNSCCs are more frequently associated with the oropharynx, whilst tobacco-associated cancers arise in the oral cavity, larynx, or hypopharynx.[74] Tobacco use is also a prognostic factor in HNC; HPV-HNSCC patients with a history of tobacco use are correlated with a worse clinical outcome than non-tobacco users. DNA damage to the *p53* gene with tobacco use allows accumulation of mutations which facilitate tumour progression.[8]

A variety of detection methods are in current use including PCR-based strategies, type-specific in situ hybridization (ISH) techniques, and immunohistochemical detection of surrogate biomarkers (e.g. p16 protein). PCR methods normally target the L1 region (MY09/11; PGMY09/11; GP5+/GP6+).[75–77]

Prevention and treatment of HPV-related HNSCC

Primary prevention efforts are focused on preventing oral infection, especially in men, who have a three-fold increased chance of HPV-HNSCC compared to women.[60] As oropharyngeal HPV infection is associated with sexual behaviours, reducing genital HPV infection through vaccination would in turn reduce the incidence of oral HPV infection. This is independent of the direct effect of the vaccine on oropharyngeal HPV infection. Although, the molecular mechanism underlying vaccine efficacy in the head and neck region would not be different from that in the anogenital tract.[78] Currently there are three commercially available prophylactic vaccines: Gardasil® (HPV 6, 11, 16, 18), Gardasil® 9 (HPV 6, 11, 16, 18, 31, 33, 45, 52, and 58), and Cervarix® (HPV 16, 18). In studies directed towards oral HPV

infection, bivalent vaccination reduced the prevalence of oral HPV 16/18 infections by 93% four years after vaccination.[79]

As indicated in the introduction, primary detection of premalignant lesions within tonsillar crypts is hindered by the lack of Pap smear equivalents available. In addition to this, secondary prevention of HPV-associated HNSCC is hindered by there not being an identifiable HPV-induced precursor lesion and lack of data on treatments for those lesions in the HNC region.[61,80]

Treatments for HNSCCs include chemotherapy, radiation, and surgery and may be used solo or in combination, depending on the stage of cancer.[81,82] HPV-HNSCC is associated with an improved prognosis and response to treatment.[61,82] In addition, distinction can be made at the subtype level in response to treatment and overall survival. Evidence has emerged that the overall survival rate for patients with tumours harbouring high-risk HPV subtypes other than HPV 16 is significantly lower than HPV 16 associated HNSCCs.[83] Even though there is a need for clinical distinction between HPV subtypes, treatment approaches should not be deintensified for all HPV-HNSCC as some patients appear to have aggressive disease.[84]

HPV genetic variants

Human papillomavirus is a highly conserved DNA virus that displays a high degree of proofreading ability with a low mutation rate. The accumulation of single nucleotide polymorphisms (SNPs) and indels that are fixed within a lineage has taken millions of years.[22,85] There is evidence of HPV 16 and 18 having diverged with the migration of *Homo sapiens* out of Africa and spreading to other continents.[86] The lineages initially corresponded to geographical locations: European, North-American, Asian-American and African.[87] Evidence for geographic distribution of other HPV variants is less clear.[88] Due to HPV having coevolved alongside humans, some isolates may persist in certain individuals based on their genetic background.[4]

A multitude of studies have begun to examine the association of HPV types and lineages with higher persistence and thus, a greater chance of progression to cancer.[89–94] Non-European variants are two to three-fold more likely to be associated with high-grade cervical lesions than is found for European variants of HPV 16.[95] Similarly, non-European variants of HPV 18 may be more common than expected in cancer specimens and high-grade cervical lesions.[96,97]

The current classification of HPV variant lineages and sublineages is based on an alphanumeric system and is linked to the original classification by geographical association.[89]

Although HPV 18 is one of the two more medically significant HPV types, only a handful of studies have described HPV 18 whole genome sequencing results. Attempts have been made to make an association between HPV 18 sublineages and specific ethnic groups, however none have been

successful thus far.[98,99] The largest study to date identified a diverse set of HPV 18 variants and obtained the complete genome sequence of 52 unique HPV 18 genomes through Sanger sequencing but was unable to assign risk to cause cancer to certain lineages.[93]

However, with next generation sequencing the field of HPV genomics has rapidly advanced. In the largest HPV whole genome study to date, over 3200 HPV 16 genomes were sequenced with the same aim of assigning cancer risk to HPV 16 variant lineages. This study was successful in assigning sublineage risk to ethnicity: Caucasian white women with sublineage HPV 16 A1/A2 were at higher risk of cervical squamous intraepithelial neoplasia 3 (CIN3+) compared to women of other genetic backgrounds; whilst Asian and Hispanic women had a higher risk associated with HPV 16 sublineages A4 and D2/D3.[90]

At a finer level of distinction, certain genes may have been under positive selection during evolutionary events, causing some HPV variants to differ in carcinogenic potential.[4] Researchers have attempted to evaluate the functional significance of sequence variation within the oncogenes and LCR of certain HPV 16 variants. Follow-up studies showed that genetic variation of the E6-coding region may possess more functional significance in the pathogenicity of HPV 16 than sequence variation of the regulatory region.[100–102] HPV 31, a close relative of HPV 16 has been investigated briefly regarding functional effects of natural sequence variation of the oncogenes. However this study found discrepancies between molecular and epidemiological data regarding variant risk, which requires further investigation.[103]

Whilst some progress has been made in associating higher persistence with certain HPV types, functional differences might not be attributed to the effect of one isolated genetic variation but to specific combinations of amino acid changes. Therefore, the increased pathogenicity related to some HPV variants could be specific to a population as a host-related factor. Nonetheless, much of the research to date has been carried out on HPV variants in cervical pathogenesis making it difficult to infer to the head and neck region.[100,101]

An aspect of viral genetic diversity of interest is within-host variance. Whilst the quasispecies phenomenon is more commonly associated with RNA viruses due to low-fidelity RNA, intracellular mutagenesis of DNA viruses, including hepatitis B and HPV have recently been described. In response to viral infection, apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like (APOBEC) proteins, play a role in the innate immune response.[104–106] APOBEC activity results in a quasispecies status of viral genomes in infected cells or tissues. In the context of HPV-mediated carcinogenesis, the clinical and biological implications of APOBEC mutagenesis is unknown.[23]

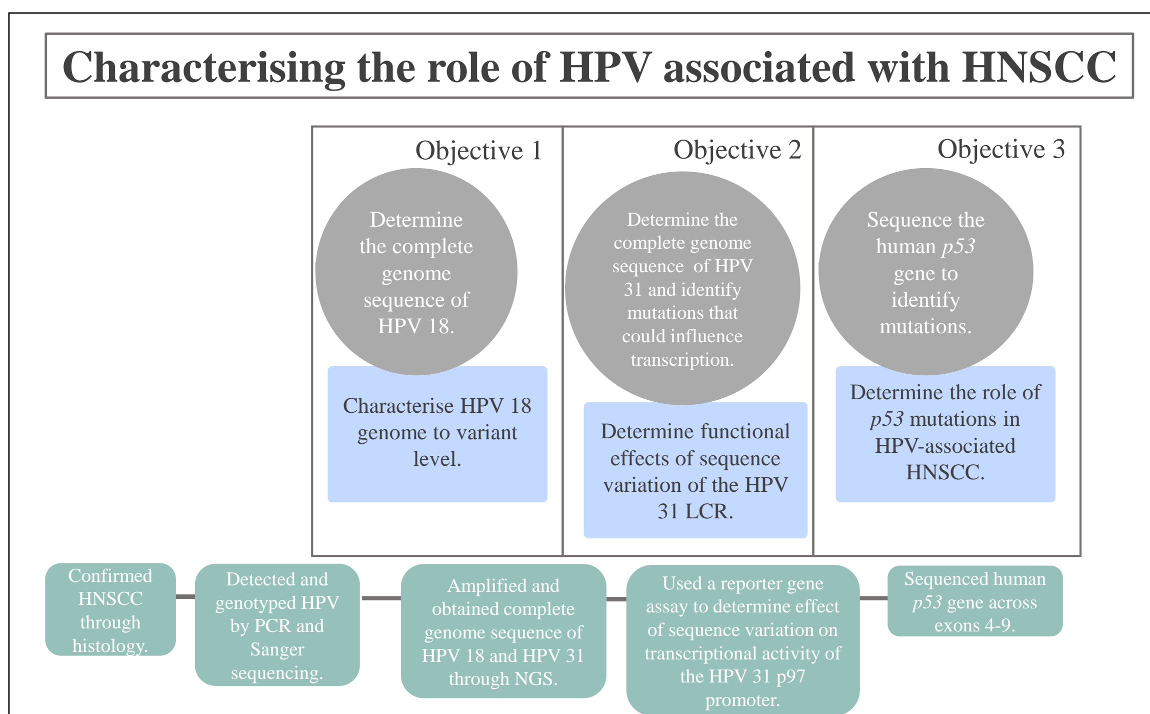
The studies reviewed here suggest the pertinence of investigating HPV variants especially in the context of HNSCC, to further understand viral evolution, epidemiology, and pathogenicity.

Rationale, aims and thesis outline

A considerable amount of research published on cervical cancer forms the paradigm for all HPV-associated cancer. Whilst the pathogenesis of HPV is similar regardless of anatomical location, there are gaps of knowledge in the field of HPV-induced HNC. HPV types differ in their propensity to cause cancer and these differences are encoded within the small and relatively conserved DNA genome. In light of a recent study published on whole genome characterisation of HPV 16 isolates from cervical cancer, more detailed and larger HNSCC associated HPV genomic studies are warranted.[90] Currently, no single study exists which characterises whole genome sequences from HNSCC. In addition, carcinogenic risk associated with HPV types, other than HPV 16 have not been extensively investigated.

An increased understanding of the genetics underlying head and neck cancers has led to HPV-associated carcinomas being classified as a distinct molecular subgroup of HNSCC. Thus it is important to investigate cancer driver genes that are associated with this virus-related epidemic. To date, no large-scale South African studies report on HPV type distribution and prevalence associated with head and neck cancer. Sequence data for South African HPV isolates from patients with HNSCC would add to the global understanding of this disease.

This thesis examines the emerging role of HPV-associated HNSCC in the context of patients treated at Universitas Academic Hospital in the Free State (FS), South Africa (SA), based on complete genome sequence and to evaluate the contribution of *p53* mutations in HNSCC. Each objective, as depicted in the box below focuses on adding to the understanding of HPV variant research and to expand knowledge of the underlying mechanisms driving HPV-associated HNSCC.



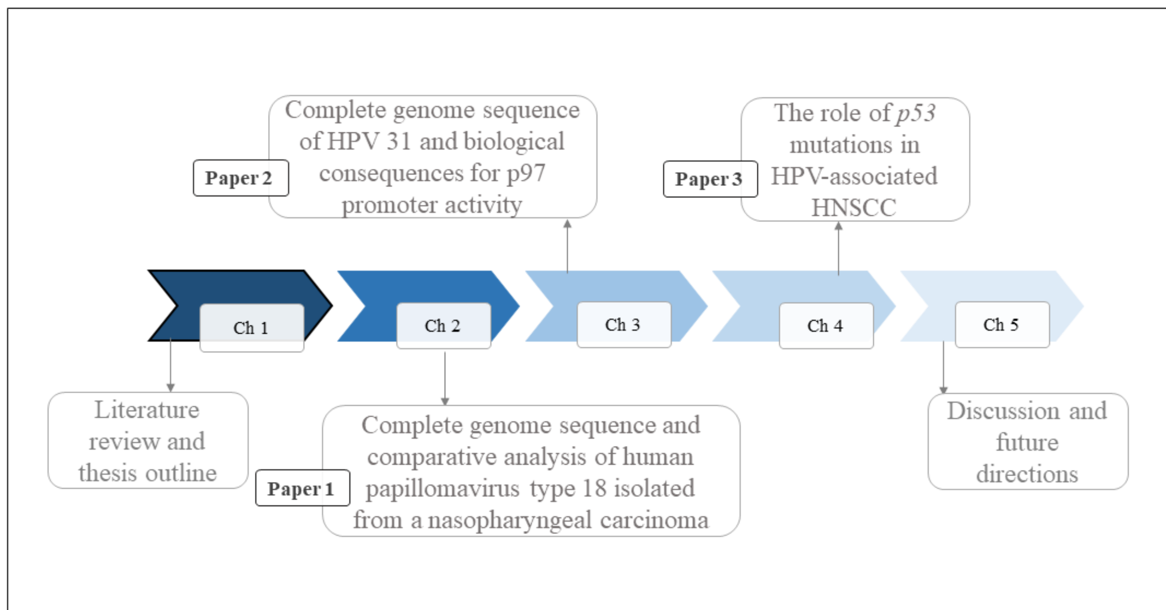
Objective 1 focused on determining the complete genome sequence of HPV isolates from patients with HNSCC in the Free State, SA, and using the sequence data to determine the presence of mutations, insertions and deletions that could influence transcription. The first complete genome sequence of HPV 18 (Chapter 2) and HPV 31 (Chapter 3) from HNSCC were characterised down to the variant level.

Objective 2 focused on investigating the influence of mutations in the HPV 31 long control region (LCR) on promoter function by means of functional assays. In this study (Chapter 3), mutagenesis and functional assays were performed using a reporter gene assay in order to determine whether a single nucleotide change and a 10 bp insertion in the LCR of a South African HPV 31 isolate have potential to modify the transcriptional activity of the p97 promoter. This study also reports the coexistence of two closely related HPV 31 quasispecies in a head and neck cancer patient.

Objective 3 focused on investigating *p53* mutations in HPV associated and HPV negative HNSCC samples. In this study (Chapter 4) the human *p53* gene was sequenced across exons 4-8, to identify mutations. *p53* mutation frequencies were significantly lower than expected in this cohort, although the functional significance of the intronic mutations observed is unknown.

Structure of thesis

The thesis is presented as three publishable papers, with a literature review and overall discussion according to the guidelines from the University of the Free State with regard to submission of a thesis in article format. To simplify formatting and presentation the references for each chapter are presented as one list at the end of the thesis. The thesis is organised in three distinct sub-sections, as depicted below.



Chapter 1 provides the literature review, background and rationale, as well as the aim and objectives of this research. The subsequent chapters are presented as a series of research articles which will be submitted for consideration for publication in selected international scientific journals. The final section of this thesis, Chapter 5, summarises the key research findings and discusses the implications of HPV-HNSCC variant research in the context of sub-Saharan Africa.

CHAPTER 2

Complete genome sequence and comparative analysis of human papillomavirus type 18 isolated from a nasopharyngeal carcinoma

Complete genome sequence and comparative analysis of human papillomavirus type 18 isolated from a nasopharyngeal carcinoma

Y Munsamy

Will be submitted for consideration for publication in *Papillomavirus Research*

Abstract

High-risk human papillomaviruses (HPV) are considered as one of the aetiologic agents of head and neck cancer. HPV 16 and HPV 18 account for most HPV-associated head and neck cancers. The complete genome of an isolate of HPV 18 (designated VBD 17/15) amplified from a nasopharyngeal carcinoma biopsy was determined using next generation sequence analysis. The genome was 7857 nucleotides in length and shared 0.15% nucleotide identity with the reference HPV 18 genome. Phylogenetic analysis based on the complete genome and using sequence data retrieved from GenBank for ten isolates representing each HPV 18 lineage, showed that VBD 17/15 clustered in lineage A1. Sequence variation within the E2 gene may have an impact on the oncogenic potential of the virus. Mutations novel to this isolate included an amino acid change in the L2 protein coding sequence, which could affect virus assembly and the infectious process, although functional differences cannot be confirmed as of yet. No risk of progression to cancer can be assigned to HPV 18 variants as there is no sampling procedure available for precancerous lesions in the head and neck region. The sequence diversity and phylogeny of the first HPV 18 isolate from a nasopharyngeal carcinoma provides the basis for future studies investigating the role of genetic variation in HPV epidemiology and head and neck carcinogenesis, especially within the Sub-Saharan African context.

Keywords

Human papillomavirus, HPV 18, head and neck squamous cell carcinoma, whole genome sequencing, HPV genomics

Introduction

Human papillomaviruses (HPV) are a family of small double-stranded circular DNA viruses with a genome of approximately 8 000 base pairs (bp) that contain between eight and nine open reading frames (ORFs).[11] Their genomes share a common organisation which includes the non-structural or early (E) genes (E1, E2, E6, and E7), the late region encoding viral structural proteins (L1, L2) and the noncoding long control region (LCR) containing the viral promoters. HPV is a well-known oncogenic virus, although of the approximately 200 types, only 13 HPV types belonging to the alpha genus are defined as high-risk (HR).[20] An aetiologic role for HPV has been established in head and neck squamous cell carcinoma (HNSCC).[5] In 2012, there were approximately 38 000 incident cases of head and neck cancer

attributable to HPV globally.[107,108] HPV 16 is the most carcinogenic HPV type, associated with approximately 50% of all cervical cancers, the majority of other HPV-related anogenital cancers, and more than 80% of HPV associated head and neck cancers.[10,106–108]

The second most frequently identified HR-HPV type, HPV 18 contributes to approximately 2.5% of head and neck cancers worldwide.[10] It is not yet known what determines the pathogenicity of high-risk types HPV 16 and 18: whether there are genetic variations linked to viral fitness or host factors that are involved, are yet to be elucidated.[112]

Whole-genome sequence analysis allows for investigation of the genome in greater detail for the discovery of novel single nucleotide polymorphisms (SNPs) or large contiguous deletions.[113] Despite HPV being considered a highly conserved DNA virus, ten different HPV 18 viral variant lineages and sublineages exist. Based on a whole-genome approach, differences of ∼1.0% define HPV variant lineages, and differences of 0.5 to 0.9% define HPV sublineages.[114] Based on complete HPV genome sequence data, three variant lineages, A, B and C, comprised of eight sublineages, A1–5, B1–3, have been defined for HPV 18.[89]

HPV types, differing in oncogenic potential and viral genetic variation within a specific type might be associated with varying risk for cancer. This might be due to difference in persistence or risk of progression to cancer. In the case of HPV 18 infections, the majority are cleared by the immune system. However, a small proportion of infections progress to cervical cancer and some studies have implicated HPV 18 genetic variation as a factor.[96] However, there are contradictory stances on whether HPV 18 variants differ in risk for cancer in the cervix.[115] Supporting this, a global study stratifying risk for cervical cancer between HPV 18 genetic variants and ethnically diverse females, also concluded that there was no role of HPV 18 (sub)lineages for discriminating cancer risk.[99] Hence the acquisition of complete genome sequence data will contribute to understanding the role of genetic variation in carcinogenesis especially within the context of the emerging HPV-HNSCC.

Questions that need to be asked include; are there differences in the genome of isolates from head and neck sites compared with isolates from cervical cancers and do these mutations contribute to viral pathogenicity? While some research has been carried out on cervical carcinomas, no single study exists which characterises whole genome sequences from HNSCC. In this study the first whole genome sequence of HPV 18 isolated from a nasopharyngeal carcinoma was determined using next generation sequencing (NGS) and the complete genome characterised for identification of genetic variations. Comparison of the genetic relationship with 125 isolates using data retrieved from GenBank was used to characterise the South African isolate.

Methods

Sample

An isolate of HPV was amplified using PCR from a biopsy collected from a patient with histologically confirmed nasopharyngeal carcinoma, treated at Universitas Academic Hospital (Bloemfontein, Free State, South Africa). The isolate was assigned laboratory number VBD 17/15. This study was approved by the University of the Free State Health Sciences Research Ethics Committee (ECUFS 137/2013D). Written informed consent for study participation was obtained from the patient.

DNA extraction and PCR for detection and genotyping of sample

DNA was extracted from fresh biopsy tissue using the QIAamp DNA Mini Kit (QIAGEN, California, United States of America) according to manufacturer's instructions. HPV was detected and genotyped using two conventional PCR assays; a nested PCR with primer pairs MY09/11 and GP5+/6+, targeting the L1 gene and an in-house multiplex hemi-nested PCR targeting the E6 gene.[73,113] A region of the β-globin gene was amplified concurrently using the primer pair PC04/GH20, as an internal control.

The PCR amplicons were genotyped using bi-directional Sanger sequencing. The resultant sequence data was edited with Chromas Pro version 1.41 (Technelysium Pty Ltd, Australia) and aligned with sequence data retrieved from GenBank from a Basic Local Alignment Search Tool (BLAST) (http://blast.ncbi.nlm.nih.gov/Blast.cgi) analysis in order to confirm the HPV type.

Determination of complete genome sequence using next generation sequencing

To amplify the full length genome in two overlapping fragments (E1 to L1 genes; L1 to E1 genes), primers were designed based on alignment of the sequence data of HPV 18 complete genomes retrieved from GenBank (Accession numbers available in Appendix B). Nucleotide sequences for each primer and position relative to HPV 18 reference strain (NC_001357) are shown in Table 2.1. Amplification was performed using the Phusion® HotStart DNA Polymerase-mediated PCR amplification kit (ThermoFisher Scientific, Massachusetts, USA) and 1 pg–10 ng template, according to manufacturer's instructions. Cycling conditions consisted of an initial incubation at 98 °C for 30 s, followed by 30 cycles of alternating 98 °C for 10 s, 64 °C (for primers F1/R1) or 65 °C (for primers F2/R2) for 30 s and 72 °C for 2 minutes 30 s. A final elongation of 5 minutes at 72 °C was included. Amplification was verified by separation of PCR products by electrophoresis on a 1% agarose gel. The amplicons were excised and purified from agarose gel using Promega Wizard® SV Gel PCR Clean-Up System kit (Promega, Wisconsin, United States of America) according to manufacturer's instructions.

Table 2.1. Sequences of each primer pair used to amplify HPV 18 in two overlapping fragments with predicted amplicon size.

Primer name	**Primer sequence**	**Annealing temperature (°C)**	**Position relative to HPV 18 reference strain (NC_001357)**	**Expected amplicon size (*bp)**
HPV_18F1	5'-GGAGATTGGAGACCAATAGTG-3'	64	2243-2263	~4438 bp
HPV_18R1	5'-CATATTGCCCAGGTACAGGAG-3'		6681-6661	
HPV_18F2	5'-ATTCTCCCTCTCCAAGTGGC-3'	65	6484-6503	~4023 bp
HPV_18R2	5'-CATCTAACATGGCCACCTTAG-3'		2501-2481	

*bp = base pairs

MiSeq library preparation and sequencing

The purified DNA was converted to a short fragmented DNA library using the Nextera XT DNA Library Preparation kit (Illumina, California, United States of America), followed by size selection with AMpure XP beads (Beckman Coulter, California, United States of America). The multiplexed libraries were analysed on a MiSeq sequencer (Illumina, California, United States of America) with the MiSeq reagent kit v3 (300 cycle) (Illumina, California, United States of America) at the University of the Free State Next Generation Sequencing Unit.

Next-generation sequencing data analysis

The raw sequencing data was converted from SFF format to FASTQ files using the sff_extract script (available as part of seq_crumbs at http://bioinf.comav.upv.es/). PRINSEQ was used to trim and filter reads based on length and quality scores (≥QC30).[114] Whole HPV 18 genome sequences from the GenBank database were used to compile unique databases and to separate contaminating sequence data using filter_by_blast (available at http://bioinf.comav.upv.es/seq_crumbs/available_crumbs.html). *De novo* assembly of the blast-filtered and unfiltered reads was performed using SPAdes v.3.7.1 into scaffolds.[115] Read mapping to the consensus sequence was conducted with Bowtie2 and appropriate file conversions were conducted with SAMtools.[116] Contiguous segments were assembled, primer sequences were removed from sequence data. Visualisation in Integrated Genomics Viewer allowed for comparison to HPV 18 sequence data retrieved from GenBank to identify areas of incomplete coverage or ambiguities.[117]

Phylogenetic relationship of HPV 18

Complete genome sequence data for 125 isolates from cervical carcinoma were retrieved from GenBank (Accession numbers are available in Appendix B). The evolutionary history was inferred by using the Maximum Likelihood method based on the Tamura-Nei model.[121] The bootstrap consensus tree inferred

from 1000 replicates is taken to represent the evolutionary history of the taxa analysed. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 0,0559)). The analysis involved 126 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated. There were a total of 7761 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Figure 2.2).[122]

Phylogenetic analyses were then performed using the above method with 10 reference strains to confirm lineages A, B and C and sublineages A1–5, B1–3. References were identified by Burk et al. (2013) by plotting pairwise comparisons within each variant lineage or between variant lineages, with an approximate cut-off of 1.0% difference between complete genomes to define major variant lineages.[89]

HPV 18 variant analysis

Variations within VBD 17/15 genome were identified by alignment of the sequence data with the HPV 18 reference strain (NC_001357), belonging to A1 lineage. The prototype or reference sequence (i.e., the cloned genome designated as the original type) is always designated variant lineage A and/or sublineage A1.[17] The HPV 18 reference was originally cloned from a cervical carcinoma from a Brazilian patient.[123] In addition, alignment of the sequence with the complete genome of ten HPV 18 variant lineage/sublineage was analysed with regard to the number of mismatched bases, to visually differentiate lineage/sublineage-specific SNPs.[89] Table 2.2 shows information for each of the ten representative genomes used in the analysis including geographic origin of sample, lineage designation, length of complete genome and GenBank accession number.

Table 2.2. HPV 18 isolates, previously identified for discrimination between lineage/sublineage-specific SNPs.[89]

	HPV Variant Lineages	GenBank accession number	Anatomical location	Country	Length of genome	GC content %	Reference
1	HPV 18 Reference Lineage A1	NC_001357	Cervix	Brazil	7857	40.4	[124]
2	HPV 18 Lineage A1	AY262282	Cervix	Unknown	7857	40.44	
3	HPV 18 Lineage A2	EF202146	Cervix	Costa Rica	7857	40.38	[86]
4	HPV 18 Lineage A3	EF202147	Cervix	Costa Rica	7857	40.41	
5	HPV 18 Lineage A4	EF202151	Cervix	Costa Rica	7857	40.33	
6	HPV 18 Lineage A5	GQ180787	Cervix	Thailand	7844	40.29	[125]
7	HPV 18 Lineage B1	EF202155	Cervix	Costa Rica	7824	40.12	[86]
8	HPV 18 Lineage B2	KC470225	Cervix	Unknown	7824	40.07	[114]
9	HPV 18 Lineage B3	EF202152	Cervix	Costa Rica	7844	40.06	[86]
10	HPV 18 Lineage C	KC470229	Cervix	Unknown	7837	40.14	[126]

Results

Genotyping of sample

The isolate was genotyped as HPV type 18 using primers targeting a region of the L1 gene and using bi-directional sequencing to obtain sequence data for this region.

Next-generation sequencing data

The mean sequence length achieved was 166.93 ± 50.68 bp, the minimum length was 35 bp, whilst the maximum length was 201 bp. The length range was 167 bp whilst the mode length was 201 bp with 135 173 sequences and an average coverage of 400 x.

To obtain the complete genome of VBD 17/15, two overlapping fragments were amplified and the resultant genome assembled into eight open reading frames with two noncoding regions: the intergenic region between the E2 and E5 genes and the long control region. The complete genome of isolate VBD 17/15 was 7857 bp in length with a GC content of 40.42%. The position of the first and end nucleotide (nt) of each gene, or region, and the length of each gene is shown in Table 2.3. The full genome was annotated indicating location of the early and late regions (Figure 2.1).

Table 2.3. Annotation of genes of isolate VBD 17/15 (Geneious V7.0 (Biomatters).

Gene/ region*	Start position nt	End position nt	Length bp
E6	105	581	477
E7	590	907	318
E1	914	2887	1974
E2	2817	3914	1098
E4	3418	3684	267
E5	3936	4157	222
L2	4244	5632	1389
L1	5430	7136	1707
LCR	7137	104	825

*E = early genes; L= late genes; LCR = long control region; nt = nucleotide; bp = base pairs;

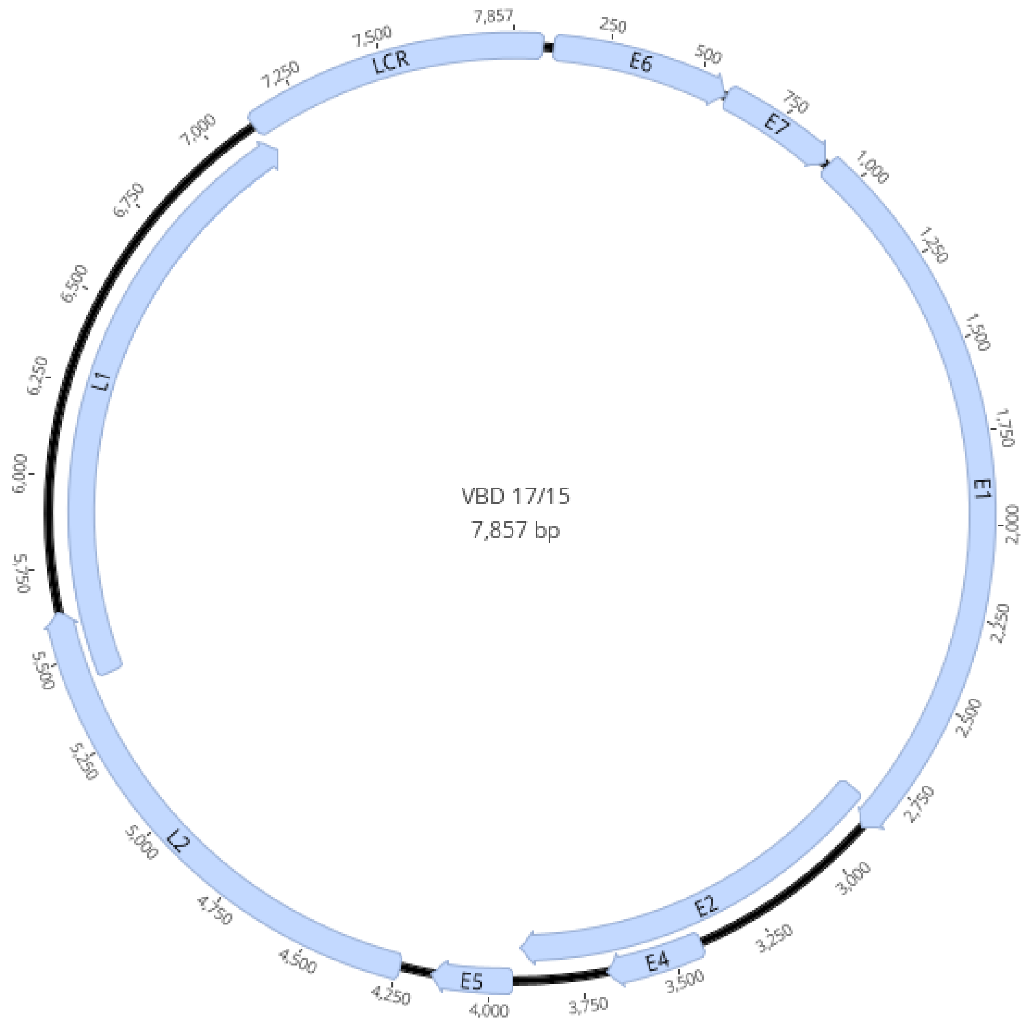


Figure 2.1. Complete genome of VBD 17/15, 7857 bp in length with 40.42% GC content, indicating the location of open reading frames E1 to E7 coding for early proteins and L1 and L2 for late proteins, and the noncoding long control region, LCR, between L1 and E6 genes. Image constructed using Geneious version 2019 (Biomatters).

Comparative phylogenetic analysis with 125 complete HPV 18 genomes

Multiple sequence alignments were conducted to determine the genetic relationship of the South African HPV isolate from a nasopharyngeal carcinoma and isolates from cervical carcinomas was determined using complete sequence data retrieved from GenBank for 125 HPV 18 cervical cancer isolates. The evolutionary history was inferred using the Tamura-Nei method (Figure 2.2).[121]

Nucleotide sequence differences across the complete HPV genome of 1.0% to 10.0% and 0.5% to 1.0% define distinct HPV variant lineages and sublineages, respectively.[22] In this study, the maximal pairwise difference between nucleotide sequences among the 126 complete HPV 18 genomes analysed was 2.13%. There are three HPV 18 variant lineages and nine distinct sublineages. In addition, lineage A

consisted of five sublineages, A1-A5. Within sublineage A1 there were 27 isolates from geographically distinct regions. VBD 17/15 clustered in HPV 18 Lineage A, with 99.8% nucleotide identity to the reference genome (Table of mean nucleotide sequence differences between each isolate is too extensive to include, available on request). VBD17/15 had the highest nucleotide homology to isolates from a Dutch cohort.[93]

The phylogenetic analysis based on VBD 17/15 and 10 isolates representing each viral variant lineage and sublineage was used to further confirm the lineage identity of the South African isolate. The evolutionary history was inferred using the Tamura-Nei method shown in Figure 2.3.

Estimates of evolutionary divergence between sequences is shown in Table 2.4. The number of base substitutions per site from between sequences are shown. Standard error estimate(s) are shown above the diagonal and were obtained by a bootstrap procedure (1000 replicates).

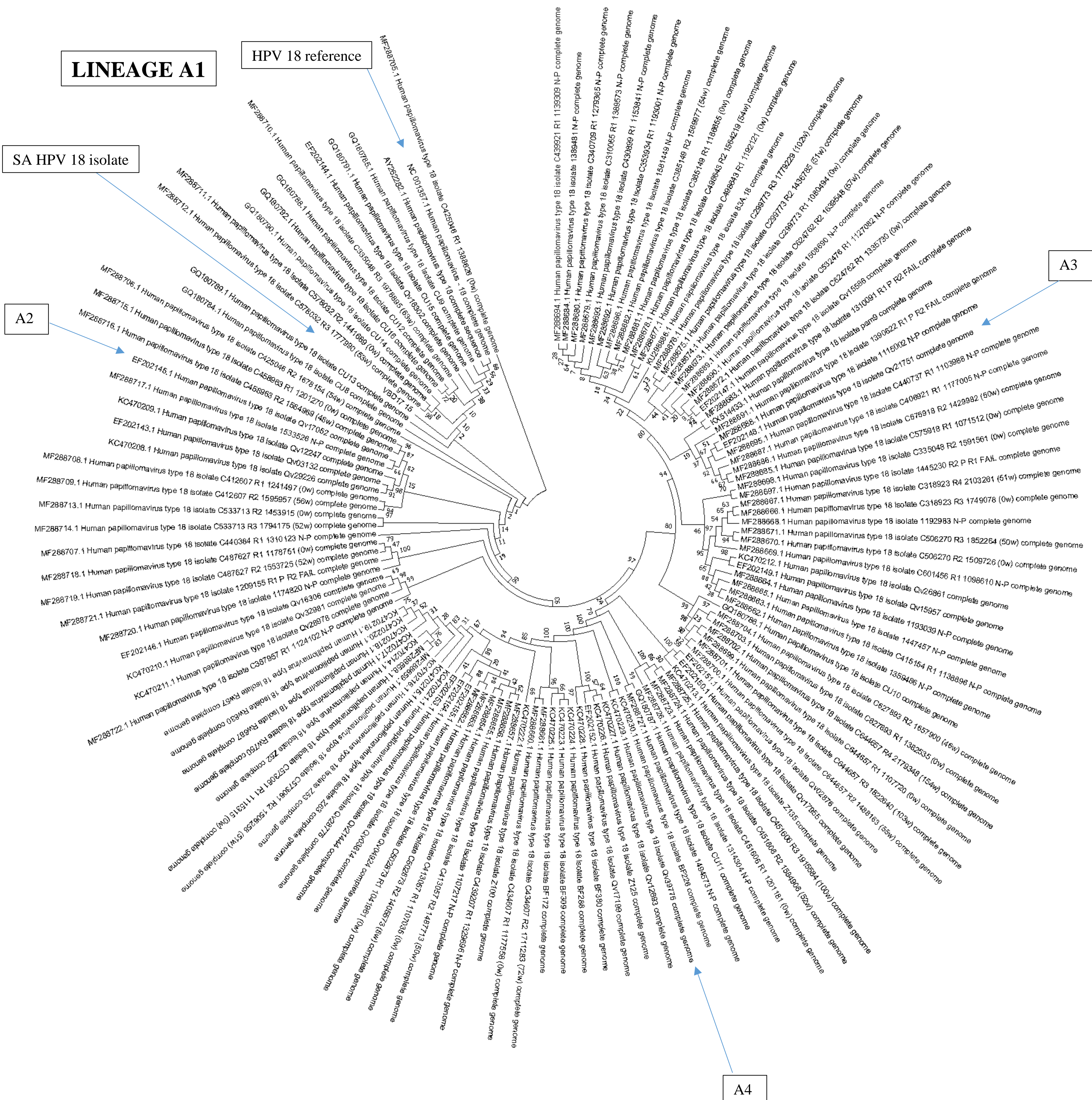


Figure 2.2. The maximum likelihood tree was inferred from a global alignment of 126 complete sequences of HPV 18 isolates from cervical carcinoma and VBD 17/15 isolated from HNSCC. A bootstrap value of 1000 replicates was employed. Tree constructed with MEGA 7.0. Each isolate is represented by a GenBank accession number.

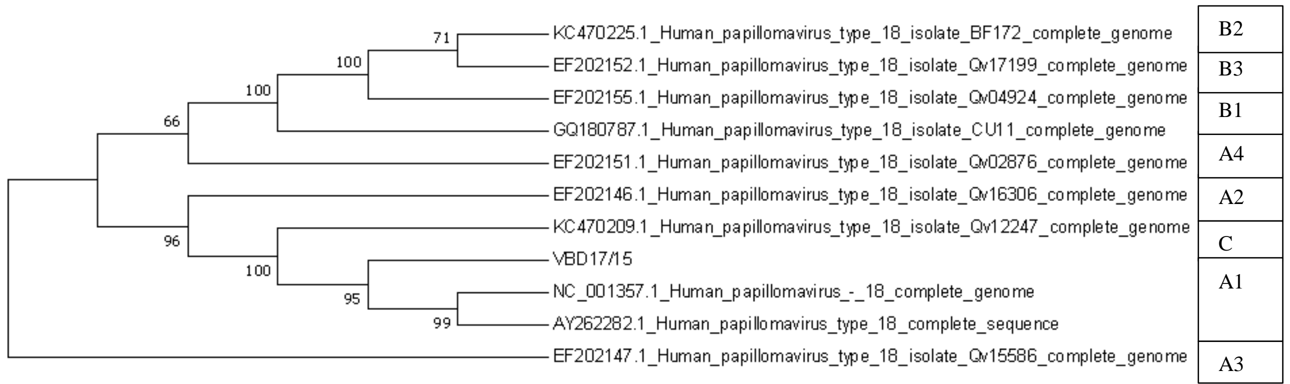


Figure 2.3. HPV 18 variant tree topology using complete genomes. The evolutionary history was inferred by using the Maximum Likelihood method based on the Tamura-Nei model. The bootstrap consensus tree inferred from 1000 replicates is taken to represent the evolutionary history of the taxa analysed. Evolutionary analyses were conducted in MEGA7.

Table 2.4. Estimates of evolutionary divergence between sequences. Standard error of distance is depicted in upper right corner.

Standard error estimate(s)											
	1	2	3	4	5	6	7	8	9	10	11
1 VBD 17/15		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.06	0.06
2 NC_001357.1	0.00		0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.06	0.06
3 KC470229.1	0.00	0.00		0.00	0.00	0.00	0.00	0.00	0.01	0.06	0.06
4 AY262282	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.01	0.06	0.06
5 EF202151.1	0.00	0.01	0.01	0.01		0.00	0.00	0.00	0.01	0.06	0.06
6 EF2021547.1	0.00	0.01	0.00	0.00	0.00		0.00	0.00	0.01	0.06	0.06
7 EF202146.1	0.00	0.01	0.00	0.00	0.01	0.00		0.00	0.01	0.06	0.06
8 GQ 180787.1	0.06	0.06	0.06	0.06	0.06	0.06	0.06		0.00	0.08	0.08
9 EF202152.1	0.07	0.08	0.07	0.07	0.08	0.07	0.07	0.02		0.08	0.08
10 KC470225.1	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.63	0.61		0.00
11 EF202155.1	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.64	0.61	0.00	
Estimates of evolutionary divergence between sequences											

Variance across each gene or region of the genome of HPV 18 lineages/sublineages, based on the ten representative genomes, ranged from 7.55% to 16.35%. The most variable region was the gene encoding the E7 protein and the most conserved was the gene encoding the E1 protein. Differences observed across the genomic regions suggest that different evolutionary rates would be expected. The number of SNPs detected and the size of each region or gene in which the SNPs were detected were taken into account to determine variance, shown in Table 2.5 and Figure 2.4.

Table 2.5. Percentage of variance calculated from the number of SNPs identified within each region/or gene investigated and the size of each region/or gene.

Gene/ region	Number of SNPs	Region size in nucleotides	% Variance
E7	52	318	16.35%
L2	192	1389	15.33%
LCR	100	825	12.12%
L1	192	1707	11.25%
E4	29	267	10.86%
E6	50	477	10.48%
E5	21	222	9.46%
E2	97	1098	8.83%
E1	149	1974	7.55%

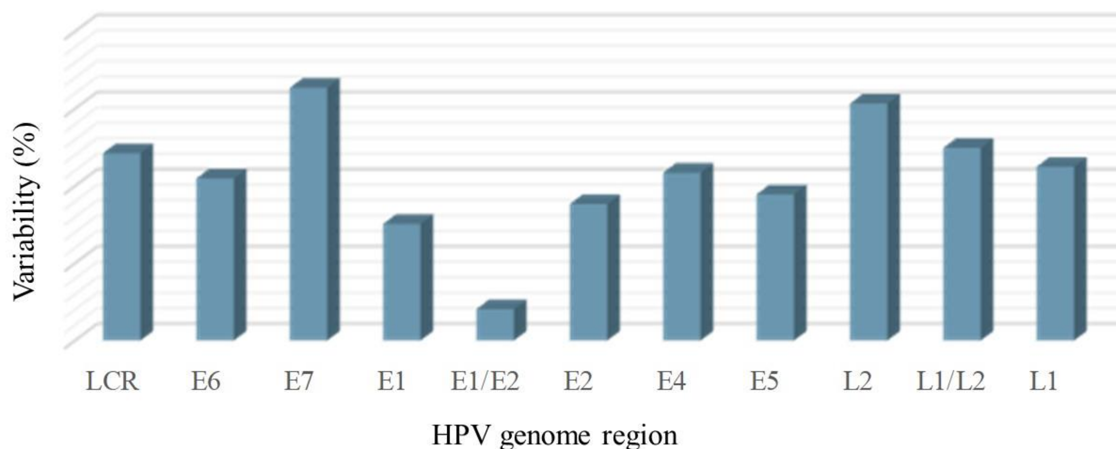


Figure 2.4. Illustration of percentage variability within different genes/regions across HPV 18 lineages.

HPV 18 Lineage A1 SNP analysis

To identify variation within the sublineage A1, the HPV 18 reference isolate (NC_001357), the representative isolate for lineage A1 (AY262282) and South African VBD 17/15 isolate were compared and a total of 14 genetic variations were identified (0.19% of genome). These variations included three transitions and 11 transversions, without insertions or deletions.

A nucleotide substitution, T4315A in VBD 17/15, was completely novel to this South African isolate from a nasopharyngeal carcinoma.

The following mutations were observed in VBD 17/15 (lineage A1), and all other lineages but not in the other two isolates belonging to sublineage A1:

- Within the L2 gene, G4772T, C5147A and T5358C
- Within the L1/L2, G5503A
- In the LCR, T7592C.

A summary of the number of amino acid substitutions that are unique to the VBD 17/15 isolate, displayed by gene, is shown in Table 2.6.

Table 2.6. Variations identified in VBD 17/15 when compared with the HPV 18 reference isolate.

Gene/ region	Gene size (bp) with nucleotide positions indicated in brackets	Nucleotide change	Point mutations (N)								Amino acid variation		
			Nucleotide position	**Total**	**Non-synonymous mutations**			**Synonymous mutations**			**Amino acid site number**	**HPV 18 reference**	**VBD 17/15**
					Total	**Ts**	**Tv**	**Total**	**Ts**	**Tv**			
E6	477 (105-581)	C-G	287	1	0			1		X	96	Proline	Arginine
E1/E2	2817-2884	T-G	2856	4	3		X						
		G-C	2857				X						
		C-G	2858					1		X			
		G-T	2859				X			X	953	Alanine	Arginine
E2	1098 (2817-3914)	C-G	3084	3	1		X						
		G-C	3085					2		X	1029	Aspartic acid	Histidine
		C-G	3275							X	1092	Proline	Arginine
L2	1389 (4244-5632)	T-A*	4315	4			X	0			1439	Arginine	Valine
		G-T*	4772				X				1591	Glycine	Valine
		C-A*	5147				X				1716		
		T-C*	5358			X							
L1/L2	5430-5632	G-A*	5503	1		X		0			1835	Glycine	Serine
L1	1707 (5430-7136)	C-G	5701	4	4		X	0					
		C-G	6460				X						
		C-G	6625				X						
		C-G	6842				X						
LCR	825 (7137-7857)	T-C*	7592	1									

*Changes within the VBD 17/15 isolate, showing variation in the A1 lineage.

∆ Change unique to VBD 17/15, in comparison to all other 125 isolates from cervical carcinomas.

HPV 18 variant lineages SNP analysis

Viral SNPs specific for HPV 18 variant lineages were identified by alignment with the reference genome and included all variant lineage sequences obtained on GenBank and VBD 17/15. In total, 675 SNPs were identified across all variant lineage/sublineages using the viral whole genome sequences. There were no insertions, however three deletion events were identified in sublineages A5, B1, B2 and B3 including a 6 bp deletion in the E2/E4 coding region; a 20 bp deletion in lineages B1 and B2 and a 7 bp deletion in lineage C in the noncoding intergenic region between the E2 and E5 genes; and a 7 bp deletion was identified in the long control region (LCR). The deletion in the E2/E4 region led to deletion of two amino acids, leucine and histidine, which appears to distinguish between the A5, B1, B2 and B3 sublineages. All viral SNPs specific for HPV 18 variant lineages/sublineages are presented in Figure 2.5.

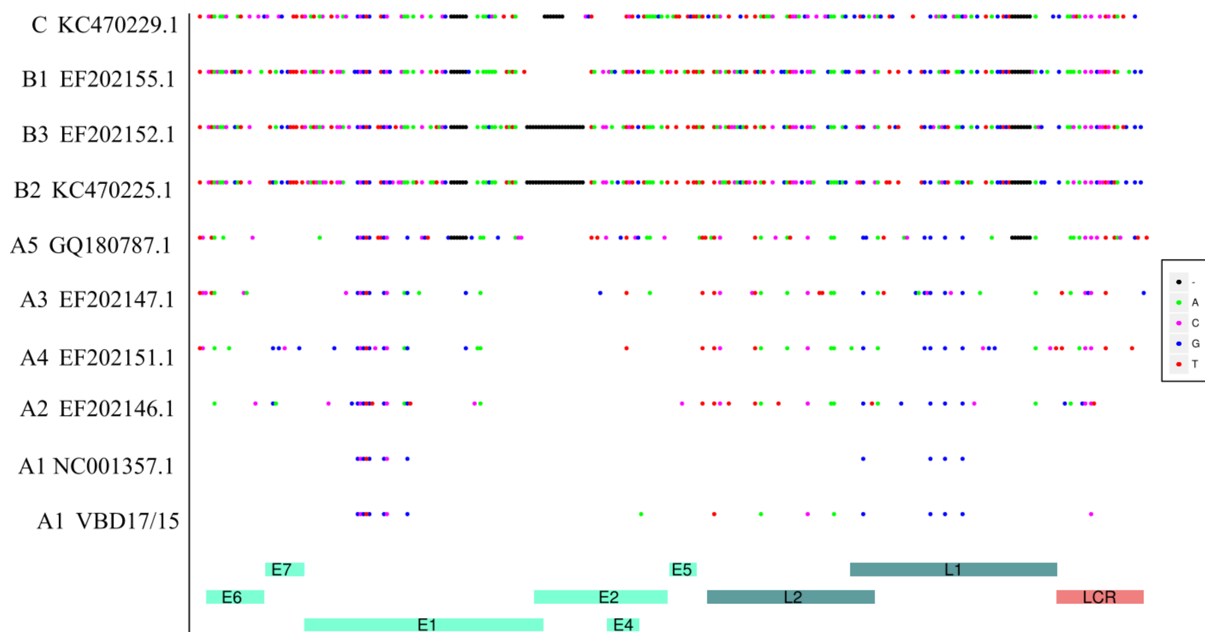


Figure 2.5. Genomic plot of single-nucleotide polymorphisms (SNPs) across human papillomavirus (HPV) 18 variant lineage/sublineage genomes in comparison to the reference HPV 18 genome (NC_001357). Each dot denotes a variable site, with nucleotide change depicted in the colour key box. Black bars represent deletions.

Discussion

Although HPV 18 is the second most prevalent HPV type associated with head and neck cancers and cervical cancers, literature describing diversity based on complete genome sequence is scarce.[10] Complete genome sequence data for a vast number of HPV 18 isolates from cervical carcinomas has become available, although the genetic characteristics of HPV 18 from HNSCC has not yet been

detailed. This underscores the need for intensive molecular studies to further characterise HPV associated with HNSCC.

Although HPV is a small, conserved DNA virus, the molecular mechanisms that underpin genomic variation and viral fitness are not fully understood.[109] Thus far, it is unclear whether HPV is implicated in cancers of other sub-sites of the head and neck region, besides the oropharynx or whether HPV presence is incidental.[62] In a previous study in our laboratory, biopsies of histologically confirmed squamous cell carcinomas of the oropharynx, nasopharynx, larynx and hypopharynx from 112 South African patients were screened using three PCR assays targeting the L1 and E6 regions of HPV and p16 immunohistochemical staining. HPV DNA was identified in 7/112 (6.3%) tumours, a relatively low HPV DNA positivity was achieved in this cohort.[73] Thus, making the analysis of the complete genome of HPV 18 from a nasopharyngeal carcinoma patient valuable. In the absence of data on HPV 18, reference is made to other HR-HPV.

This study describes the first characterisation of HPV 18 from the head and neck region and examines the genomic relationship with isolates from cervical carcinomas. Previous studies have observed that diversity within HPV 18 correlates with patterns of the evolution and spread of *Homo sapiens*.[71] According to Chen *et al*. the A lineage predominates in most regions except sub-Saharan Africa, where the B lineage is predominant. Interestingly, isolate VBD 17/15 was identified as belonging to HPV 18 Lineage A1, which predominates in eastern Asia and the Pacific.[99]

In total, 675 SNPs were identified across all variant lineage/sublineages, with a predominance of C-G transversions, Phylogenetic analysis in this study of variant sub/lineages shows that VBD 17/15 did not display a large amount of diversity with regards to isolates obtained from cervical carcinomas. With specific reference to isolate VBD 17/15, the deletion affecting the coding region, E2/E4, occurred simultaneously with deletions in the noncoding E2/E5 intergenic region and the LCR (7 bp). This combination of linked genes in non-random proportions, known as linkage disequilibrium, is resultant of papillomavirus evolution and involves genetic drift.

The deletions discriminate the A5, B1, B2, B3 and C lineages from the other variant lineages and have been described previously.[86,93,120] These deletions contribute to the genetic distance of the aforementioned lineages from lineage A, as they are not present within lineage A1, A2, A3 and A4. Many lineage B and C-specific SNPs were distributed throughout the whole genome, which reflects the phylogenetic distance of these lineages from lineage A, as shown in Figure 2.5. These observations further support the use of complete genome sequencing for the basis of HPV variant classification.

Substitutions at specific positions serve as signatures for higher risk for cervical neoplasia. For example, T20I and G63S substitutions in E7 of HPV 58 are associated with a higher risk for cervical neoplasia.[110] Although, no specific molecular signature has yet been identified to discriminate between clearing HPV 18 infections and HPV 18 that progresses to carcinoma.[93]

In this study a comparison of genetic regions of HPV 18 sublineages and the isolate from the nasopharyngeal carcinoma, the E1 gene was the most conserved region. The alignment of 126 HPV 18 isolates suggested that the E7 gene was not as highly conserved (16.35% genomic sequence variation). In a study of over 5000 HPV 16 genomes from cervical cancer, to determine whether viral genetic variation influences risk of cervical precancer and cancer, the E7 gene showed more variation in patients that did not progress to precancer/cancer. Whereas, there were no E7 genetic variants in those that had cancer. Thus, strict conservation of the E7, which disrupts the tumour suppressor functioning of the RB protein, is critical for HPV 16 carcinogenesis.[122]

In comparison with the reference isolate and A1 lineage isolates, six SNPs were identified in VBD 17/15; one of which was a novel SNP in the L2 gene. It is unlikely that these changes are a PCR artefact, as a high fidelity enzyme was used to amplify the isolate and NGS data had high quality score with sufficient read depth. Although NGS data has a higher resolution than Sanger sequencing data, it would be expected to detect either HPV co-infections or HPV type-specific variant co-infections. However, deep sequence data did not demonstrate evidence of HPV 18 variant co-infections in this sample.

Four amino acid changes were identified within the genome of VBD17/15: one in E6, one in E1/E2 and two in E2. Sequence variation within the E2 gene may have an impact on the oncogenic potential of the virus.[100]

In addition, no inference can be made between disease progression and amino acid alterations in HPV associated HNSCC as can be done for cervical cancer. The change from arginine, a charged amino acid, does not seem to affect a specific potential functional region (residue 180) although it might affect papillomavirus assembly and the infectious process.[123] Refer to Appendix C for a published list of known protein interactions with papillomavirus L2. However, it is also worth noting that functional differences might not be attributed to the effect of one isolated genetic variation but to specific combinations of amino acid changes.

Various mutations were detected in the HPV 18 coding and non-coding regions of VBD 17/15. A T7592C nucleotide substitution in the LCR, where the viral promoter and transcriptional control elements are located could influence transcriptional activity of the downstream viral oncogenes as has been described for HPV 16.[100] Although this change was not located within transcription binding sites, functional assays would be required to determine if the substitution has any effect on transcriptional regulation.

HPV 18 variants do not appear to be associated with differing risk for cervical cancer.[93,112] However when investigating HPV-HNSCC, sampling of precancerous lesions is challenging as there are no Pap smear equivalents for the head and neck region; thus hindering progress on associating variants with risk for cancer. Nonetheless a larger cohort of HPV 18 variants associated with HNSCC would provide a more exhaustive picture on the underlying mechanisms behind HPV-HNSCC carcinogenesis.

The most obvious finding to emerge from the analysis is that genetic diversity is spread throughout the HPV 18 genome, with higher levels of variation in the E1 and E2 genes. The E1 viral protein is mainly involved in replication of the HPV genome, whilst genetic diversity within the E2 gene might affect its role in viral replication and transcriptional regulation of the viral early genes and regulation of viral oncogene expression.[12]

Prior studies have noted the importance of substitutions of amino acids on the viral capsid (L1). The host cellular immune response may be altered across variants that differ in the L1 region, affecting vaccination efficacy. It can be speculated that if there were amino acid substitutions in the L1 of an HPV type, there might not be immunological protection against a subsequent infection with other variants of the same genotype.[129] Changes in key viral proteins, such as the L1 in this case, show that there are either neutral or diversifying selection pressure during viral evolution.[93, 124]

Another significant aspect of HPV-related cancers is integration in the human genome. In the case of cervical cancers caused by HPV 16, the virus can be integrated or remain in episomal form, whilst almost all cancers caused by HPV 18 have the virus integrated at previously described common hotspots.[46] The fact that we were able to amplify the whole genome in two overlapping fragments, suggests that this HPV 18 isolate was not integrated. However further investigation of integration in multiple patients is needed to determine if integration is a key event in HNSCC and if and how viral genetic variation might relate to integration.

Much of the knowledge on HPV is based on cervical cancer and has shed light on HPV infection of the oral cavity; however significant differences exist between these two sites, causing dissimilarities in clinical progression between the two sites. To develop an exhaustive picture of HPV associated HNSCC, multifactorial studies are needed to investigate both viral and host factors. The findings in the study may be limited by the obtaining of one complete HPV 18 genome, due to low HPV positivity in this South African cohort.[73] A larger global cohort of HPV 18 variant lineages associated with head and neck cancers can help elucidate the potential mechanisms underlying differences in carcinogenicity at the HPV type, lineage, and nucleotide levels especially in HNSCC.

Conclusion

This study describes the first complete genome analysis of HPV 18 from a nasopharyngeal carcinoma patient. Mutations novel to this isolate include an amino acid change within the L2 protein. The findings of this study contribute to our understanding of HPV 18 associated with HNSCC and lays the groundwork for future research into HPV-HNSCC genomic characterisation in Sub-Saharan Africa.

Data Availability

The datasets generated and/or analysed during the current study are available from the corresponding author.

Acknowledgements

Y Munsamy held a scholarship by the National Research Foundation (Grant UID: 94968) and funding by the Poliomyelitis Research Foundation (Grant number: 15/68). The research was supported by the National Research Foundation (Grant number: 90556).

Competing Interests Statement

The authors declare no competing interests.

Author Contributions

FB conceived and supervised the experiments; YM designed and performed the experiments; TS contributed to genotyping the isolate; PB and YM analysed the NGS data; RS contributed biopsy tissue; YM wrote the manuscript. FB revised and edited the paper.

CHAPTER 3

Complete genome sequence of HPV 31 and biological consequences for p97 promoter activity

Complete genome sequence of HPV 31 and biological consequences for p97 promoter activity

Y Munsamy

Will be submitted for consideration for publication in *Virology Journal*

Abstract

High-risk human papillomavirus type 31, although detected less frequently than types 16 and 18, is considered one of the risk factors for head and neck cancer. Information regarding HPV 31 and its association with head and neck and cervical cancers is limited. Previous studies suggest that HPV 31 polymorphisms in the long control region (LCR) may alter the oncogenic potential of the virus. This study reports the first complete genome of a South African HPV 31 isolate associated with head and neck cancer (HNC). Quasispecies harbouring minor nucleotide variations in the whole genome sequence were observed in this South African laryngeal isolate. Sequence variations relative to the HPV 31 prototype sequence were identified. The pBlue-Topo® vector, a reporter gene system was used to investigate the possible influence of these variations on the LCR promoter activity in vitro by monitoring expression of a reporter gene. Using mutagenesis to create two different fragments, β-galactosidase assays were used to monitor the effect of nucleotide variations on the p97 promoter. Increased β-galactosidase expression was observed in mutants, when compared to the South African HPV 31 LCR isolate. Enhanced transcriptional activity was observed with the mutant that possessed a single nucleotide change within the YY1 transcription factor binding site. Sequence variation within the LCR of HPV 31 isolates may have a functional effect on viral p97 promoter activity, however, increased oncogenicity may be due to a combination of factors other than sequence variation in this noncoding region.

Keywords

Human papillomavirus, HPV 31, head and neck squamous cell carcinoma, LCR, transcriptional activity, next-generation sequencing, genetic diversity, HPV quasispecies

Introduction

Human papillomaviruses (HPVs) include over 200 different types and infection by one of the 13 high-risk types increases the risk of developing cervical, anogenital, or head and neck cancers.[5,21] These small double-stranded circular DNA viruses of approximately 8 000 base pairs (bp) contain between eight to nine open reading frames (ORFs), with dual promoters. The E6 and E7 genes encode the HPV oncoproteins that promote cell cycle progression and viral DNA replication.[11] The long control region

(LCR) contains the viral promoter, p97 and transcriptional elements that regulate expression of the viral oncogenes, as well as transcription factor binding sites and the viral origin of replication.[131] HPV types are distinguished by a genetic difference of around 10% in the L1 gene sequence of the genome.[21,22] Whilst HPV 16 and HPV 18 account for the majority of HPV-associated head and neck squamous cell carcinomas (HNSCC), contribution of other oncogenic types (31/33/45/52/58) to causing disease should not be ignored.[32] Although the virus is relatively genetically stable there is differentiation into genomic variants. At the variant lineage and sublineage level, there are nucleotide differences of 1-10% and 0.5-1.0%, respectively.[17,89] HPV 16 is the closest relative of HPV 31, however there are differences in carcinogenicity between these two types.[104] HPV 31 is divided into three major variant lineages and seven sublineages, including: sublineages A1, A2, B1, B2, C1-3.[89] Across these HPV 31 sublineages, carcinogenicity and viral persistence differs, ranging from HPV 16-like to less aggressive behaviour.[132] In trying to locate genomic regions that are responsible for differences in oncogenic potential, functional significance of sequence variation among HPV 31 LCR variants showed that different variant lineages vary in transcriptional activities.[133]

With the advent in next-generation sequencing pipelines, the field of HPV genomics is rapidly advancing. There has been an expansion of obtained complete HPV 16 genomes from cervical carcinomas. Currently no complete genomes are available for HPV 31 isolates from HNSCC on GenBank. To address our research interests, the complete genome of an HPV 31 isolate from head and neck cancer was determined by next generation sequencing and the isolate characterised. A reporter gene system was used to determine if mutations identified in the noncoding LCR of an HPV 31 isolate, had any influence on biological activity. These mutations were sequentially introduced into a reporter vector expressing β-galactosidase activity in a mammalian cell culture system, to determine differences in the transcriptional activities.

Methods

Sample preparation and next generation sequencing

A laryngeal carcinoma biopsy collected from a patient in the South African province of the Free State, and histologically confirmed as a HNSCC was submitted for testing for HPV. HPV was detected using the multiplex PGMY primers targeting the L1 gene.[73,76] The isolate was designated VBD 13/14 and genotyped as HPV type 31 using bi-directional DNA sequencing and alignment with GenBank sequence data. Informed consent was obtained from the patient. Ethics approval for the study was obtained: ECUFS 137/2013D.

To obtain the complete viral genome, overlapping fragments were generated using primer pair HPV_31F1; HPV_31R1 (5'-GTCCCAAATGGTACAATGGG-3'; 5'-TTCACCAACCGTGCCTGATC-3') and HPV_31F2; HPV_31R2 (5'-TTGCAAACCACCTATTGGAG-3'; 5'-GATTTACCTGTATTAGGTGCACC-3').

All primers used in this study were designed based on an HPV 31 reference sequence (GenBank accession no J04353). To generate overlapping amplicons, Phusion® High-Fidelity DNA Polymerase enzyme mix (ThermoFisher Scientific, Massachusetts, USA) was utilized according to manufacturer's instructions. Each reaction mixture contained 10 ng template DNA, 10 µl 5x HF buffer, 2.5 µl of each primer (10 µM), 1 µl 10 mM dNTPs, 0.5 µl Phusion DNA polymerase and water to a final volume of 50 µl. The amplification was carried out as follows: 98 °C for 30 s, followed by 30 cycles of alternating 98 °C for 5 s, 64 °C or 65 °C (depending on primer pair, F2_R2 or F1_R1, respectively) for 30 s and 72 °C for 2 minutes 30 s. A final elongation of 5 minutes at 72 °C was included. PCR products were verified by gel electrophoresis on a 1% agarose gel, purified using Promega Wizard® SV Gel PCR Clean-Up System kit (Promega, Wisconsin, United States of America) according to manufacturer's instructions. MiSeq library preparation and sequencing was conducted at the University of the Free State Next Generation Sequencing Unit. The mean sequence length achieved was 170.17 ± 48.03 bp, the minimum length was 35 bp, whilst the maximum length was 201 bp. The length range was 167 bp whilst the mode length was 201 bp with 168 894 sequences and an average coverage of 400 x. The primer sequences were removed from the acquired nucleotide sequences. The acquired nucleotide sequences were assembled and analysed using HPV 31 prototype reference (J04353) retrieved from GenBank, originally isolated from a cervical dysplasia.[134]

Phylogenetic Analysis and Variant Lineage/Sublineage Identification

Complete genome sequence data for 23 isolates from cervical carcinoma were retrieved from GenBank (Accession numbers are available in Appendix D). The evolutionary history was inferred by using the Maximum Likelihood method based on the Tamura-Nei model.[121] The bootstrap consensus tree inferred from 1000 replicates is taken to represent the evolutionary history of the taxa analysed. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 0,0559)). The analysis involved 126 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated. There were a total of 7761 positions in the final dataset. Evolutionary

analyses were conducted in MEGA7.[122] Variations within VBD 13/14 were identified by alignment with the reference isolate.

Amplification of HPV 31 LCR

Using the complete genome sequence data for VBD 13/14, additional primers were designed to amplify the full length HPV 31 LCR (Table 3.1), to generate a fragment from nucleotide (nt) position 7069 to 180. Phusion® High-Fidelity DNA Polymerase enzyme mix (ThermoFisher Scientific, Massachusetts, USA) creates blunt ends, therefore, A-tailing of the PCR product was carried out by incubation of the product in 0.2 mM dATP solution (New England BioLabs, Massachusetts, USA) for 30 minutes at 70 °C. The PCR product was then dephosphorylated by incubation in the presence of 5 U Antarctic phosphatase (New England BioLabs, Massachusetts, USA) for 30 minutes at 37 °C. The reaction was stopped by incubation for 10 minutes at 70 °C.

Construction of reporter plasmid

A β-galactosidase reporter construct was prepared using the pBlue-Topo® promoterless vector (ThermoFisher Scientific, Massachusetts, USA), with the VBD 13/14 LCR amplicon containing the p97 promoter. For ligation and the following transformation into *E. coli* cells, the ratio of insert to vector was 1:1 and the procedure was carried out according to manufacturer's instructions. The LCR fragment was placed upstream of the β-galactosidase reporter gene. Before conducting site-directed mutagenesis, the pBlue-Topo® LCR clones were sequenced to verify the insertion of the LCR fragment. The confirmed plasmid was designated pBlue_VBD13/14. Site-directed mutagenesis of the insert in pBlue_VBD13/14 was performed to introduce specific mutations.

Site-directed mutagenesis

Site-directed mutagenesis by PCR was performed to introduce a specific insertion and to introduce a mutation in the LCR region in pBlue_VBD13/14. A schematic representation of the full-length LCR clones is depicted in Figure 3.1. A 10 bp mutation absent in HPV 31 lineage B was inserted (5'-TACTATTTTA-3'), using primers described in Table 3.1. The resultant clone was designated pBlue_SDM1. The sequence variation at a Yin-Yang 1 (YY1) transcription factor binding site, was introduced using primers containing the T-C transition. This T nucleotide at position 7564 is a lineage B associated SNP, whilst lineages A and C contain a C at that position. This clone was designated pBlue_SDM3.

Site-directed mutagenesis was performed using Phusion® High-Fidelity DNA Polymerase enzyme mix (ThermoFisher Scientific, Massachusetts, USA), for mutagenic primer-directed replication of both plasmid strands. The final concentration of each primer was 0.5 µM, whilst in the reaction to construct pBlue_SDM3, a final concentration of 0.13 µM was used. The basic procedure utilised the pBlue-

Topo® vector with the LCR fragment insert (pBlue_VBD13/14) and two synthetic oligonucleotide primers, one of which contains the desired mutation. The amplification profile consisted of an initial incubation at 98 °C for 30 s, followed by 30 cycles of alternating 98 °C for 5 s, 53 °C or 65 °C (depending on primers used) and elongation at 72 °C for 4 minutes 30 s. A final elongation of 10 minutes at 72 °C was included. Amplification was verified by separation of PCR products by electrophoresis on a 1% agarose gel.

Following thermocycling, the products were treated with *Dpn*I endonuclease, which is specific for methylated and hemimethylated DNA in order to digest the parental DNA template and select for mutation-containing synthesised DNA. The reaction was set up as follows: 45 µl PCR product, 5 µl 10x *Dpn*I buffer, 1 µl *Dpn*I (New England BioLabs, Massachusetts, USA) for 16 hours at 37 °C. The amplicons were electrophoresed on 0.8% agarose gel and purified using Promega Wizard® SV Gel PCR Clean-Up System kit (Promega, Wisconsin, USA) according to manufacturer's instructions. The purified product was eluted in 30 µl nuclease free water, dried completely at 60 °C in an Eppendorf concentrator plus (Hamburg, Germany) for 15 minutes, then resuspended in 13 µl nuclease free water. For a one-step phosphorylation and ligation, linear template (13 µl) was ligated with 1,5 µl T4 ligase, 2 µl ligase buffer and phosphorylated with 2 µl 10 mM ATP (New England BioLabs, Massachusetts, USA) and 1,5 µl polynucleotide kinase (New England BioLabs, Massachusetts, USA) in one reaction, for 16 hours at 4 °C. Top10 *E. coli* cells (Invitrogen, California, USA) were transformed with the mixture, plasmid-grade DNA was extracted and a restriction digest was performed with *Bam*HI (New England BioLabs, Massachusetts, USA) and *Hind*III (New England BioLabs, Massachusetts, USA) to select for positive transformants. After obtaining positive clones, the constructs were validated by Sanger sequencing.

Table 3.1. Properties of primers to amplify HPV 31 LCR and to conduct site-directed mutagenesis on the reporter plasmid.

Primer purpose	Primer name	Primer sequence	Annealing temperature (°C)	Position relative to South African HPV 31 isolate (VBD 13/14)	Expected amplicon size
LCR fragment	HPV31_LCR F	5'-CATGTGTCTGTATGTGTATG-3'	51	7069-7088	~989 bp
	HPV31_LCR R	5'-CATCGTAGGGTATTTCCAATG-3'		180-160	
Site-directed mutagenesis	HPV31_SDM1 F	5'-TAAACTATTGTTCCTACTTGTCC-3'	53	7282-7304	~8800 bp
	HPV31_SDM1 R	5'-TAATAGTATGTTACTAATAGGGT-3'		7261-7283	
	HPV31_SDM3 F	5'-CATGCTAGTACAACTATGCTGATACAG-3'	65	7454-7484	~8800 bp
	HPV31_SDM3 R	5'-TTTAAACAATGCAACCGAAAA-3'		7466-7446	

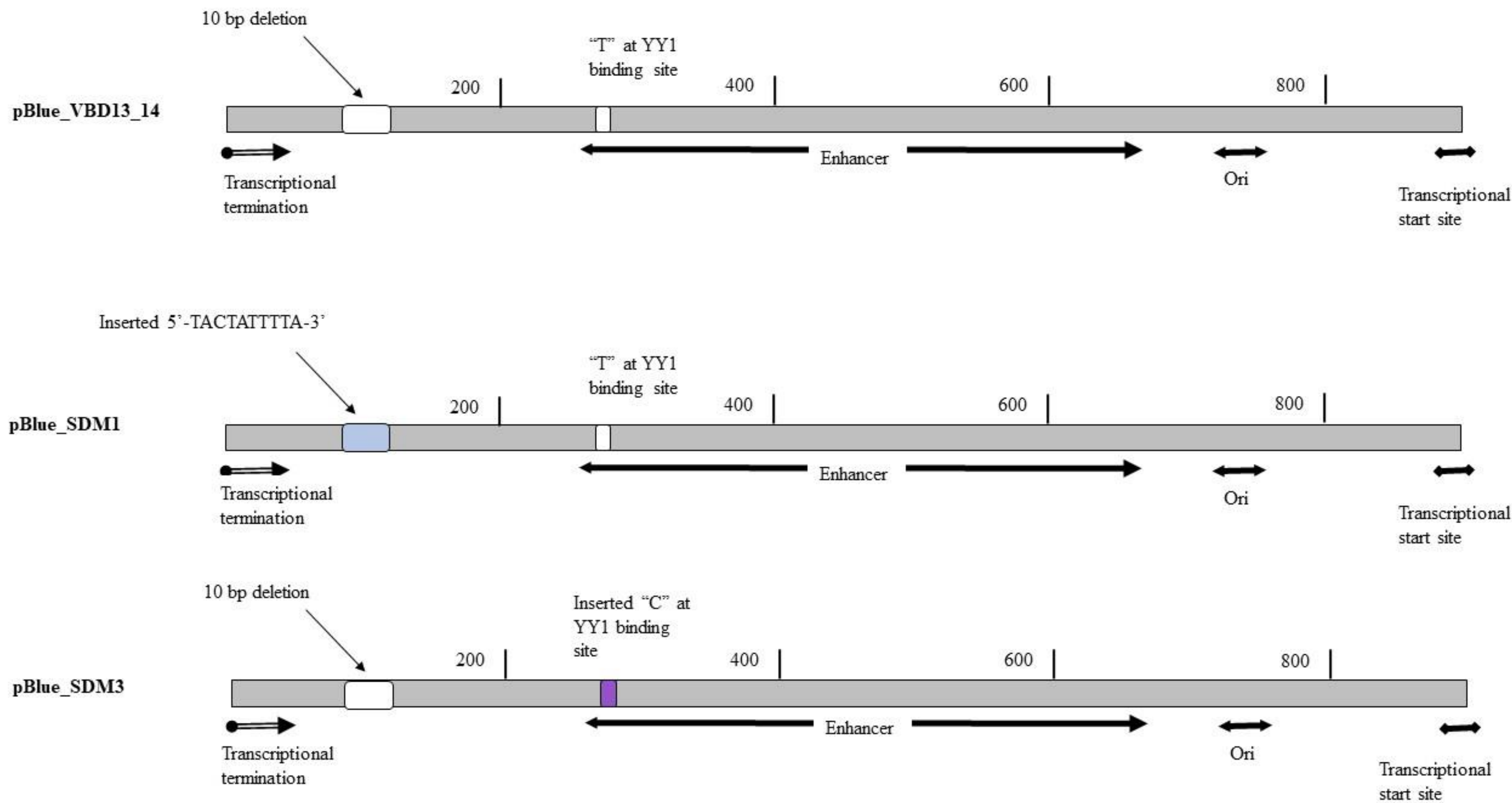


Figure 3.1. Schematic representation of the full-length LCR variants cloned upstream of the β-galactosidase gene in the reporter vector pBlue-Topo®. YY1 binding site, Yin-Yang 1 binding site; Enhancer, keratinocyte-specific enhancer domain; Ori, origin of replication of the HPV 31 circular genome.

Cell culture

Baby hamster kidney 21 (BHK) (ATCC ® CCL-10™) cells were grown in cell culture flasks in an incubator with 95% relative air humidity and 5% CO_2 in Dulbecco's Modified Eagle Medium (DMEM) (Lonza, Verviers, Belgium) with 10% foetal bovine serum (FBS) (Delta products, Johannesburg, South Africa), 1% L-Glutamine (L-Glut) (Sigma Aldrich, Ayrshire, United Kingdom), 1% non-essential amino acids (NEAA) (Lonza, Verviers, Belgium), and 1% penicillin/streptomycin (Pen/Strep) antibiotics (Sigma Aldrich, Ayrshire, United Kingdom). According to their doubling time, cells were passaged every 3–5 days to keep it in the logarithmic growth phase. For passaging, the culture media was removed and the cells were washed with 1 x phosphate buffered saline (PBS). Cells were dissociated from the culture flask surface using trypsin (Lonza, Verviers, Belgium) and incubating the cells at 37 °C for 5 minutes. Trypsinized cells were resuspended in fresh, pre-warmed media and transferred into new flasks, at a seeding rate of approximately 1 x 10^5 cells/ml.

Transfection of BHK cells

BHK cells were seeded at 1 x10^5 cells/ml in a 24 well plate to be 80% confluent at transfection. Untransfected BHK cells were used as negative controls. The transfection mixture contained 1 µg plasmid DNA (pBlue_VBD13/14, pBlue_SDM1 or pBlue_SDM3), 1.5 µl Lipofectamine™ 3000 reagent (Invitrogen, Karlsbad, USA) and 2 µl P3000. Transfection was performed according to the recommendations of the manufacturer. Each transfection experiment was carried out independently at least three times.

Transfection efficiency determine by β-galactosidase staining

Transfected cells were stained for β-galactosidase activity according to the manufacturer's instructions (β -galactosidase staining kit, Mirus, Madison, WI) 48 h post transfection. All reagents were provided within the kit. Staining was conducted using 5-bromo-4-chloro-3-indolyl-b-D-galactopyranoside (X-Gal) for LacZ activity. Washed cells were prefixed in 0.2% glutaraldehyde solution at room temperature for 5 min. The cells were washed three times with PBS pH 7.4 and then incubated in cell staining working solution in a moist chamber and protected from light at 37 °C overnight (16 h). After the incubation the cells were washed three times with PBS and stained cells visualised using an Olympus CKX53 microscope at 10x magnification.

Transcriptional activity measured by β-galactosidase assay

Forty-eight hours post-transfection, cells were harvested, lysed and incubated in assay reagent according to the manufacturer's instruction. Levels of active β-galactosidase expressed from BHK cells transfected with plasmids expressing *lacZ* gene were determined using the β-galactosidase assay kit according to the instructions of the manufacturer (Invitrogen, California, USA). The protein

concentration of the cell lysate was determined using the Qubit™ protein assay according to the manufacturer's instructions (Invitrogen, California, USA). The specific activity of the cell lysate, determined in a total volume of 8×10^5 nanolitres, was calculated as follows: specific activity = nmoles of ONPG hydrolyzed/t/mg protein, (where nmoles of ONPG hydrolyzed = (OD_{420}) (8×10^5 nanolitres) / (4500nl/$nmoles^{-cm}$) (1cm), and where 4500 = the extinction coefficient, t = the time of incubation in minutes at 37 °C, and mg protein = the amount of protein assayed.

Results

Complete genome sequence and phylogenetic analysis of HPV 31 VBD 13/14

HPV was detected using the multiplex PGMY primers targeting the L1 gene. The isolate was designated VBD 13/14 and genotyped as HPV type 31 using bi-directional DNA sequencing and alignment with GenBank sequence data.

The first complete genome of an HPV type 31 isolate from HNSCC, detected in a patient in South Africa (isolate VBD 13/14), was determined by NGS and characterised. Briefly, the complete genome of VBD 13/14 was 7877 bp in length with a GC content of 37.2%. Annotation of the complete HPV 31 isolate revealed eight coding regions (E6, E7, E1, E2, E4, E5, L2, L1) and two noncoding regions (intergenic E2/E5 region and LCR).

There was no sequence data available for HPV 31 isolates from HNSCC on GenBank. To determine the genetic relationship of the SA isolate and all available complete HPV 31 genomes from cervical cancer from other parts of the world submitted previously to the GenBank, a phylogenetic tree was constructed using Geneious version 2019.0 (Biomatters) and classified to lineages according to Chen *et al*.[126] Accession numbers for isolates from cervical carcinomas retrieved from GenBank available in Appendix D.

Isolates within lineage A formed two distinct sublineages, designated A1 and A2. Sublineage A2 included HPV 31 isolates with sequences closely related to the sequence of an isolate from cervical carcinoma from Thailand (GenBank accession no. HQ537675.1). Lineage B was subdivided into two sublineages (B1 and B2) (Figure 3.2). Lineage C was divided into three sublineages (C1-3). The South African isolate clustered in sublineage B2 and was dissimilar to other isolates available on GenBank.

Comparison of the available complete sequences shows no distinct SNPs exclusive to HPV-HNSCC. Sequence variations between VBD 13/14 and the reference HPV isolate were then investigated. There were a total of 59 sequence variations between VBD 13/14 and the HPV 31 prototype reference; comprising of 40 transitions, 17 transversions and two 10 bp deletions at positions 3998 (E5 gene) and

7298 (LCR). However a novel 10 bp deletion (GCTTTTGCCA) was observed in the E5 gene of VBD 13/14. The deletion in the E5 gene has not been described previously. Visualisation of deep sequence reads with Integrative Genomics Viewer (IGV) showed that 70% of the reads contained an E5 deletion, suggesting a co-infection of HPV 31 viral variants. This deletion led to a frame-shift, although the functionality of the protein is not known.

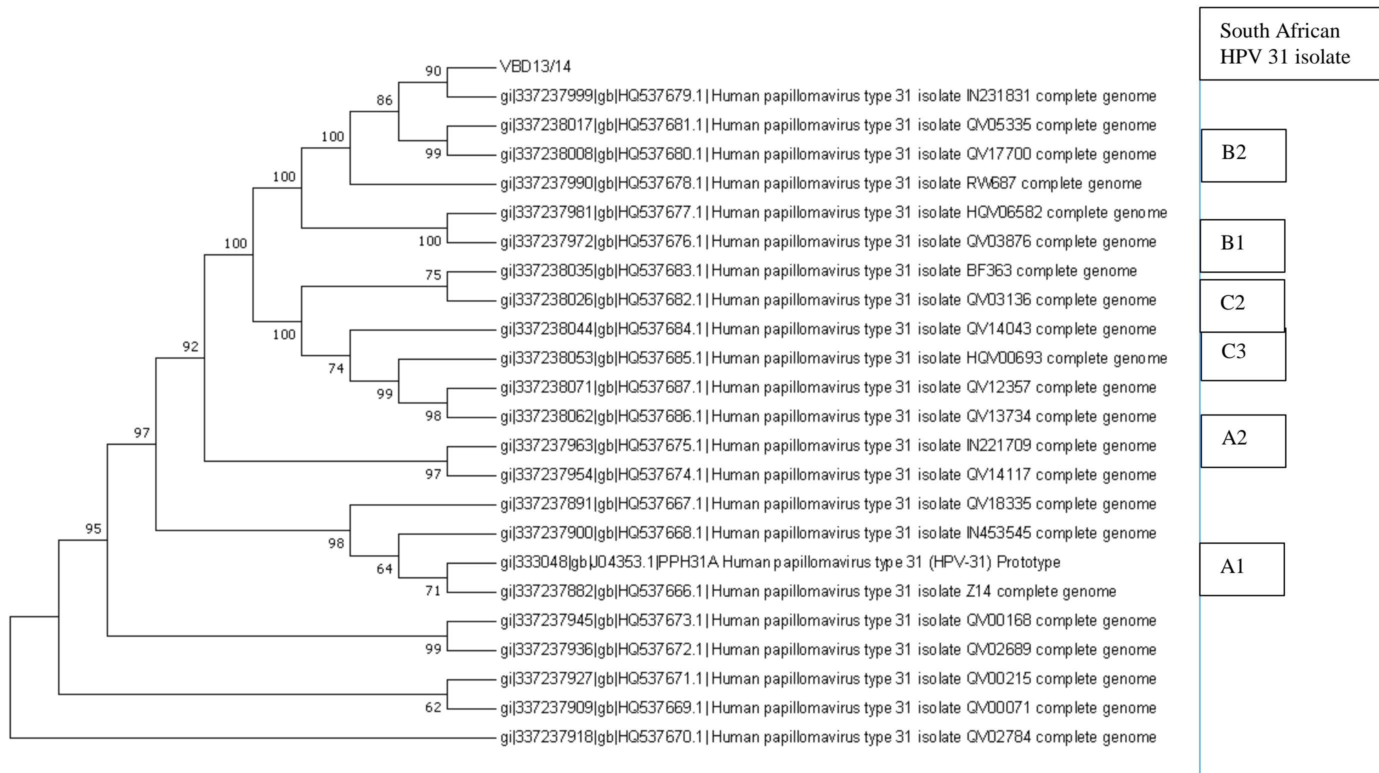


Figure 3.2. The maximum likelihood tree was inferred from a global alignment of complete sequences of HPV 31 isolates from cervical carcinoma and VBD 13/14 isolated from HNSCC. A bootstrap value of 1000 replicates was employed. Tree constructed with MEGA 7.0. Each isolate is represented by a GenBank accession number.

In order to explore natural nucleotide sequence variation in the LCR of HPV 31, mutations were selected representing lineage A on the basis of phylogenetic analysis. Altogether 16 changes were identified in the LCR, relative to the prototype as detailed in Table 3.2.

Table 3.2. Sequence differences in the noncoding LCR relative to the prototype HPV 31.

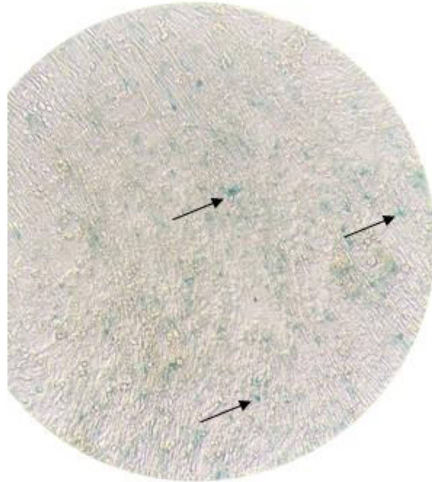
HPV Prototype* nucleotide position	VBD 13/14	Transcription factor binding site affected
7355-7364 TACTATTTTA	Deletion	None
7412 A	G	None
7430 G	C	None
7442 G	A	None
7452 C	A	None
7507 GA	AC	None
7515 G	A	None
7532 C	T	None
7564 C	T	Yin-Yang 1
7583 G	A	None
7633 T	C	None
7768 C	T	None
7812 C	A	None
7923 T	G	None

**(accession number: J04353.1)*

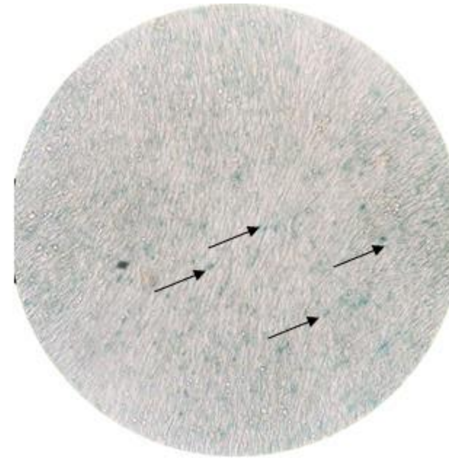
Transfection efficiency determined by β-galactosidase staining

To localize nucleotide alterations that could potentially affect transcriptional activity, mutants were constructed to reflect changes identified relative to the prototype HPV 31. These constructs were used to transfect BHK cells and the transfection efficiency of BHK cells was evaluated by X-gal staining. A higher transfection efficiency of approximately 60% was observed in the cells transfected with pBlue_SDM3 compared with an approximate 10% transfection efficiency of cells transfected with pBlueVBD13_14 and pBlue_SDM1. The untransfected BHK cells showed slight staining due to endogeneous β–galactosidase present in cells. This was subsequently accounted for in the activity assays by subtracting the endogenous staining background.

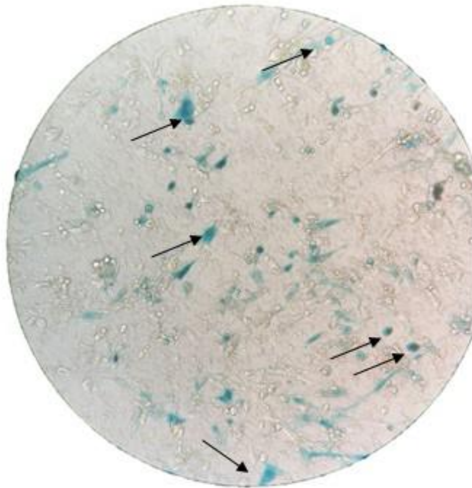
pBlue_VBD13_14



pBlue_SDM1



pBlue_SDM3



Untransfected

Figure 3.3. β-galactosidase staining for determination of transfection efficiency of transfected BHK cells. BHK cells transfected with pBlueVBD13_14, pBlue_SDM1 and pBlue_SDM3 and stained for β-galactosidase. Untransfected BHK cells with slight staining that shows endogeneous β-galactosidase activity. Black arrows indicate some of the stained transfected BHK cells per field.

Transcriptional activity of HPV 31 LCR mutants

Several nucleotide changes were observed in the South African HPV 31 LCR when compared to a reference isolate (Genbank Accession number J04353.1). To determine if selected nucleotide changes resulted in altered transcriptional activity of the HPV 31 LCR, the South African HPV 31 LCR was cloned into a promoterless β-galactosidase reporter vector pBlue-Topo®. The resultant construct was sequenced to rule out PCR-generated mutations and subsequently site-directed mutagenesis was used to generate constructs with changes in the LCR that were representative of the lineage A reference isolate. All constructs were sequenced to confirm the presence of the selected mutations and then used to transfect BHK cells.

In summary three constructs were prepared. pBlue_SDM1 incorporated the following nucleotides, TACTATTTTA at position 7355. pBlue_SDM3 incorporated a C nucleotide at position 7564 whilst pBlue_VBD13/14, represented the native SA HPV 31 LCR, associated with lineage B. Comparing the specific activity detected in cells transfected with plasmids pBlue_SDM1, pBlue_SDM3 and pBlue_VBD13/14, there was a higher activity in cells transfected with the plasmid containing a single nucleotide change in a YY1 binding site (pBlue_SDM3) (Figure 3.4). Overall, the average β-galactosidase activity expressed from the *lacZ* gene of pBlue_SDM1 and pBlue_SDM3 displayed similar levels of β-galactosidase expression which was almost double that of pBlue_VBD13/14 (Figure 3.4). Tables of raw data to calculate specific activity and table of mean and standard deviation are available in Appendix E.

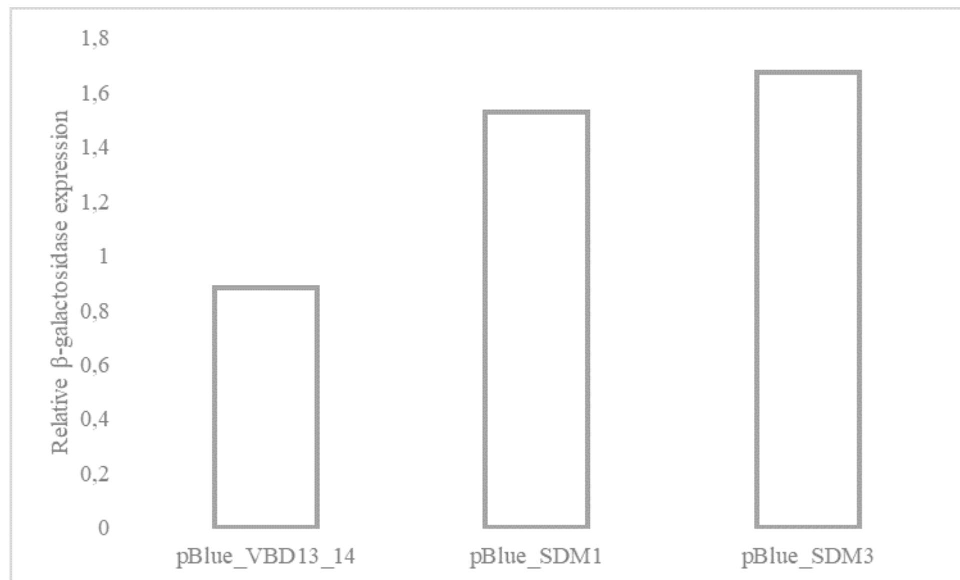


Figure 3.4. Transcriptional activity of HPV 31 full-length LCR variants. Relative β-galactosidase activities of BHK cells transfected with reporter constructs containing different HPV 31 LCR variants. Data shown represent the means of at least three independent transfection experiments.

Discussion

HPV 16 is considered the most carcinogenic HPV type as it is associated most frequently with cervical carcinomas. What differentiates the closely related HPV 31 from HPV 16 in terms of oncogenicity is unclear. Sequence variation in the E6 and E7 oncogenes and in the LCR, which controls their transcription, may have functional significance. Intratypic variation complicates the study of differences in carcinogenic risk of high-risk HPV infections.

This study presents the first whole genome sequence of an HPV 31 isolate from a head and neck cancer patient, adding to the limited number of complete HPV 31 genomes available on GenBank.

The topology of the HPV 31 phylogenetic tree confirmed the presence of three lineages. As lineage-specific SNPs were not evenly dispersed throughout the HPV 31 genomes analysed, this study further supports using complete genome sequencing for HPV variant classification. Phylogenetic analysis of isolate VBD 13/14 showed that the isolate clusters in lineage B. Limited nucleotide sequence divergence was observed between the three HPV 31 lineages. Phylogenetic analysis shows that HPV 31 acquired minor nucleotide sequence changes, which over time led to the differentiation of the lineages. Figure 3.2 demonstrates that although lineage C is the most diverse, lineages A and B are more prevalent.

As HPV isolates have coevolved with humans, some isolates may persist in certain individuals based on their genetic background.[4] Although the geographical and ethnic association of HPV 16 and HPV 18 have previously been described, these relationships for HPV 31 are not well established. In cervical carcinoma with Caucasian and African-American women as the majority, infections with HPV 31 lineage B variants, like the one identified in this study, were more likely to resolve than those with lineage C variants. However, the mechanisms involving variant-specific race-associated clearance is not yet known.

Although the DNA-based HPV genome is generally regarded as stable, high level of variations across histological grades and HPV types have been observed in cervical carcinoma studies. Within-host variations of the HPV genome have been described in cervical carcinoma, however, the impact on cancer development is unclear. In this study, it is postulated that the individual was co-infected with two HPV 31 lineage B variants. A mixed infection of similar variants may change clinical aspects of the virus. Comparison of VBD 13/14 with HPV 31 isolates from cervical carcinomas was performed to identify SNPs exclusive to HPV-HNSCC. The deletion in the E5 gene has not been described before and may represent a distinct SNP exclusive to HPV-HNSCC that requires a larger cohort for further investigation.

Ultra depth of sequence reads from NGS yielded high resolution, allowing quantification (70% of the reads contained this polymorphism) of variants in this mixed infection. Low quality reads had been discarded before initial analysis and a high read depth with around 2000 x coverage at this position was achieved, although further confirmation using Sanger sequencing may be warranted.

This deletion led to a frame-shift, which would most likely result in a non-functional protein. The primary target of E5 is a membrane protein or receptor that then acts to alter the levels or activities of cell cycle regulators. In addition, the E5 may play a role in retaining proliferative activity following differentiation, in order to produce high volumes of virus.[135] These findings raise questions about whether this deletion poses a selective advantage or not for this particular variant. Future studies should extend to investigating infectivity and pathogenicity of HPV 31 E5 deletion variants.

In locating regions of the HPV 31 genome that would influence oncogenicity, this study investigated the genetically variable LCR. The LCR regulates transcriptional activity of the viral oncogenes, thus variation in this region may result in altered expression of the oncogenes. Site-directed mutagenesis used to construct variants with changes that are associated with the prototype sequence (lineage A) allowed for comparison of reporter gene activity. Increased β-galactosidase expression was observed in the mutant constructs compared to the VBD 13/14 LCR. The highest level of β-galactosidase expression was observed with pBlue_SDM3. The C7564T change occurred at a YY1 binding site. YY1 physically

interacts with a number of proteins regulating cell proliferation and apoptosis, including p53, Mdm2, and RB, all of which play a role in modulating tumourigenesis.[136] In certain HPV 16 isolates, point mutations or deletions of YY1 binding sites in the LCR were found to have enhanced transcriptional activity.[100] The SNP and indel are representative of lineages A and C, respectively. From these results, it can be suggested that lineages A and C might have a higher expression of viral oncogenes than lineage B. In cervical cancer studies, HPV 31 lineage C is more persistent than A and/or B.[132,137] However, HPV 31 lineages A/B are more commonly associated with the development of cervical intraepithelial neoplasia 3 (CIN3).[132] Precursor lesions have not been identified in head and neck cancers, therefore it is not known which HPV 31 lineages are associated with persistence or are more carcinogenic in the context of the upper aerodigestive tract.

It is important to note that variation within the LCR is not the only mechanism which leads to HPV types and intratypic variants having different oncogenic capacities. Natural sequence variation of the HPV 31 E6 protein may be involved in the observed differences in the oncogenic potential between HPV 31 variants.[95] In addition, combinations of amino acid changes with the oncoproteins as well as host factors may also influence oncogenic capacity.

The findings of this study are limited by the low prevalence of HPV associated with HNSCC in this population; comparative studies could not be completed on a large cohort of naturally occurring HPV 31 variants. Further studies should create deletion mutations, to identify specific regions that are responsible for increased transcriptional activity rather than isolating single changes.

Conclusion

This study reports the first complete genome of a South African HPV 31 isolate associated with HNC. In addition, the coexistence of two closely related HPV 31 quasispecies was found in a head and neck cancer patient. Results suggest that natural variation in the LCR of HPV 31 isolates has a functional effect on viral p97 promoter activity. The study contributes to our understanding of the multi-factorial nature of the oncogenic potential of HPV variants and lays the groundwork for pinpointing specific genomic regions that are responsible for increased oncogenicity.

Funding

Yuri Munsamy was supported by the National Research Foundation, and the Poliomyelitis Research Foundation. The project was funded by the National Research Foundation.

Competing interests

The authors have no competing interests to declare.

Author Contributions

FB conceptualised and supervised the study; YM designed and performed the experiments and analysed and interpreted the results; TS contributed to genotyping the isolate; PB and YM analysed the NGS data; RS provided biopsy tissue; YM wrote the manuscript. FB revised and edited the paper.

CHAPTER 4

The role of *p53* mutations in HPV-associated HNSCC

The role of *p53* mutations in HPV-associated head and neck squamous cell carcinomas

Y Munsamy

Will be submitted for consideration for publication in *Cancer Epidemiology, Biomarkers & Prevention* or *Journal of Cancer Research and Clinical Oncology*

Abstract

Although *p53* mutations in human cancers have been extensively studied, the significance and role of *p53* in the aetiology of head and neck squamous cell carcinomas (HNSCC) is not completely understood. The involvement of high-risk human papillomaviruses in HNSCC is an alternative mechanism to inactivate the p53 protein function. To investigate *p53* mutations in HNSCC, DNA extracted from 25 histologically confirmed HNSCC was amplified using primers spanning and targeting exons 4-9. In addition, p16 immunostaining as an indication of HPV infection was performed on each tissue section. Exposure to alcohol and smoking as well as HPV status were assessed. High-risk HPV was detected in 4/25 (16%) tumours. p16 as an indirect marker of HPV infection did not correlate well with HPV detection using PCR. Sequencing confirmed 13 *p53* mutations in nine tumours; with a tendency towards G:C > T:A changes. The functional significance of the intronic mutations observed within *p53* should be further investigated. *p53* mutation frequencies were significantly lower than expected. It is unknown whether the genomes of participants of African descent are too diverse from the reference genome to accurately use frequency and functional data of known mutations.

Keywords

p53, head and neck squamous cell carcinoma, human papillomavirus, mutation

Introduction

Tumours of the head and neck are associated with exposure to carcinogenic agents, such as tobacco or oncogenic viruses such as human papillomavirus (HPV) infection.[6,138] An increased understanding of the genetics underlying head and neck cancers has led to HPV-associated carcinomas being classified as a distinct molecular subgroup of HNSCC. The *p53* gene functions as a tumour suppressor and is responsible for cell cycle arrest or apoptosis.[57] Of the 50-100 candidate cancer driver genes involved in head and neck squamous cell carcinomas (HNSCC), *p53* is the most frequently mutated.[50] Between 46-73% of HNSCC cases contain mutations.[55,56]

Both protein-based and DNA-based methods can be used to detect *p53* mutations. Wild type and mutant *p53* differ in expression levels, which is what immunohistochemistry capitalises on to detect mutations. However it is unable to detect nonsense or frameshift mutations that result in p53 protein truncation.[131]

There are 11 exons and 10 introns that span the 20-Kb *p53* gene, located on chromosome 17p13.[132] The majority of studies analyse exons 5-8, 5-9 or 4-9 by DNA sequencing and a limited number of studies include all exons.[133] The central portion of the p53 protein, encoded by exons 5-8 is known as the DNA binding domain (DBD). In order to carry out the tumour suppressor function, this region needs to retain its tetrameric conformation. The majority of mutations are located within this thermodynamically and kinetically unstable domain. Oncogenic mutations lower the stability of the DBD even further as a result of structural changes, causing it to rapidly unfold at body temperature while retaining native conformation at sub-physiological temperature.[142,143]

Varying levels of disruption of the protein function are observed, depending on where mutations are located.[135] Missense mutations are expected to have a loss of function effect on the p53 protein. Mutations that occur within the core domain may completely disrupt p53 protein function. As in other human cancers, missense mutations primarily within the DBD account for 75% of all mutations in the *p53* gene and confer both dominant negative and poorly understood gain of function properties.[57,145]

Hotspots have been described for specific cancers with regard to splice site mutations and single base deletions. For example, laryngeal tumours frequently harbour mutations within exon 5 that affect the S2' protein domain [codons 132-135]. Although the L2/L3 region does not affect the protein structure when mutated, it does affect DNA binding capacity.[137] It is speculated that other mutations outside these DNA contact points may not have as severe repercussions.[135]

In HNSCC tumours that do not harbour *p53* mutations, p53 function may be inactivated by various other mechanisms.[138] The *p53* pathway is a target in HPV-related malignancies but is almost never mutated. *p53* is degraded by a trimeric complex, composed of *p53*, E6 and host E6 associated protein (E6AP).[39] High-risk HPV E6 can degrade *p53* even in the absence of E6AP. The inactivation of *p53* compromises the integrity of the replicated DNA and causes DNA damage and chromosomal instability, cell proliferation or tumour development.[139,140] E6 degradation of *p53* protein is not functionally equivalent to a *p53* mutation, therefore HPV is incidental and not causal in the development of these HNSCCs.[135] There is a frequent underestimation of *p53* mutations in HPV-HNSCC, although the dual presence of HPV DNA and *p53* mutations ranges from 0% to 42%. For HNSCCs with nondisruptive *p53* mutations, transcriptionally active HPV could heighten the overall effect of abrogation in p53 function.[135]

p53 status is used as prognostic and predictive marker in various cancers. Wild type *p53* status may be a universal marker for a better outcome for all HNSCCs.[141] HPV-positive HNSCCs that arise in the oropharynx, are less commonly associated with tobacco or alcohol exposure, show enhanced sensitivity to radiation therapy and are consistently associated with favourable patient outcomes compared with non–HPV-related HNSCCs.[135]

Another tumour suppressor, p16, is a surrogate marker of high-risk HPV infection as it is overexpressed by functional inactivation of RB. When the tumour is not HPV-driven, staining of p16 is weak or absent.[5] It is recognised as a prognostic marker as patients that are p16-positive have significantly better outcomes than patients with p16-negative disease.[36]

There were two primary aims of this study: 1. To investigate *p53* mutations in a subset of HNSCC samples 2. To ascertain the relationship between HPV status and *p53* mutations and p16 immunostaining.

Materials and Methods

Samples

Twenty five tumour samples from histologically confirmed HNSCCs were processed with the intent of sequencing the *p53* gene. Samples were selected based on anatomical location; the oropharynx. An additional four samples were included based on known HPV-positive status from other regions of the head and neck. Clinical characteristics of participants is shown in Table 4.1. Smoking and alcohol exposure was obtained through a questionnaire in a related study. Ethics approval for the study was obtained: ECUFS 137/2013D.

HPV detection and p16 staining

DNA was extracted from fresh biopsy tissue using the QIAamp DNA Mini Kit (QIAGEN, California, USA) according to manufacturer's instructions. For detection of HPV DNA, isolated DNA was tested with a consensus PCR targeting the L1 region of the viral genome using PGMY primers and sequencing of the PCR product to genotype the isolate.[76] All specimens that were found to contain an identifiable high-risk HPV type were genotyped. Tissue sections were also stained for p16 expression.[73]

p53 sequencing

A *p53* gene fragment, encompassing exons 4–9, was amplified from purified tumour DNA by PCR. Nested PCR was performed using two pairs of primers to obtain amplicons for sequencing. Primers utilised are described in Table 4.2.[133,142] PCR reactions were performed using GoTaq®flexi kit (Promega, Madison, USA) according to manufacturers' instructions. The PCR reactions were cycled using the following conditions: initial denaturation, 95 °C for two minutes followed by 30 cycles of denaturation at 95 °C for 30 s, specified annealing temperature for 30 s and elongation at 72 °C for one minute and final elongation at 72 °C for seven minutes, with a hold at 4 °C. The amplified products were visualised by agarose gel electrophoresis (1% gel). Positive amplicons from the nested PCR were purified from gel with the Promega Wizard® SV Gel PCR Clean-Up System kit (Promega, USA) for subsequent nucleotide sequence analysis.

Bi-directional sequencing of the amplicons was performed with both the nested PCR primers and sequencing primers (Table 4.2) using the BigDye® Terminator v3.1 Cycle Sequencing kit (Applied

Biosystems). PCR products were analysed by an 8-capillary automated sequencer (ABI PRISM® 3100 Genetic Analyzer, Applied Biosystems), based on the Sanger method. Positions of all primers are illustrated in Figure 4.1. A portion of the samples were resequenced to rule out Sanger sequencing as a source of error when a mutation was detected, although not all samples were resequenced due to time constraints.

p53 mutation analysis

The nucleotide sequence of the *p53* gene was edited with Geneious Prime and aligned with the reference sequence, from NC_000017.10 (hg38) on chr17:7687599...7668459, from *p53* 5'UTR to 3'UTR (reverse base).[136] Electropherograms showing SNP positions are available in Appendix F. Variations were analysed with the International Agency for Research on Cancer (IARC) TP53 database (R19, August 2018 version) to check whether the variation was a known polymorphism and to possibly obtain frequency and functional data for each variation.[58]

Table 4.1. Clinical characteristics of the head and neck cancer patients, including HPV status, p16 immunostaining and exposure to alcohol and smoking.

VBD No	Age	Sex	Site	Histology	HPV status	p16 status	Extent of p16 immunostaining	Smoking status	Alcohol
13/14	74	M	Larynx	Carcinoma *in situ*	HPV 31	Positive	>90%	No	Yes
24/14	59	M	Oropharynx	Moderately-differentiated SCC	Negative	Positive	60%	Yes	Yes
30/14	58	M	Oropharynx	Moderately-differentiated SCC	Negative	Negative	N/A	Yes	Yes
47/14	56	M	Tonsil	Moderately-differentiated SCC	HPV 16	Positive	Not recorded	Yes	Yes
50/14	58	F	Tonsil	Moderately-differentiated SCC	Negative	Positive	>90%	Yes	Yes
12/15	60	M	Oropharynx (tongue base)	Moderately-differentiated SCC	Negative	Positive	60%	Yes	Yes
17/15	39	M	Nasopharynx	Poorly differentiated SCC	HPV 18	Positive	70%	No	No
27/15	76	M	Oropharynx	Moderately-differentiated SCC	Negative	Negative	N/A	Yes	Yes
30/15	62	M	Oropharynx	Well-differentiated SCC	Negative	Negative	N/A	Yes	Yes
56/15	55	M	Oropharynx	Well-differentiated keratinising SCC	Negative	Negative	N/A	Yes	Yes
58/15	71	M	Oropharynx	Moderately-differentiated SCC	Negative	Negative	N/A	Yes	Yes
61/15	85	M	Oropharynx (posterior pharyngeal wall)	Moderately-differentiated SCC	Negative	Positive	80%	Yes	Yes
64/15	51	M	Oropharynx (tonsil)	Moderately-differentiated SCC	Negative	Negative	N/A	Yes	Yes
72/15	55	M	Oropharynx	Moderately-differentiated SCC	Negative	Negative	N/A	Yes	Yes

VBD No	Age	Sex	Site	Histology	HPV status	p16 status	Extent of p16 immunostaining	Smoking status	Alcohol
74/15	58	M	Oropharynx	Poorly differentiated SCC	Negative	Negative	N/A	Yes	Yes
04/16	62	M	Oropharynx	Moderately-differentiated SCC	Negative	Negative	N/A	Yes	Yes
13/16	56	F	Oropharynx	Moderately-differentiated SCC	Negative	Positive	60%	Yes	Yes
24/16	54	M	Oropharynx (tongue base)	Moderately-differentiated SCC	Negative	Negative	N/A	Yes	Yes
38/16	62	M	Oropharynx	Moderately-differentiated SCC	Negative	Positive	>90%	Yes	Yes
62/16	54	M	Oropharynx	Moderately-differentiated SCC	Negative	Positive	90%	Yes	Yes
67/16	47	M	Oropharynx	Moderately-differentiated keratinising SCC	Negative	Negative	N/A	Yes	Yes
09/17	37	M	Larynx	Carcinoma in situ	HPV 16	Positive	100%	Yes	Yes
26/17	64	M	Oropharynx	Moderately-differentiated SCC	Negative	Negative	N/A	Yes	Yes
44/17	47	M	Tonsil	Well-differentiated SCC	Negative	Negative	N/A	Yes	Yes
08/18	80	M	Tongue base	Poorly differentiated SCC	Negative	Negative	N/A	Yes	Yes

N/A: not applicable

Table 4.2. Primers used for amplification of *p53* gene from exon 4-9 and primers used for sequencing PCR amplicons. Nucleotide position relative to *p53* gene, NC_000017.10 (hg38).

Primer use	Primer name	Reference	Sequence	Nucleotide position	Annealing temperature	Amplicon size
1. First round PCR	E4/C4 F	Bosch et al., 2004	5'-GCTAGAGACCTGGTCCTCT-3'	11286	57 °C	2822 bp
	E9//C1 R	Bosch et al., 2004	5'-CGGCATTTTGAGTGTTAGAC-3'	14107		
2. Nested PCR and sequencing primers	E4/C3 F	Bosch et al., 2004	5'-GACTGCTCTTTTCACCCATC-3'	11303	57 °C	1474 bp
	2997 R	Bosch et al., 2004	5'-CCACTGACAACCACCCTTAAC-3'	12776		
3. Hemi-Nested PCR and sequencing primers	E7/C4 F	Bosch et al., 2004	5'-CTGGCCTCATCTTGGGCCTG-3'	13277	58 °C	831 bp
	E9//C1 R	Bosch et al., 2004	5'-CGGCATTTTGAGTGTTAGAC-3'	14107		
Sequencing primers	P236 F	*IARC, 2010	5'-TGTTCACTTGTGCCCTGACT-3'	12314	55 °C	249 bp
	P271 R	IARC, 2010	5'-CAGCCCTGTCGTCTCTCCAG-3'	12562		
	P326 F	IARC, 2010	5'-TGAGGACCTGGTCCTCTGAC-3'	11286	51 °C	394 bp
	P327 R	IARC, 2010	5'-AGAGGAATCCCAAAGTTCCA-3'	11679		

*International Agency for Research on Cancer

Figure 4.1. Location of primers across *p53* gene, NC_000017.10 (hg38) on chr17:7687599...7668459. Figure drawn with Geneious version 2019.0 (Biomatters). No sequence data was obtained for Intron 7.

Results

Clinical characteristics of head neck cancer patients

The entire cohort of 25 patients consisted of 23 men and 2 women (Table 4.1). Fifty-six (%) were negative for p16. Other clinical and pathologic features correlated with HPV and p16 status are shown.

Mutations in the *p53* gene in the head and neck region

In order to determine the frequency of mutations of the *p53* gene in HNSCC samples, exons 4–9 were sequenced. No sequence data was obtained for Intron 7. Reads were mapped to NC_000017.10 (hg38), to find variations/SNPs. Electropherograms showing mutant HNSCC samples are available in Appendix F.

Figure 4.2 A shows the distribution of samples according to anatomical location. Samples were selected for investigation for *p53* mutations based on the primary tumour site being the oropharynx (76% of samples). Four non-oropharyngeal HPV positive samples were included for comparison, although there were no *p53* mutations detected in this site (Figure 4.2 B). Mutations in the *p53* gene exons 4-9 were detected in 9/25 (36%) of the tumours that were tested (Figure 4.2 C). *p53* mutations were more common in non-oropharyngeal samples (44%) than oropharyngeal samples (12%). There were no HPV positive samples within the oropharynx, thus no correlation could be made on the coexistence of HPV and *p53* mutations within the oropharynx (Figure 4.2 D). More than half of all substitutions were G:C to A:T changes, followed by G:C to T:A (23%) and there was one deletion event (Figure 4.2 E). The majority of samples were HPV negative (84%) and of those, 71% did not contain mutations. Although there were only four HPV positive samples included in this study, three of the four (75%) contained *p53* mutations (Figure 4.2 F).

Thirteen mutations were identified in intronic regions and the highest mutational frequencies were found in intron 5. The SNP location and type of mutation is summarised for all samples that displayed a mutation in Table 4.3. Existing data shows that most mutations occur within the DBD. The effect of intron 5 mutations on p53 capacity to bind DNA is not known. Samples 09/17 and 30/14 both displayed a transition at the same location, which has not been described before. Interestingly, 30/14 was HPV negative whilst 09/17 was HPV positive. None of the mutations identified in this study were known polymorphisms and therefore no frequency and functional data could be obtained from the IARC database.

Table 4.3. Subset of HNSCC samples that contained *p53* mutations. Position relative to *p53* gene extracted from chromosome 17, GenBank (NC_000017.10) (hg38).

VBD No	*p53* mutations across exons 4-9						
	Position	**SNP location**	**Mutation type**	**Mutation**	**Transition/ Transversion**	**Base change**	**HPV status**
13/14	12310	Intron 5	Substitution	Wildtype	Transition	G-A	Positive
30/14	12084	Intron 5	Substitution	Wildtype	Transition	G-A	Negative
12/15	12293	Intron 5	Substitution	Wildtype	Transition	G-A	Negative
	13520	Intron 7	Substitution	Wildtype	Transition	C-T	Negative
	13540	Intron 7	Substitution	Wildtype	Transition	C-T	Negative
17/15	13511	Intron 8	Substitution	Wildtype	Transition	C-T	Positive
61/15	12084	Intron 5	Substitution	Wildtype	Transition	G-A	Negative
	12207	Intron 5	Substitution	Wildtype	Transition	A-G	Negative
	12210	Intron 5	Deletion	Wildtype	N/A	GAA	Negative
64/15	12299	Intron 5	Substitution	Wildtype	Transition	G-A	Negative
09/17	12084	Intron 5	Substitution	Wildtype	Transition	T-C	Positive
44/17	12317	Intron 5	Substitution	Wildtype	Transition	G-A	Negative
08/18	12284	Intron 5	Substitution	Wildtype	Transition	G-A	Negative

N/A: Not Applicable

Comparison of *p53* gene status with patients' clinical characteristics

Correlations of age, primary anatomical site and smoking and alcohol exposure were made with regard to *p53* status (Table 4.4). There was a predominance of males in this cohort, and only two females. Therefore no correlations could be made regarding sex and *p53* gene status. Although there was a predominance of younger patients (≤60 years), mutations were observed almost equally in older patients. Thirty percent of all smoking patients had a *p53* mutation, whilst there was a higher correlation with alcohol exposure and *p53* mutation (96%). Almost all patients had exposure to smoking and alcohol, however the data did not distinguish between ex-smokers and current smokers, nor were there accurate estimates of number of packs of cigarettes smoked per day by the number of years the person has smoked (pack years) or similar data for alcohol units consumed.

Table 4.4. Characteristics of patients and association of *p53* mutation.

Characteristics	**Total no.**	***p53* status**
		Mutant no. (%)
Sex		
Male	23	39%
Female	2	0%
Age at diagnosis		
≤60 years	16	38%
≥60 years	9	33%
Smoking	23	30%
Alcohol	24	20%
Primary site		
Nasopharynx	1	100%
HPV-negative oropharynx	21	29%
HPV-positive oropharynx	0	0%
Larynx	2	100%
Cell differentiation		
Well	2	50%
Moderate	16	25%
Poor–undifferentiated	3	66%
Carcinoma	2	100%
Well-differentiated keratinising SCC	1	0%
Moderately-differentiated keratinising SCC	1	0%
HPV status		
HPV negative	21	29%
HPV positive	4	75%

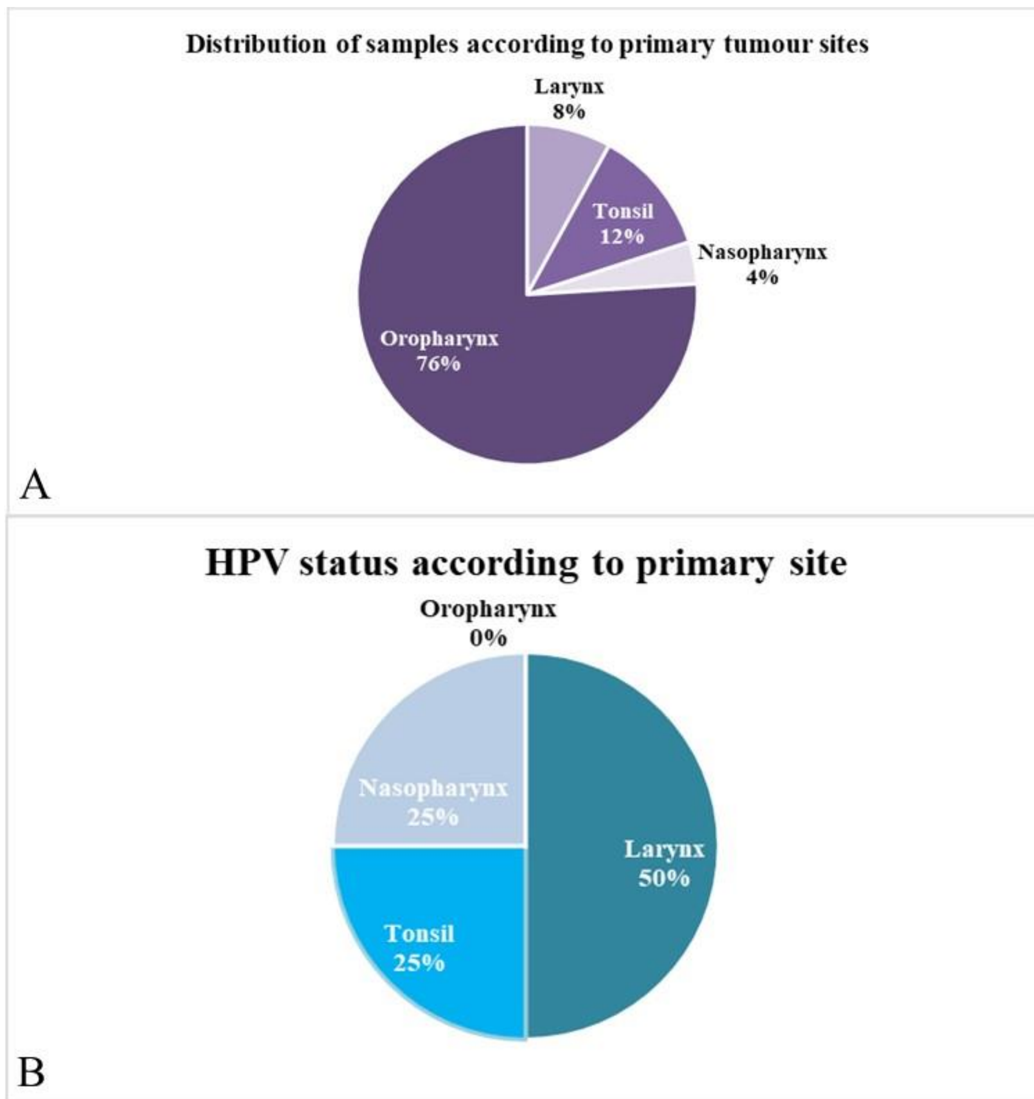


Figure 4.2. (A) Distribution of samples according to primary site and (B) HPV status according to primary site in this cohort.

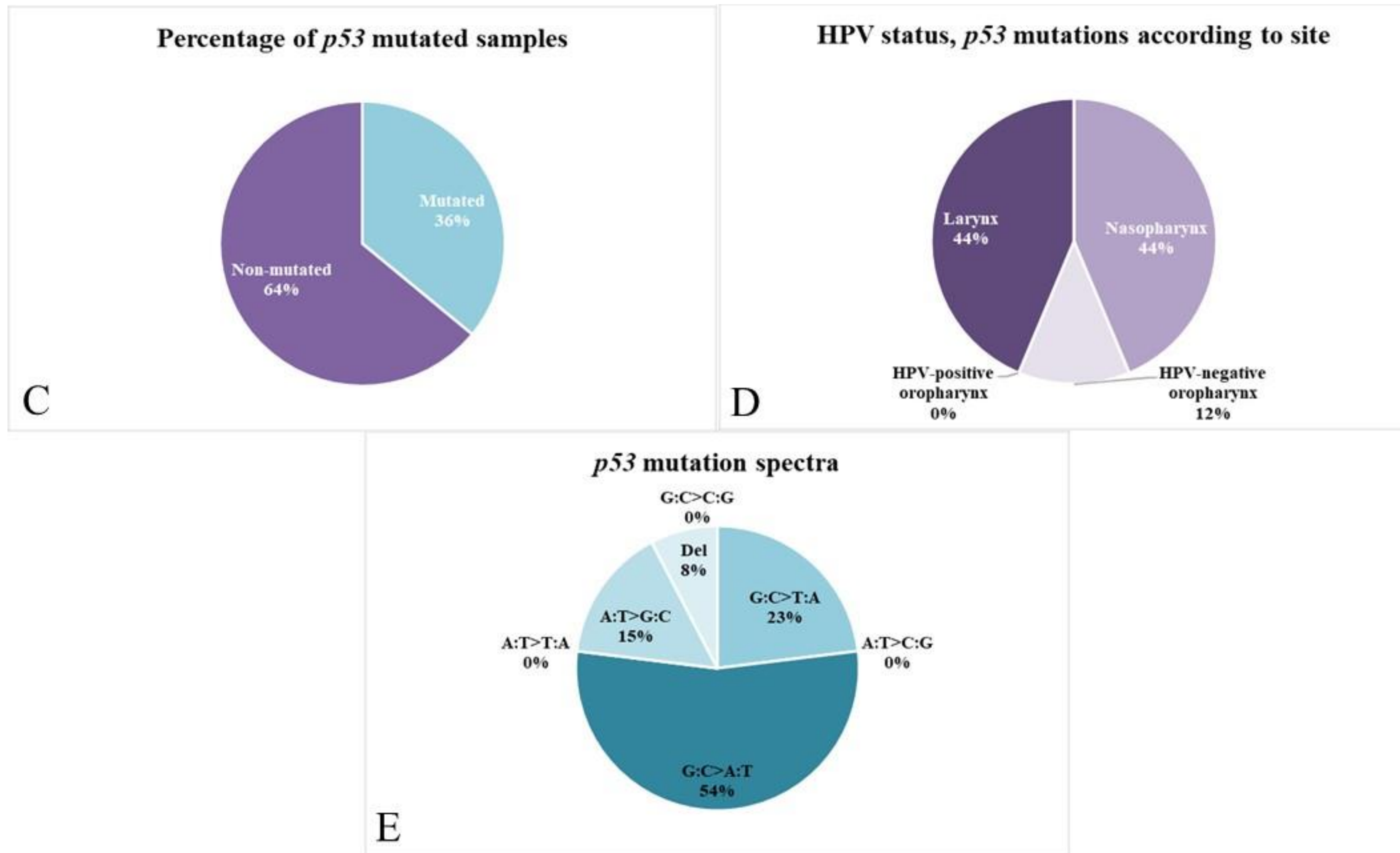


Figure 4.2. (C) Percentage of samples with *p53* mutations (D) *p53* mutation distribution according to anatomical site, considering HPV status (E) Spectra of *p53* mutations observed.

F

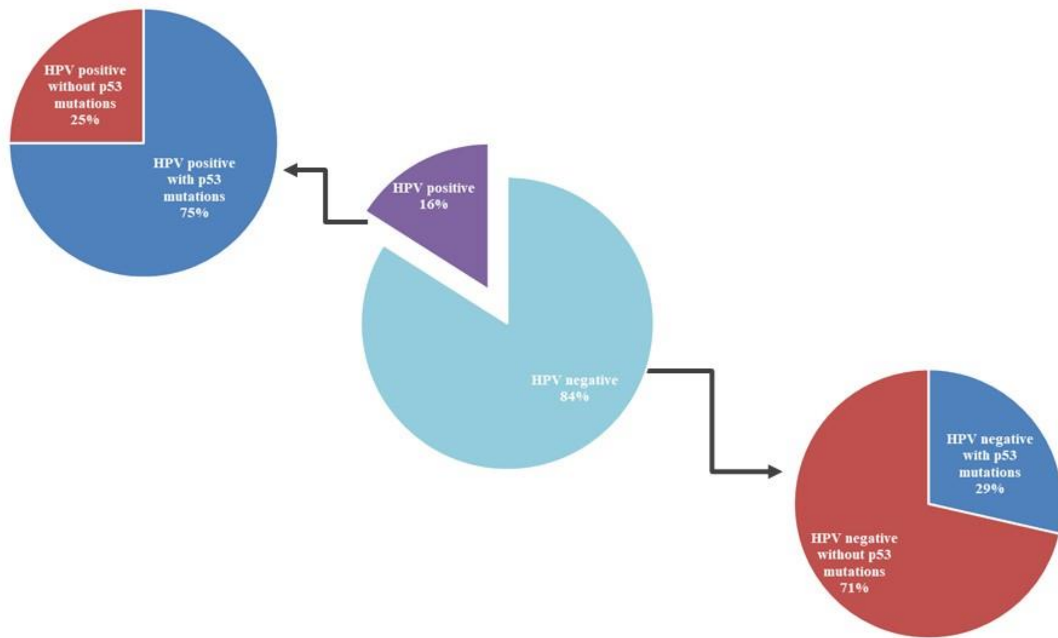


Figure 4.2. (F) *p53* mutation distribution according to HPV status.

Correlation of p16 staining and HPV status

Using p16 staining to observe cell morphology, the majority of samples were moderately-differentiated squamous cell carcinoma (SCC). p16 was considered to be positive when strong staining was observed in 60% of the tumour cells. Of the 25 HNSCC tumours, 11 (44%) showed a positive staining for p16. In summary, 7/11 (64%) of tissues the staining was extensive (>70%), in 3/11 (27%) tissues stained weakly positive (60%) for p16. p16 staining was not recorded for one sample (47/14). p16 staining was very extensive or extensive in the four tissues that were positive for HPV DNA, while none of the cases presenting with a negative or weak p16 staining were HPV positive. The correlation between p16 staining and HPV status is shown in Figure 4.3. The majority of samples (58%) were negative by both HPV detection PCR as well as p16 staining. There was a weak correlation between positivity by HPV detection PCR and p16 staining (17%). In 25% of cases where samples were HPV negative, with positive p16 staining, the p16 might have been mutated by other means. Although oropharyngeal samples were HPV negative, some of the samples stained positive for p16 expression (Table 4.1). Keratinising squamous cell carcinomas were expected of HPV-negative HNSCC with mutated *p53*. However, samples 56/15 and 67/16 did not contain *p53* mutations within exons 4-9. Although mutations could exist outside the regions that were examined. In addition, other HPV-negative samples did not show keratinising cell morphology.

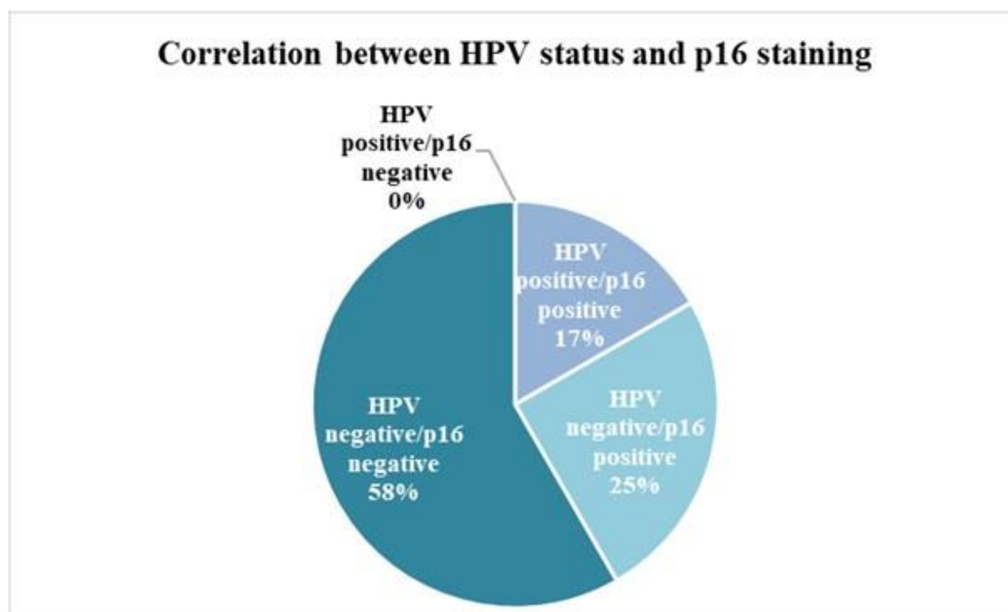


Figure 4.3. Correlation between HPV status and p16 staining.

Discussion

Although *p53* mutations are ubiquitous in human cancers and have been extensively researched, there is a paucity of information regarding *p53* mutations in head and neck cancers. As HPV and environmental carcinogens act to disrupt the *p53* pathway through two distinct mechanisms, there is an inverse relationship between HPV positivity and the existence of *p53* mutations.[5] In the case of high-risk HPV infection, tumour suppressors are inactivated by the E6 gene, thus there is little or no selective advantage for these genes to be mutated during the course of oncogenesis. HPV negatively affects the expression of *p53*, pRB and consequently there is an overexpression of p16.[143] *p53* can be mutant in some subsets of HPV-related HNSCC; these carcinomas are usually more aggressive than those with wild type *p53*.[144]

In this cohort, there was a predominance of men, which was expected as HNSCC is more than 2-fold higher in men than women, regardless of whether the cancer is HPV-driven or not.[66]

p16 is proposed to be a marker for risk stratification in oropharyngeal squamous cell carcinomas. A strong and diffuse pattern of p16 immunostaining is considered a highly sensitive surrogate marker for the identification of HPV-driven tumours.[145] The majority of samples in this study (58%) were negative by both HPV testing as well as p16 staining. However there were samples that were HPV negative, with positive p16 staining. There were a few samples (17%) that tested positive for HPV DNA and stained positive for p16 staining. In certain instances, HPV positive tumours with p16 positive staining presented a mutation in *p53*.

Approximately 15% to 20% of p16 positive oropharyngeal cases are HPV negative by *in situ hybridisation* (ISH).[64] In this subset of samples 25% of samples were p16 positive, although HPV was tested by PCR instead of ISH. HPV negative samples that stained positive for p16 may have other molecular mechanisms causing p16 overexpression. p16 positive, HPV negative oropharyngeal cases do not differ significantly in outcome, however, HPV positive tumours, p16 negative tumours fare far worse.[145] There wasn't a strong correlation with HPV positivity and p16 in this study. As p16 can be overexpressed in the absence of HPV infection, the use of p16 as a biomarker in a clinical setting should be used with caution.

Further distinction was made regarding other molecular and clinicopathological factors. HPV-positive HNSCC is normally associated with males with a lower tobacco and alcohol exposure, small tumours at first diagnosis and younger than patients with HPV-negative HNSCC.[6] Interestingly, of the 23 patients who had exposure to smoking, only 30% harboured *p53* mutations. There was a predominance of G:C > T:A mutations identified in this study. This mutational pattern is characteristic of the carcinogen benzo(A)pyrene diol epoxide, from cigarette smoke.[146,147]

Most of the patients in our study were smokers, but the association between smoking and *p53* mutation was not significant, perhaps because the data-collection forms did not distinguish between former smokers and current smokers as well as pack years. Whilst almost all patients had exposure to alcohol, only 20% of those samples had *p53* mutations. A limitation of this study is that the data did not distinguish number of alcohol units, use or abuse of alcohol, finer correlations could not be made.

The consensus on whether risk factors such as use of tobacco or alcohol and exposure to chemicals can affect the frequency of *p53* mutations in patients with HNSCC is still divisive. Previous studies demonstrated that the frequency of *p53* mutations is positively associated with tobacco exposure, whilst others do not show an association.[141]

It is well known that there is an inverse relationship HPV DNA positivity in the oropharynx and the presence of a *p53* mutation.[5] Oropharyngeal samples were investigated for the presence of *p53* mutations as this region is more likely to host HPV-associated carcinoma. Furthermore, the oropharynx is less likely to be exposed to environmental carcinogens. However, there was a low prevalence of HPV in this cohort and specifically in the oropharynx, where no HPV DNA was detected. Therefore the proportion of tumours with mutant *p53* (36%) was lower than what is normally detected in exons 4-9. A mutation frequency in HNSCC cases not driven by HPV ranges from 75 to 85%.[55,56] However, there was a coexistence of *p53* mutations and HPV in three of the four tumours in this study. For HNSCCs with nondisruptive *p53* mutations, transcriptionally active HPV could heighten the overall effect of abrogation in p53 function.[135]

It is not uncommon that the mutations identified in this study were within intronic regions that span the DNA binding domain. The majority of *p53* mutations (90%) are located in introns, outside splice sites, or in noncoding exons.[57] A high frequency of mutations were observed within intron 5 and the consequences of these mutations on the p53 capacity to bind DNA is unknown. Variants could be further analysed using online tools to analyse the potential formation or removal of splice-sites or how variants affect splice enhancers or repressors. In this study, the focus was identification of mutations to determine if further analysis is warranted.

Thus far, *p53* mutation studies have focused on sequencing hotspot regions, specifically around the DNA-binding domain. It has previously been observed that there are about 2-5 % mutations reported outside exons 5-8, therefore this study might have missed mutations that were outside the sequenced region.[141] Although, it is also a possibility that other genes in the *p53* pathway are targeted or that these tumours follow p53-independent routes of malignant progression.[157]

Applying next-generation sequencing (NGS) to the entire gene shows that there is an underestimation of mutations in the oligomerization domain located in the C-terminus of the protein (codons 323–356). Interestingly, mutations outside the DNA-binding domain do not occur at the same frequency and depend on the type of cancer.[58]

This enhanced sensitivity of NGS however comes at a relatively high cost, especially in the sub-Saharan African setting. Another limitation to consider in this study, is that Sanger sequence data was compared to the human reference genome from a single individual. The study participants were mainly of African descent, thus their genetic background does not match that of the reference. A recent paper describes deep sequence data of 900 individuals of African descent, where the African pan-genome contains close to 10% more DNA than the current human reference genome.[158] The effects of a polymorphism can be subtle and varies according to genetic background. The lack of genetic diversity in the reference genome might have contributed to there not being any functional nor frequency data available for the mutations identified in this study. Adding to the lack of data on *p53* mutations in this population, is the fact that intronic mutations have not been tested widely in functional assays or studied for effects on cancer risk.[57]

Research currently recognises the critical role played by p53 protein in suppressing tumour development. However, there is a considerable amount of work ahead in terms of mode of inactivation in head and neck cancers and resultant diversity of phenotypes.

Conclusion

HPV related HNSCC represents a distinct subgroup compared to virus-unrelated HNSCC. There was a low prevalence of HPV in this cohort. All of the mutations within this study were detected within intron 5. Intronic mutations have not been well characterised regarding p53 function abrogration. As p16 can be overexpressed in the absence of HPV infection, the use of p16 as a biomarker in a clinical setting should be used with caution. Population studies will require larger sample sizes of genetically diverse individuals. Next-generation sequencing will allow for more comprehensive studies of polymorphisms.

Acknowledgments

Y Munsamy held a scholarship by the National Research Foundation (Grant UID: 94968) and funding by the Poliomyelitis Research Foundation (Grant number: 15/68). The research was supported by the National Research Foundation (Grant number: 90556).

Competing Interests Statement

The authors declare no competing interests.

Author Contributions

FB conceived and supervised the experiments; YM designed and performed the experiments; TS contributed to genotyping the isolates; YM analysed data; RS contributed biopsy tissue; YM wrote the manuscript. FB revised and edited the paper.

.

CHAPTER 5

Conclusions and Future Directions

Conclusions and Future Directions

Papillomaviruses have a long history with human hosts and it is likely that almost all humans will be infected at some point in their lives; mostly having asymptomatic infections that are cleared by the immune system. However, an infection with one of 13 high-risk HPV types in the upper aerodigestive tract is the leading cause of head and neck squamous cell carcinomas (HNSCC). Despite phylogenetic relatedness, HPV types are known to differ in their tropisms and oncogenic potentials. However much less is known about intratypic variants of HPV with regards to oncogenicity and the clinical outcomes of infection. These varying genetic traits of the virus are all encoded within the relatively small 8 kb circular double-stranded DNA genome. So far, studies on intratypic variation have focused mainly on HPV 16 and HPV 18, showing that some variants are more aggressive than others. Further compounding our understanding of this virus, is the challenge in transposing information from cervical cancer studies to HNSCC.

Countries, in which the age rate of HPV-attributable head and neck cancer is relatively high (over 1.25 per 100 000), are located in Northern America and Europe.[108] Whereas in sub-Saharan Africa, knowledge on the role of HPV in HNC is very limited. Actually, 84% of the existing information on HPV and HNC is derived from studies in Europe and North America.[32] To date, no large-scale South African studies report on HPV type distribution and prevalence associated with head and neck cancer. There are currently no published whole genome HPV sequences isolated from HNC.

This thesis fulfilled the aims of the study to investigate genetic characterisation of HPV associated with HNSCC by performing deep sequencing analyses of viral whole-genome sequences in clinical specimens and evaluated the contribution of *p53* mutations to HNSCC.

A recent paper from our research group reported on testing of biopsy samples from histologically confirmed HNSCC patients for detection of HR-HPV in a South African cohort.[73] HPV positive samples identified using partial sequence analysis were included in this study for the first complete HPV genomes from HNSCC to be characterised.

In Chapter 2, the complete genome sequence of HPV 18 from a nasopharyngeal carcinoma was characterised. HPV 18 is widely accepted as the second most carcinogenic HPV type after HPV 16. HPV 18 isolate designated VBD 17/15 characterised in this study, was identified as belonging to lineage A1. Based on data from cervical cancers associated with HPV, HPV 18 isolates belonging to lineages A1 and A2 may have greater risk of inducing tumour formation. So far, there is no conclusive role of HPV 18 sublineages in discriminating cancer risk.[93] Further studies with a larger cohort of samples should provide evidence of the risk of HPV 18 isolates, especially within sublineage A, to cause cancer in the head and neck region.

Previous studies of intratypic evolution studies of HPV variants have not dealt with complete genome sequences, but have used the partial regions of the viral genome that were generally limited to the E6, E7, and LCR regions. The limitation of sequencing of small regions of the genome is that closely related viral variants are almost indistinguishable from one another. In this study SNPs identified using complete genome data, allow for accurate differentiation into sublineages. Complete genome sequence characterisation allowed detection of mutations and deletions and although a theoretical role of these changes can be discussed, the functional, or biological, significance of these will require further investigation. For instance, a previously reported 7 bp LCR deletion was detected in VBD 17/15. This deletion might be an effect of linkage disequilibrium (LD), a mutation that occurs simultaneously with mutations at other loci due to divergence of papillomaviruses with human hosts.

The sequence variation identified within the E2 and LCR may possibly lead to altered biological functions and affect the clinical outcome of infection. It is important to determine the role of the other viral genes, such as E2 and E5, that modulate the HPV oncogenes and in combination have an effect on the tumourigenesis process. Further research should be undertaken to investigate the novel mutation within the L2; whether it has implications on papillomavirus assembly or not. With an increase in the number of complete genomes being sequenced, novel mutations like the ones described in this study might possibly be a lineage-defining variation.

In the absence of tests for identification of precancerous lesions, this study was not able to identify molecular signatures in VBD 17/15 associated with HNSCC. Molecular signatures have been described for isolates from cervical carcinoma, where precursor lesions have been identified. Whether there are different signatures associated with HNSCC is unknown at this stage. Further compounding the linking of molecular signatures with HNSCC, is the lack of a Pap smear equivalent and the inability to sample tissues from the precancerous stage to carcinoma in the upper aerodigestive tract.

The results in Chapter 2 indicate that there are sequence variations in HPV isolates from HNSCC. However, published reports regarding functional analysis of sequence variation of non HPV 16/18 oncogenic types are rare. The functional effect of mutations within the LCR of an HPV 31 isolate was described in Chapter 3.

The deep coverage obtained with NGS data enabled the identification of an HPV 31 lineage B quasispecies co-infection. Detection of multiple sequences in one sample would not have been possible with Sanger sequencing due to low sensitivity. The co-infection was confirmed by the identification and visualisation of a polymorphism in the E5 gene occurring in shared sequence reads, representing two separate HPV 31 variant lineage molecules. The deep read depths allowed estimation of the percent infection with the two HPV 31 variants; where 70% of the reads contained an E5 deletion. It is unknown what this deletion in the E5 gene represents for viral fitness as this polymorphism has not been described before. Comparisons with other HPV 31 complete genomes from the head and neck region will

determine whether these polymorphisms are a signature of HNC isolates. It would be interesting to create E5 deletion mutants to determine the effect on oncogenicity as well as whether immune responses are diminished.

In the absence of multiple HPV 31 LCR isolates for comparison of viral promoter activity, site-directed mutagenesis was adopted to construct variants with changes that are associated with the prototype sequence (lineage A). Expression of a reporter gene was used as an indication of promotor function. The highest degree of β-galactosidase expression was observed with pBlue_SDM3, a clone in which a mutation was introduced to represent a mutation, C7564T, within the YY1 binding site.

The findings of this study were limited as the HPV-attributable fraction in this population was low; comparative studies based on a large cohort of naturally occurring HPV 31 variants would enhance our knowledge on the significance of novel mutations. Although evidence exists on HPV 31 LCR variants influencing transcriptional activity; oncogenic potential is not restricted to LCR genetic variation and there might be other factors at play.

With regard to evidence provided in both Chapters 2 and 3, the following insights can be proposed:

Integration is normally a hallmark of cervical cancer and sometimes even seen in pre-cancer stages, evidenced by large deletions in NGS datasets. Unlike other studies where whole genomes were amplified by large sets of overlapping fragments, whole genomes from this study (VBD 17/15 and VBD13/14) were derived in two overlapping fragments. Thus, if these isolates were integrated into the host genome it would not have been amplified in full. Whether integration plays a role in HNSCC warrants further investigation. The presence of HPV type co-infections by PCR and Sanger sequencing were not detected in this cohort. However reports by other HPV genomics researchers and evidence in this study from NGS data show that co-infection at the viral variant level can occur.[23,90,159] How these co-infections influence the course of carcinogenesis is not clear. Future research should focus on the interaction of these viral variants and their influence on the course of carcinogenesis.

Having analysed infections at the level of viral lineages, the final part of this thesis went on to investigate host factors relating to HNSCC. In the absence of a precursor lesion, it has been suggested that the tumour suppressor *p53* gene be used as a prognostic and predictive marker. The current paradigm recognises that non-viral HNSCCs have mutations in *p53*, whilst HPV-associated tumours usually do not possess mutations. However, HPV still interferes with normal p53 functioning, by binding and forming a complex that leads to the degradation of p53. Although, in non-viral HNSCC, HR-HPV enhances the effect of smoking on development of tumours.

It was hypothesised that HPV-positive HNSCCs especially in the oropharynx, would not show *p53* mutations. In this cohort, there was a low prevalence of HPV and therefore the oropharynx was devoid of HPV infection. Although, as expected, *p53* mutations were more common in non-oropharyngeal

samples (44%) than oropharyngeal samples (12%). There were four non-oropharyngeal samples that were included based on known HPV positivity and of those, the majority (75%) contained *p53* mutations. Most mutations within the *p53* gene are intron-based, within splice sites or exist in noncoding regions. Therefore it was not uncommon that all of the mutations within this study were detected within intron 5. Intronic mutations have not been well characterised regarding the biological consequences of these mutations. Both tissue and type of cancer influence the frequency of *p53* mutations. Although one of the consequences of *p53* mutations is a loss of function of DNA binding and transcription, there are other consequences of mutant missense p53 proteins in a cancer cell.[59]

There was a predominance of G:C to A:T changes, a substitution pattern that is associated with a tobacco smoke carcinogen. Pack years as well as alcohol units were not factored into the questionnaire from a related study. Although HPV-HNSCC is normally associated with younger males, there weren't enough females in this cohort to make statistically significant correlations. Although there was a predominance of younger patients (≤60 years), *p53* mutations were observed almost equally in older patients. A higher *p53* mutation frequency was expected, although there may be other cancer driver genes involved in tumourigenesis of this cohort of HNSCC samples. In addition, other sites not sequenced in this study might harbour *p53* mutations.

With respect to establishing the relationship between HPV status and p16 immunostaining, it was found that although a quarter of samples tested were p16 positive, the correlation with HPV detection by PCR was low. Using p16 as a marker for risk stratification should be used with caution in clinical settings as other molecular mechanisms might cause p16 overexpression.

Overall, NGS technologies should be employed in sequencing of cancer driver genes in human cancers. None of the mutations identified in this study were known polymorphisms and therefore no frequency and functional data could be obtained from the IARC database. It is possible, therefore, that people of African descent are genetically diverse from the human reference genome and there were mutations in this cohort that were missed. There is still a significant amount of work that needs to be conducted on describing mutations within participants of African descent to add to the IARC database. The aetiologic factors of HNSCC complicates determining how the *p53* pathway is affected especially in HPV-HNSCC that are associated with smoking. Understanding of how these cancers arise can lead to more targeted therapeutics and indicators of prognosis. As of yet, there is considerable work to be conducted in understanding human cancer genomics in sub-Saharan Africa.

It is evident that oncogenicity is a complex process that requires further investigation of both host factors and viral genetics. A natural progression of this work is to analyse whether the novel mutations identified in these South African isolates are unique to this region or whether they are polymorphisms identified specific for HNC.

Although cervical cancer is on the decline in Western countries, head and neck cancers due to HPV are rising. It is projected that with the high HPV burden in the cervix and the lack of knowledge on HPV an increasing trend in unsafe oral sex behaviours with a concomitant rise in cancer especially in South Africa (SA) is expected. In SA, bivalent vaccination (Cervarix®) is being provided to female primary school learners, therefore infection by types 16/18 might eventually be reduced. As of yet, the quadrivalent vaccines (Gardasil®) are not used as routinely and the nonavalent vaccine (Gardasil® 9) is not approved for use in SA. With the decline in HPV 16/18 infections, there might be a rise in HPV 31 infections.

Considerable work will need to be conducted to develop a Pap smear equivalent or to identify suitable biomarkers, as this seems to be the limiting factor in detecting tumours in the head and neck region. In sampling precancerous tissues, one can follow progression to carcinoma. Larger studies are warranted to characterise HPV with NGS data to detect molecular signatures indicative of high-risk-to-cause-cancer in the head and neck region. Additionally, earlier detection might help patient prognosis as well as determine if treatment should be modified. Although the lack of tests for precancerous lesions limit early detection, HPV testing of biopsies has prognostic implications.

The present study adds to the growing body of literature elucidating the underlying molecular mechanisms of HPV infection in HNC in sub-Saharan Africa, especially in light of high HPV burden in the cervix. When looking at the complexity of the human genome, it seems more of a tangible goal to intensify studies on the <8 kb HPV genome in an attempt to understand viral HNSCC. The first HPV 18 and 31 isolates were characterised at the whole-genome level and novel mutations were found in these isolates. In addition, evidence of genetic diversity contributing to an increase of transcriptional activity is presented. HPV-HNSCC is complicated by synergistic interaction with host factors. Evidence of *p53* mutations was shown, however there is a significant amount of research to be conducted on cancer driver genes in head and neck cancer.

REFERENCES

1. Kamangar, F., Dores, G. M. & Anderson, W. F. Patterns of Cancer Incidence, Mortality, and Prevalence Across Five Continents: Defining Priorities to Reduce Cancer Disparities in Different Geographic Regions of the World. *J. Clin. Oncol.* **24**, 2137–2150 (2006).
2. Pfster, D. G. *et al.* Head and Neck Cancers. Natl. Compr. Canc. Netw. 13, 847-856 (2016).
3. Jemal, A. *et al.* Annual Report to the Nation on the Status of Cancer, 1975–2009, Featuring the Burden and Trends in Human Papillomavirus (HPV)–Associated Cancers and HPV Vaccination Coverage Levels. *JNCI J. Natl. Cancer Inst.* **105**, 175–201 (2013).
4. Pimenoff, V. N., de Oliveira, C. M. & Bravo, I. G. Transmission between Archaic and Modern Human Ancestors during the Evolution of the Oncogenic Human Papillomavirus 16. *Mol. Biol. Evol.* **34**, 4–19 (2017).
5. Gillison, M. L. Evidence for a Causal Association Between Human Papillomavirus and a Subset of Head and Neck Cancers. *J. Natl. Cancer Inst.* **92**, 709–720 (2000).
6. Marur, S., D'Souza, G., Westra, W. H. & Forastiere, A. A. HPV-associated head and neck cancer: a virus-related cancer epidemic. *Lancet Oncol.* **11**, 781–789 (2010).
7. Combes, J.-D. & Franceschi, S. Human papillomavirus genome variants and head and neck cancers: a perspective. *Infect. Agent. Cancer* **13**, (2018).
8. Powell, N. G. & Evans, M. Human papillomavirus-associated head and neck cancer: oncogenic mechanisms, epidemiology and clinical behaviour. *Diagn. Histopathol.* **21**, 49–64 (2015).
9. Morgan, I., DiNardo, L. & Windle, B. Integration of Human Papillomavirus Genomes in Head and Neck Cancer: Is It Time to Consider a Paradigm Shift? *Viruses* **9**, 208 (2017).
10. Ndiaye, C. *et al.* HPV DNA, E6/E7 mRNA, and p16INK4a detection in head and neck cancers: a systematic review and meta-analysis. *Lancet Oncol.* **15**, 1319–1331 (2014).
11. zur Hausen, H. Papillomavirus infections - a major cause of human cancers. *Biochimica et Biophysica Acta* **1288**, F55–F78 (1996).
12. Stanley, M. A. Epithelial Cell Responses to Infection with Human Papillomavirus. *Clin. Microbiol. Rev.* **25**, 215–222 (2012).
13. Wilson, V. G., West, M., Woytek, K. & Rangasamy, D. Papillomavirus E1 Proteins: Form, Function, and Features. *Virus Genes* **24**, 275–290 (2002).
14. Tsakogiannis, D. *et al.* Molecular and phylogenetic analysis of the HPV 16 E4 gene in cervical lesions from women in Greece. *Arch. Virol.* **157**, 1729–1739 (2012).
15. DiMaio, D. & Petti, L. M. The E5 proteins. *Virology* **445**, 99–114 (2013).
16. Fehrmann, F. & Laimins, L. A. Human papillomaviruses: targeting differentiating epithelial cells for malignant transformation. *Oncogene* **22**, 5201–5207 (2003).

17. de Villiers, E.-M., Fauquet, C., Broker, T. R., Bernard, H.-U. & zur Hausen, H. Classification of papillomaviruses. *Virology* **324**, 17–27 (2004).
18. Doorbar, J. The papillomavirus life cycle. *J. Clin. Virol.* **32**, 7–15 (2005).
19. Syrjanen, S. The role of human papillomavirus infection in head and neck cancers. *Ann. Oncol.* **21**, vii243–vii245 (2010).
20. Bouvard, V. Human Papillomaviruses. (2009).
21. Bzhalava, D., Eklund, C. & Dillner, J. International standardization and classification of human papillomavirus types. *Virology* **476**, 341–344 (2015).
22. Bernard, H.-U. *et al.* Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology* **401**, 70–79 (2010).
23. Hirose, Y. *et al.* Within-Host Variations of Human Papillomavirus Reveal APOBEC Signature Mutagenesis in the Viral Genome. *J. Virol.* **92**, e00017-18 (2018).
24. Doorbar, J. *et al.* The Biology and Life-Cycle of Human Papillomaviruses. *Vaccine* **30**, F55–F70 (2012).
25. Hebner, C., Beglin, M. & Laimins, L. A. Human Papillomavirus E6 Proteins Mediate Resistance to Interferon-Induced Growth Arrest through Inhibition of p53 Acetylation. *J. Virol.* **81**, 12740–12747 (2007).
26. Schiller, J. T., Day, P. M. & Kines, R. C. Current understanding of the mechanism of HPV infection. *Gynecol. Oncol.* **118**, S12–S17 (2010).
27. McMillan, N. A. J., Payne, E., Frazer, I. H. & Evander, M. Expression of the α6 Integrin Confers Papillomavirus Binding upon Receptor-Negative B-Cells. *Virology* **261**, 271–279 (1999).
28. Shafti-Keramat, S. *et al.* Different Heparan Sulfate Proteoglycans Serve as Cellular Receptors for Human Papillomaviruses. *J. Virol.* **77**, 13125–13135 (2003).
29. Li, M., Beard, P., Estes, P. A., Lyon, M. K. & Garcea, R. L. Intercapsomeric Disulfide Bonds in Papillomavirus Assembly and Disassembly. *J. Virol.* **72**, 8 (1998).
30. Bousarghin, L., Touze, A., Sizaret, P.-Y. & Coursaget, P. Human Papillomavirus Types 16, 31, and 58 Use Different Endocytosis Pathways To Enter Cells. *J. Virol.* **77**, 3846–3850 (2003).
31. Day, P. M. & Schelhaas, M. Concepts of papillomavirus entry into host cells. *Curr. Opin. Virol.* **4**, 24–31 (2014).
32. Kreimer, A. R., Clifford, G. M. & Boyle, P. Human Papillomavirus Types in Head and Neck Squamous Cell Carcinomas Worldwide: A Systematic Review. *Cancer Epidemiol. Biomark. Prev.* 10 (2005).
33. Biryukov, J. & Meyers, C. Superinfection Exclusion between Two High-Risk Human Papillomavirus Types during a Coinfection. *J. Virol.* **92**, e01993-17 (2018).
34. Aydin, I. *et al.* Large Scale RNAi Reveals the Requirement of Nuclear Envelope Breakdown for Nuclear Import of Human Papillomaviruses. *PLoS Pathog.* **10**, e1004162 (2014).

35. Calton, C. M. *et al.* Translocation of the papillomavirus L2/vDNA complex across the limiting membrane requires the onset of mitosis. *PLoS Pathog.* **13**, e1006200 (2017).
36. Munger, K. & Jones, D. L. Human Papillomavirus Carcinogenesis: an Identity Crisis in the Retinoblastoma Tumor Suppressor Pathway. *J. Virol.* **89**, 4708–4711 (2015).
37. Moody, C. A. & Laimins, L. A. Human papillomavirus oncoproteins: pathways to transformation. *Nat. Rev. Cancer* **10**, 550–560 (2010).
38. Munger, K. *et al.* Mechanisms of Human Papillomavirus-Induced Oncogenesis. *J. Virol.* **78**, 11451–11460 (2004).
39. Ganguly, N. & Parihar, S. P. Human papillomavirus E6 and E7 oncoproteins as risk factors for tumorigenesis. *J. Biosci.* **34**, 113–123 (2009).
40. Mellin, H. *et al.* Human papillomavirus type 16 is episomal and a high viral load may be correlated to better prognosis in tonsillar cancer. *Int. J. Cancer* **102**, 152–158 (2002).
41. Akagi, K. *et al.* Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. *Genome Res.* **24**, 185–199 (2014).
42. Bodelon, C., Untereiner, M. E., Machiela, M. J., Vinokurova, S. & Wentzensen, N. Genomic characterization of viral integration sites in HPV-related cancers: Genomic Characterization of HPV Integration Sites. *Int. J. Cancer* **139**, 2001–2011 (2016).
43. Wentzensen, N., Vinokurova, S. & Doeberitz, M. von K. Systematic Review of Genomic Integration Sites of Human Papillomavirus Genomes in Epithelial Dysplasia and Invasive Cancer of the Female Lower Genital Tract. *Cancer Res.* **64**, 3878–3884 (2004).
44. Hu, Z. *et al.* Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism. *Nat. Genet.* **47**, 158–163 (2015).
45. Vinokurova, S. *et al.* Type-Dependent Integration Frequency of Human Papillomavirus Genomes in Cervical Lesions. *Cancer Res.* **68**, 307–313 (2008).
46. The Cancer Genome Atlas Research Network. Integrated genomic and molecular characterization of cervical cancer. *Nature* **543**, 378–384 (2017).
47. Couturier, J., Sastre-Garau, X., Schneider-Maunoury, S., Labib, A. & Orth, G. Integration of Papillomavirus DNA near myc Genes in Genital Carcinomas and Its Consequences for Proto-Oncogene Expression. *J. Virol.* **65**, 5 (1991).
48. Jackson, R. *et al.* Functional variants of human papillomavirus type 16 demonstrate host genome integration and transcriptional alterations corresponding to their unique cancer epidemiology. *BMC Genomics* **17**, (2016).
49. Hein, A. L., Ouellette, M. M. & Yan, Y. Radiation-induced signaling pathways that promote cancer cell survival (Review). *Int. J. Oncol.* **45**, 1813–1819 (2014).
50. Leemans, C. R., Snijders, P. J. F. & Brakenhoff, R. H. The molecular landscape of head and neck cancer. *Nat. Rev. Cancer* **18**, 269–282 (2018).

51. Kobayashi, K. *et al.* A Review of HPV-Related Head and Neck Cancer. *J. Clin. Med.* **7**, 241 (2018).
52. Hayes, D. N., Van Waes, C. & Seiwert, T. Y. Genetic Landscape of Human Papillomavirus–Associated Head and Neck Cancer and Comparison to Tobacco-Related Tumors. *J. Clin. Oncol.* **33**, 3227–3234 (2015).
53. Romagosa, C. *et al.* p16Ink4a overexpression in cancer: a tumor suppressor gene associated with senescence and high-grade tumors. *Oncogene* **30**, 2087–2097 (2011).
54. Reed, A. L. *et al.* High Frequency of p16 (CDKN2/MTS-1/INK4A) Inactivation in Head and Neck Squamous Cell Carcinoma. *Cancer res*. **56**, 3630-3633 (1996).
55. Agrawal, N. *et al.* Exome Sequencing of Head and Neck Squamous Cell Carcinoma Reveals Inactivating Mutations in NOTCH1. *Science* **333**, 1154–1157 (2011).
56. Stransky, N. *et al.* The Mutational Landscape of Head and Neck Squamous Cell Carcinoma. *Science* **333**, 1157–1160 (2011).
57. Olivier, M., Hollstein, M. & Hainaut, P. TP53 Mutations in Human Cancers: Origins, Consequences, and Clinical Use. *Cold Spring Harb. Perspect. Biol.* **2**, a001008–a001008 (2010).
58. Bouaoun, L. *et al. TP53* Variations in Human Cancers: New Lessons from the IARC TP53 Database and Genomics Data: Human Mutation. *Hum. Mutat.* **37**, 865–876 (2016).
59. Levine, A. J. Targeting Therapies for the p53 Protein in Cancer Treatments. *Annu. Rev. Cancer Biol.* **3**, (2019).
60. Louie, K. S., Mehanna, H. & Sasieni, P. Trends in head and neck cancers in England from 1995 to 2011 and projections up to 2025. *Oral Oncol.* **51**, 341–348 (2015).
61. Chaturvedi, A. K. Epidemiology and Clinical Aspects of HPV in Head and Neck Cancers. *Head Neck Pathol.* **6**, 16–24 (2012).
62. Gillison, M. L. *et al.* Human Papillomavirus and Diseases of the Upper Airway: Head and Neck Cancer and Respiratory Papillomatosis. *Vaccine* **30**, F34–F54 (2012).
63. Näsman, A. *et al.* Incidence of human papillomavirus (HPV) positive tonsillar carcinoma in Stockholm, Sweden: An epidemic of viral-induced carcinoma? *Int. J. Cancer* **125**, 362–366 (2009).
64. Smeets, S. J. *et al.* A novel algorithm for reliable detection of human papillomavirus in paraffin embedded head and neck cancer specimen. *Int. J. Cancer* **121**, 2465–2472 (2007).
65. D'Souza, G. & Dempsey, A. The role of HPV in head and neck cancer and review of the HPV vaccine. *Prev. Med.* **53**, S5–S11 (2011).
66. Chaturvedi, A. K., Engels, E. A., Anderson, W. F. & Gillison, M. L. Incidence Trends for Human Papillomavirus–Related and –Unrelated Oral Squamous Cell Carcinomas in the United States. *J. Clin. Oncol.* **26**, 612–619 (2008).
67. Faggons, C. E. *et al.* Human papillomavirus in head and neck squamous cell carcinoma: A descriptive study of histologically confirmed cases at Kamuzu Central Hospital in Lilongwe, Malawi. *Malawi Med. J.* **29**, 142 (2017).

68. Matsha, T. Human papillomavirus associated with oesophageal cancer. *J. Clin. Pathol.* **55**, 587–590 (2002).
69. Richter, K. L., Van Rensburg, E. J., Van Heerden, W. F. P. & Boy, S. C. Human papilloma virus types in the oral and cervical mucosa of HIV-positive South African women prior to antiretroviral therapy: HPV types in the mucosa of SA women with AIDS. *J. Oral Pathol. Med.* **37**, 555–559 (2008).
70. Davidson, C. L., Richter, K. L., Van der Linde, M., Coetsee, J. & Boy, S. C. Prevalence of oral and oropharyngeal human papillomavirus in a sample of South African men: A pilot study. *S. Afr. Med. J.* **104**, 358 (2014).
71. Chikandiwa, A., Pisa, P. T., Sengayi, M., Singh, E. & Delany-Moretlwe, S. Patterns and trends of HPV-related cancers other than cervix in South Africa from 1994–2013. *Cancer Epidemiol.* **58**, 121–129 (2019).
72. Schäfer, G. *et al.* The role of inflammation in HPV infection of the Oesophagus. *BMC Cancer* **13**, 185 (2013).
73. Sekee, T. R. *et al.* Human papillomavirus in head and neck squamous cell carcinomas in a South African cohort. *Papillomavirus Res.* **6**, 58–62 (2018).
74. Sturgis, E. M. & Cinciripini, P. M. Trends in head and neck cancer incidence in relation to smoking prevalence: An emerging epidemic of human papillomavirus-associated cancers? *Cancer* **110**, 1429–1435 (2007).
75. Camargo, M. *et al.* Detection by PCR of human papillomavirus in Colombia: Comparison of GP5+/6+ and MY09/11 primer sets. *J. Virol. Methods* **178**, 68–74 (2011).
76. Coutlee, F. *et al.* Use of PGMY Primers in L1 Consensus PCR Improves Detection of Human Papillomavirus DNA in Genital Samples. *J. Clin. Microbiol.* **40**, 902–907 (2002).
77. Gravitt, P. E. *et al.* Improved Amplification of Genital Human Papillomaviruses. *J CLIN MICROBIOL* **38**, 5 (2000).
78. Gillison, M. L. Human papillomavirus-associated head and neck cancer is a distinct epidemiologic, clinical, and molecular entity. *Semin. Oncol.* **31**, 744–754 (2004).
79. Herrero, R. Human Papillomavirus and Oral Cancer: The International Agency for Research on Cancer Multicenter Study. *CancerSpectrum Knowl. Environ.* **95**, 1772–1783 (2003).
80. Kreimer, A. R. *et al.* Efficacy of a bivalent HPV 16/18 vaccine against anal HPV 16/18 infection among young women: a nested analysis within the Costa Rica Vaccine Trial. *Lancet Oncol.* **12**, 862–870 (2011).
81. Adams, A., Wise-Draper, T. & Wells, S. Human Papillomavirus Induced Transformation in Cervical and Head and Neck Cancers. *Cancers* **6**, 1793–1820 (2014).
82. Chai, R. C., Lambie, D., Verma, M. & Punyadeera, C. Current trends in the etiology and diagnosis of HPV-related head and neck cancers. *Cancer Med.* **4**, 596–607 (2015).

83. Bratman, S. V. *et al.* Human Papillomavirus Genotype Association With Survival in Head and Neck Squamous Cell Carcinoma. *JAMA Oncol.* **2**, 823 (2016).
84. Maxwell, J. H., Grandis, J. R. & Ferris, R. L. HPV-Associated Head and Neck Cancer: Unique Features of Epidemiology and Clinical Management. *Annu. Rev. Med.* **67**, 91–101 (2016).
85. Rector, A. *et al.* Ancient papillomavirus-host co-speciation in Felidae. *Genome Biol.* **8**, R57 (2007).
86. Chen, Z., DeSalle, R., Schiffman, M., Herrero, R. & Burk, R. D. Evolutionary Dynamics of Variant Genomes of Human Papillomavirus Types 18, 45, and 97. *J. Virol.* **83**, 1443–1455 (2009).
87. Ong, C.-K. *et al.* Evolution of Human Papillomavirus Type 18: an Ancient Phylogenetic Root in Africa and Intratype Diversity Reflect Coevolution with Human Ethnic Groups. *JVI* **67**, 6424–6431 (1993).
88. Calleja-Macias, I. E. *et al.* Worldwide Genomic Diversity of the High-Risk Human Papillomavirus Types 31, 35, 52, and 58, Four Close Relatives of Human Papillomavirus Type 16. *J. Virol.* **79**, 13630–13640 (2005).
89. Burk, R. D., Harari, A. & Chen, Z. Human papillomavirus genome variants. *Virology* **445**, 232–243 (2013).
90. Mirabello, L. *et al.* HPV16 Sublineage Associations With Histology-Specific Cancer Risk Using HPV Whole-Genome Sequences in 3200 Women. *J. Natl. Cancer Inst.* **108**, djw100 (2016).
91. Chen, A. A., Gheit, T., Franceschi, S., Tommasino, M. & Clifford, G. M. Human Papillomavirus 18 Genetic Variation and Cervical Cancer Risk Worldwide. *J. Virol.* **89**, 10680–10687 (2015).
92. van der Weele, P., Meijer, C. J. L. M. & King, A. J. Whole-Genome Sequencing and Variant Analysis of Human Papillomavirus 16 Infections. *J. Virol.* **91**, (2017).
93. van der Weele, P., Meijer, C. & King, A. High Whole-Genome Sequence Diversity of Human Papillomavirus Type 18 Isolates. *Viruses* **10**, 68 (2018).
94. Chen, Z. *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). *Virol.* **516**, 86–101 (2018).
95. Xi, L. F. *et al.* Human Papillomavirus Type 16 and 18 Variants: Race-Related Distribution and Persistence. *J. Natl. Canc. Inst.* **98**, 1045–1052 (2006).
96. Burk, R. D. *et al.* Distribution of Human Papillomavirus Types 16 and 18 Variants in Squamous Cell Carcinomas and Adenocarcinomas of the Cervix. Cancer Res. **63**, 7215–7220 (2003).
97. Villa, L. L. *et al.* Molecular variants of human papillomavirus types 16 and 18 preferentially associated with cervical neoplasia. *J. Gen. Virol.* **81**, 2959–2968 (2000).
98. Xi, L. F. *et al.* Risk for High-Grade Cervical Intraepithelial Neoplasia Associated with Variants of Human Papillomavirus Types 16 and 18. *Cancer Epidemiol. Biomarkers Prev.* **16**, 4–10 (2007).
99. Chen, A. A., Gheit, T., Franceschi, S., Tommasino, M. & Clifford, G. M. Human Papillomavirus 18 Genetic Variation and Cervical Cancer Risk Worldwide. *J. Virol.* **89**, 10680–10687 (2015).

100. Veress, G., Szarka, K., Dong, X. P. & Gergely, L. Functional significance of sequence variation in the E2 gene and the long control region of human papillomavirus type 16. *J. Gen. Virol.* **80**, 1035–1043 (1999).

101. Kurvinen, K., Yliskoski, M., Saarikoski, S., Syrjänen, K. & Syrjänen, S. Variants of the long control region of human papillomavirus type 16. *Eur. J. Cancer* **36**, 1402–1410 (2000).

102. Kämmer, C., Warthorst, U., Torrez-Martinez, N., Wheeler, C. M. & Pfister, H. Sequence analysis of the long control region of human papillomavirus type 16 variants and functional consequences for P97 promoter activity. *J Gen Virol* **81**, 1975–1981 (2000).

103. Ferenczi, A. *et al.* Phylogenetic and functional analysis of sequence variation of human papillomavirus type 31 E6 and E7 oncoproteins. *Infect. Genet. Evol.* **43**, 94–100 (2016).

104. Gee, P. *et al.* APOBEC1-Mediated Editing and Attenuation of Herpes Simplex Virus 1 DNA Indicate That Neurons Have an Antiviral Role during Herpes Simplex Encephalitis. *J. Virol.* **85**, 9726–9736 (2011).

105. Suspene, R. *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.* **85**, 7594–7602 (2011).

106. Lucifora, J. *et al.* Specific and Nonhepatotoxic Degradation of Nuclear Hepatitis B Virus cccDNA. *Science* **343**, 1221–1228 (2014).

107. Torre, L. A. *et al.* Global cancer statistics, 2012: Global Cancer Statistics, 2012. *CA. Cancer J. Clin.* **65**, 87–108 (2015).

108. de Martel, C., Plummer, M., Vignat, J. & Franceschi, S. Worldwide burden of cancer attributable to HPV by site, country and HPV type: Worldwide burden of cancer attributable to HPV. *Int. J. Cancer* **141**, 664–670 (2017).

109. Guan, P. *et al.* Human papillomavirus types in 115,789 HPV-positive women: A meta-analysis from cervical infection to cancer. *Int. J. Cancer* **131**, 2349–2359 (2012).

110. de Sanjose, S. *et al.* Human papillomavirus genotype attribution in invasive cervical cancer: a retrospective cross-sectional worldwide study. *Lancet Oncol.* **11**, 1048–1056 (2010).

111. Taylor, S., Bunge, E., Bakker, M. & Castellsagué, X. The incidence, clearance and persistence of non-cervical human papillomavirus infections: a systematic review of the literature. *BMC Infect. Dis.* **16**, (2016).

112. Mirabello, L. *et al.* The Intersection of HPV Epidemiology, Genomics and Mechanistic Studies of HPV-Mediated Carcinogenesis. *Viruses* **10**, 80 (2018).

113. Liu, Y., Pan, Y., Gao, W., Ke, Y. & Lu, Z. Whole-Genome Analysis of Human Papillomavirus Types 16, 18, and 58 Isolated from Cervical Precancer and Cancer Samples in Chinese Women. *Sci. Rep.* **7**, (2017).

114. Chen, Z. *et al.* Evolution and Taxonomic Classification of Human Papillomavirus 16 (HPV16)-Related Variant Genomes: HPV31, HPV33, HPV35, HPV52, HPV58 and HPV67. *PLoS ONE* **6**, e20183 (2011).

115. Arias-Pulido, H., Peyton, C. L., Joste, N. E., Vargas, H. & Wheeler, C. M. Human Papillomavirus Type 16 Integration in Cervical Carcinoma In Situ and in Invasive Cervical Cancer. *J. Clin. Microbiol.* **44**, 1755–1762 (2006).

116. de Roda Husman, A.-M., Walboomers, J., van den Brule, A., Meijer, C. J. L. M. & Snijders, P. The use of general primers GP5 and GP6 elongated at their 3' ends with adjacent highly conserved sequences improves human papillomavirus detection by PCR. *J. Gen. Virol.* **76**, 1057–62 (1995).

117. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).

118. Nurk, S. *et al.* Assembling Single-Cell Genomes and Mini-Metagenomes From Chimeric MDA Products. *J. Comput. Biol.* **20**, 714–737 (2013).

119. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

120. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).

121. Tamura, K. & Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**, 512–526 (1993).

122. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).

123. Boshart, M., Gissmann, L., Ikenburg, H. & Kleinheinz, A. A new type of papillomavirus DNA, its presence in genital cancer biopsies and in cell lines derived from cervical cancer. *The EMBO Journal* **3**, 7 (1151).

124. Cole, S. T. & Danos, O. Nucleotide sequence and comparative analysis of the human papillomavirus type 18 genome. *J. Mol. Biol.* **193**, 599–608 (1987).

125. Lurchachaiwong, W. *et al.* Whole-Genome Sequence Analysis of Human Papillomavirus Type 18 from Infected Thai Women. *Intervirology* **53**, 161–166 (2010).

126. Chen, Z. *et al.* Evolution and Taxonomic Classification of Alphapapillomavirus 7 Complete Genomes: HPV18, HPV39, HPV45, HPV59, HPV68 and HPV70. *PLoS ONE* **8**, e72565 (2013).

127. Mirabello, L. *et al.* HPV16 E7 Genetic Conservation Is Critical to Carcinogenesis. *Cell* **170**, 1164-1174.e6 (2017).

128. Wang, J. W. & Roden, R. B. S. L2, the minor capsid protein of papillomavirus. *Virology* **445**, 175–186 (2013).

129. Arroyo, S. L. *et al.* Human Papillomavirus (HPV) genotype 18 variants in patients with clinical manifestations of HPV related infections in Bilbao, Spain. *Virol. J.* **9**, 258 (2012).

130. Tenjimbayashi, Y. *et al.* Whole-genome analysis of human papillomavirus genotypes 52 and 58 isolated from Japanese women with cervical intraepithelial neoplasia and invasive cervical cancer. *Infect. Agent. Cancer* **12**, (2017).

131. Doorbar, J., Egawa, N., Griffin, H., Kranjec, C. & Murakami, I. Human papillomavirus molecular biology and disease association: Human papillomavirus. *Rev. Med. Virol.* **25**, 2–23 (2015).

132. Schiffman, M. *et al.* A Population-Based Prospective Study of Carcinogenic Human Papillomavirus Variant Lineages, Viral Persistence, and Cervical Neoplasia. *Cancer Res.* **70**, 3159–3169 (2010).

133. Ferenczi, A. *et al.* Sequence variation of human papillomavirus Type 31 long control region: Phylogenetic and functional implications. *J. Med. Virol.* **85**, 852–859 (2013).

134. Lorincz, A. T., Lancaster, W. D. & Temple', G. F. Cloning and Characterization of the DNA of a New Human Papillomavirus from a Woman with Dysplasia of the Uterine Cervix. *J VIROL* **58**, 5 (1986).

135. Fehrmann, F., Klumpp, D. J. & Laimins, L. A. Human Papillomavirus Type 31 E5 Protein Supports Cell Cycle Progression and Activates Late Viral Functions upon Epithelial Differentiation. *J. Virol.* **77**, 2819–2831 (2003).

136. Deng, Z., Cao, P., Wan, M. M. & Sui, G. Yin Yang 1: A multifaceted protein beyond a transcription factor. *Transcription* **1**, 81–84 (2010).

137. Xi, L. F. *et al.* Persistence of newly detected human papillomavirus type 31 infection, stratified by variant lineage. *Int. J. Cancer* **132**, 549–555 (2013).

138. Zhou, G., Liu, Z. & Myers, J. N. *TP53* Mutations in Head and Neck Squamous Cell Carcinoma and Their Impact on Disease Progression and Treatment Response. *J. Cell. Biochem.* **117**, 2682–2692 (2016).

139. Taylor, D. *et al.* Immunohistochemical detection of p53 protein accumulation in head and neck cancer: Correlation with p53 gene alterations. *Hum. Pathol.* **30**, 1221–1225 (1999).

140. Lamb, P. & Crawford, L. Characterization of the human p53 gene. *Mol. Cell. Biol.* **6**, 1379–1385 (1986).

141. Bosch, F. X. *et al.* Head and neck tumor sites differ in prevalence and spectrum ofp53 alterations but these have limited prognostic value. *Int. J. Cancer* **111**, 530–538 (2004).

142. Halazonetis, T. D. & Kandil, A. N. Conformational shifts propagate from the oligomerization domain of p53 to its tetrameric DNA binding domain and restore DNA binding to select p53 mutants. *EMBO J.* **12**, 5057–5064 (1993).

143. Butler, J. S. & Loh, S. N. Folding and misfolding mechanisms of the p53 DNA binding domain at physiological temperature. *Protein Sci.* **15**, 2457–2465 (2006).

144. Westra, W. H. *et al.* Inverse Relationship between Human Papillomavirus-16 Infection and Disruptive p53 Gene Mutations in Squamous Cell Carcinoma of the Head and Neck. *Clin. Cancer Res.* **14**, 366–369 (2008).

145. Petitjean, A. *et al.* Impact of mutant p53 functional properties on *TP53* mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database. *Hum. Mutat.* **28**, 622–629 (2007).

146. Cho, Y., Gorina, S., Jeffrey, P. & Pavletich, N. Crystal Structure of a p53 Tumor Suppressor-DNA Complex: Understanding Tumorigenic Mutations. *Science* **265**, 346–355 (1994).

147. Rothenberg, S. M. & Ellisen, L. W. The molecular pathogenesis of head and neck squamous cell carcinoma. *J. Clin. Invest.* **122**, 1951–1957 (2012).

148. Havre, P. A., Yuan, J., Hedrick, L., Cho, K. R. & Glazer, P. M. p53 Inactivation by HPV16 E6 Results in Increased Mutagenesis in Human Cells. Cancer Res. **63**, 7215–7220 (1995).

149. Schaeffer, A. J. *et al.* E6 and E7 Oncoproteins Induce Distinct Patterns of Chromosomal Aneuploidy in Skin Tumors from Transgenic Mice. *Cancer Res.* **64**, 538–546 (2004).

150. Poeta, M. L., Forastiere, A., Ridge, J. A., Saunders, J. & Koch, W. M. TP53 Mutations and Survival in Squamous-Cell Carcinoma of the Head and Neck. *N. Engl. J. Med.* **10** (2007).

151. International Agency for Research on Cancer. Detection of TP53 mutations by direct sequencing. *IARC TP53 Database* (2010). Available at: http://p53.iarc.fr/ProtocolsAndTools.aspx.

152. Paquette, C. *et al.* Evidence That Alpha-9 Human Papillomavirus Infections are a Major Etiologic Factor for Oropharyngeal Carcinoma in Black South Africans. *Head Neck Pathol.* **7**, 361–372 (2013).

153. Walline, H. M. *et al.* Integration of high-risk human papillomavirus into cellular cancer-related genes in head and neck cancer cell lines: HPV integration into cancer genes in HNSCC cell lines. *Head Neck* **39**, 840–852 (2017).

154. Lewis, J. S. *et al.* p16 Positive Oropharyngeal Squamous Cell Carcinoma:An Entity With a Favorable Prognosis Regardless of Tumor HPV Status: *Am. J. Surg. Pathol.* **34**, 1088–1096 (2010).

155. Takeshima, Y. *et al.* p53 mutations in lung cancers from non-smoking atomic-bomb survivors. *The Lancet* **342**, 1520–1521 (1993).

156. Lee, L. N. *et al.* Exon 8 mutation of p53 gene associated with nodal metastasis in non-small-cell lung cancer. *Am. J. Respir. Crit. Care Med.* **150**, 1667–1671 (1994).

157. Berns, K. *et al.* A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* **428**, 431–437 (2004).

158. Sherman, R. M. *et al.* Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* **51**, 30–35 (2019).

159. Cullen, M. *et al.* Deep sequencing of HPV16 genomes: A new high-throughput tool for exploring the carcinogenicity and natural history of HPV16 infection. *Papillomavirus Res.* **1**, 3–11 (2015).

APPENDICES

Appendix A

Letter of ethics approval

IRB nr 00006240
REC Reference nr 230408-011
IORG0005187
FWA00012784

02 December 2015

MRS MUNSAMY
DEPARTMENT OF MEDICAL VIROLOGY
FACULTY OF HEALTH SCIENCES
UFS

Dear Mrs Munsamy

ECUFS NR 137/2013D (SUBSTUDY OF ECUFS NR 137/2013)
MRS MUNSAMY
DEPARTMENT OF MEDICAL VIROLOGY
PROJECT TITLE: CHARACTERISATION OF THE HUMAN PAPILLOMAVIRUS GENOME AND P53 MUTATIONS IN HEAD AND NECK SQUAMOUS CELL CARCINOMAS

1. You are hereby kindly informed that, at the meeting held on 01 December 2015, the Ethics Committee reviewed the above research project. Research may not be conducted before the following condition(s) has/have been met and the Ethics Committee grants final approval for the project:

   - *The signed permission letters from the Free State Department of Health must be submitted before final approval will be granted.*

   PLEASE NOTE: This ethics letter must accompany your application for approval to the Department of Health for their consideration, along with submitting the online application.

2. Upon receipt of the above feedback/document(s), the Ethics Committee will issue a final approval letter. Only thereafter may the study be conducted.

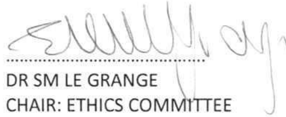
3. All relevant documents e.g. signed permission letters from the authorities/institutions; amendments to the protocol, questionnaires etc. have to be submitted to the Ethics Committee before the study may be conducted.

4. Kindly use the ECUFS NR as reference in correspondence to the Ethics Committee Secretariat.

5. Thus, this letter only serves as conditional approval.

6. The Ethics Committee functions in compliance with, but not limited to, the following documents and guidelines: The SA National Health Act. No. 61 of 2003; Ethics in Health Research: Principles, Structures and Processes (2015); SA GCP(2006); Declaration of Helsinki; The Belmont Report; The US Office of Human Research Protections 45 CFR 461 (for non-exempt research with human participants conducted or supported by the US Department of Health and Human Services- (HHS), 21 CFR 50, 21 CFR 56; CIOMS; ICH-GCP-E6 Sections 1-4; The International Conference on Harmonization and Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH Tripartite), Guidelines of the SA Medicines Control

Ethics Committee
Office of the Dean: Health Sciences
T: +27 (0)51 401 7795/7794 | F: +27 (0)51 444 4359 | E: ethicsfhs@ufs.ac.za
Block D, Dean's Division, Room D104 | P.O. Box/Posbus 339 (Internal Post Box G40) | Bloemfontein 9300 | South Africa
www.ufs.ac.za



Council as well as Laws and Regulations with regard to the Control of Medicines, Constitution of the Ethics Committee of the Faculty of Health Sciences.

Yours faithfully

....

DR SM LE GRANGE
CHAIR: ETHICS COMMITTEE

Ethics Committee
Office of the Dean: Health Sciences
T: +27 (0)51 401 7795/7794 | F: +27 (0)51 444 4359 | E: ethicsfhs@ufs.ac.za
Block D, Dean's Division, Room D104 | P.O. Box/Posbus 339 (Internal Post Box G40) | Bloemfontein 9300 | South Africa
www.ufs.ac.za

Appendix B

Table indicating accession numbers of sequences used in primer design and comparative phylogenetic analysis in Chapter 2.

AY262282.1	MF288672.1	MF288700.1	KC470225.1
GQ180785.1	MF288673.1	MF288701.1	MF288660.1
EF202144.1	MF288689.1	MF288702.1	MF288661.1
GQ180784.1	MF288690.1	MF288703.1	KC470229.1
GQ180789.1	MF288685.1	MF288704.1	KC470230.1
MF288705.1	MF288686.1	EF202150.1	
MF288706.1	MF288687.1	EF202151.1	
MF288707.1	MF288688.1	KC470213.1	
MF288710.1	MF288695.1	MF288723.1	
MF288708.1	MF288691.1	MF288724.1	
MF288709.1	MF288674.1	MF288725.1	
GQ180790.1	MF288675.1	MF288726.1	
GQ180788.1	MF288676.1	GQ180787.1	
GQ180791.1	MF288677.1	MF288727.1	
MF288711.1	MF288678.1	EF202152.1	
MF288712.1	MF288681.1	KC470226.1	
MF288713.1	MF288682.1	KC470228.1	
MF288714.1	MF288679.1	KC470227.1	
GQ180792.1	MF288680.1	EF202153.1	
EF202145.1	MF288684.1	EF202154.1	
MF288717.1	MF288694.1	MF288652.1	
MF288715.1	MF288692.1	MF288653.1	
MF288716.1	MF288693.1	MF288654.1	
KC470208.1	MF288696.1	MF288655.1	
KC470209.1	MF288697.1	MF288657.1	
NC_001357.1	MF288698.1	MF288656.1	
EF202143.1	EF202149.1	EF202155.1	
MF288718.1	KC470212.1	KC470214.1	
MF288719.1	MF288662.1	MF288658.1	
MF288720.1	MF288663.1	MF288659.1	
MF288721.1	MF288664.1	KC470216.1	
EF202146.1	MF288665.1	KC470215.1	
KC470210.1	MF288666.1	KC470217.1	
KC470211.1	MF288667.1	KC470219.1	
MF288722.1	MF288668.1	KC470220.1	
EF202147.1	MF288669.1	KC470218.1	
EF202148.1	MF288670.1	KC470221.1	
KU298886.1	MF288671.1	KC470222.1	
KX514433.1	GQ180786.1	KC470223.1	
MF288683.1	MF288699.1	KC470224.1	

Appendix C

Table of published list of functional regions of papillomavirus L2 from Wang & Roden, 2013.[128]

Table 2
Published list of known protein interactions with PV L2.

Stage of PV virus life cycle	Protein that interacts with L2	Interacting region on L2	Does the protein recognize a specific motif/consensus sequence on L2?	Purpose of interaction	References
Prior to infectious cell entry	Furin	9–12	R-X-K/R-R	Cleavage of L2	Richards et al. (2006)
	Cyclophilin B (CyPB)	90–110	N/A	Assist in capsid conformational change for secondary receptor uptake	Bienkowska-Haba et al. (2009)
	Annexin A2 Heterotetramer	108–120	Neutralizing epitope region 108–120	Putative secondary receptor	Woodham et al., (2012)
Vesicular trafficking and endosome escape	Cyclophilin B (CyPB)	90–110	N/A	Assist in capsid disassembly in the late endosome	Bienkowska-Haba et al. (2012)
	Gamma-secretase (GS)	Unknown	Unknown	Facilitates L2/vDNA endosomal escape	Karanam et al. (2009)
	Sortin Nexin 17 (SNX17)	245–257	F/YxNPxF/Y		Bergant Marusic et al. (2012) Bergant and Banks (2013)
Nuclear transport	Syntaxin 18 (BPV1 only)	43–47	DQ/KILQ/K	Transport of L2 toward nucleus	Bossis et al. (2005) Laniosz et al., (2007a)
	Heat shock cognate protein 70 (Hsc70)	Unknown	N/A		Florin et al. (2004)
	Beta-actin	25–45	N/A		Yang et al. (2003b)
	Dynein motor proteins (DYNLT1 & 3)	456–460, 457–461 (HPV16 L2)	R/KR/KXXR/K		Florin et al. (2006) Schneider et al. (2011)
	Karyopherins (Kapβ_2 and Kapβ_3)	1–9 or 454–462 (NLS signals at the amino and carboxyl termini respectively)	N/A		Bordeux et al. (2006), Darshan et al. (2004), Klucevsek et al. (2006), Fay et al., (2004), Sun et al. (1995)
Viral gene regulation, morphogenesis and assembly	ND-10 domain and binding to Daxx	390–420	N/A	Recruits other viral proteins and viral DNA to initiate viral gene transcription or assembly?	Becker et al. (2003) Florin et al. (2002b)
	PV E2	1–50 and 301–400	N/A	L2 brings E2 to the ND-10 domains (possibly to initiate viral gene transcription or replication)	Okoye et al. (2005) Heino et al. (2000)
	PV L1	412–455	PXXP	Viral capsid assembly	Finnen et al. (2003) Lowe et al. (2008)
	TBX2 and TBX3	L2 C-terminus region	N/A	L2 brings TBX2/3 to the viral genome's LCR where TBX2/3 then represses of viral gene transcription	Schneider et al. (2013)
Other known interaction protein partners but with undefined function	• PATZ • TIP60 • TIN-AG-RP • PLINP	Unknown	Unknown	Unknown	Gornemann et al. (2002)
	SUMO (small ubiquitin-related modifiers)	34–37	PVKE	Unknown	Marusic et al. (2010)

Appendix D

Table indicating accession numbers of sequences used in primer design and comparative phylogenetic analysis in Chapter 3.

HQ537679.1
HQ537681.1
HQ537680.1
KX514424.1
HQ537678.1
HQ537677.1
HQ537676.1
KU298890.1
HQ537672.1
HQ537669.1
KU298889.1
KU298888.1
J04353.1
HQ537666.1
HQ537668.1
HQ537667.1
HQ537675.1
HQ537674.1
HQ537670.1
HQ537671.1
HQ537673.1
KX638481.1
HQ537687.1
HQ537686.1
HQ537685.1
KX514430.1
HQ537683.1
HQ537684.1
HQ537682.1

Appendix E

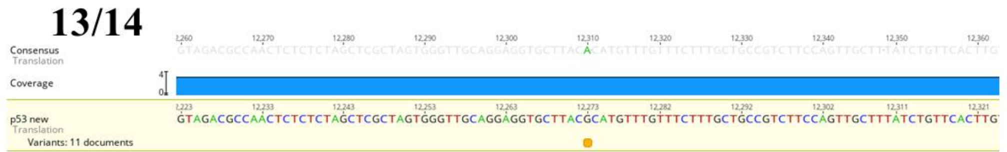
Tables of raw data to calculate specific activity.

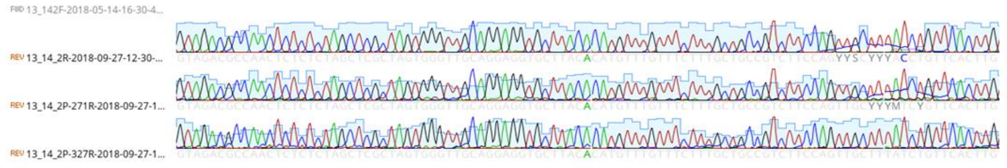
Replicate	Plasmid construct	Protein concentration of cell lysate (µg/ml)	Protein concentration dilution factor corrected *200 (mg/ml)	OD420	nmoles ONPG	nmoles ONPG/min	Protein conc µg	Adjust to nmoles/min/mg	Specific activity	
				Cell lysate volume	Cell lysate volume	Cell lysate volume	Cell lysate volume	Cell lysate volume		
				5 µl	5 µl	5 µl	5 µl	5 µl	5 µl	
1	**pBlue LCR 5**	14,9	2,98	0,021	3,73	0,12	14,9	67,11	8,35	0,87
	SDM 1 clone 4	11	2,2	0,031	5,51	0,18	11	90,91	16,7	1,74
	SDM 3 Clone 1	11,1	2,22	0,023	4,09	0,14	11,1	90,09	12,28	1,28
	Untransfected	14,8	2,96	0,024	4,27	0,14	14,8	67,57	9,61	1
2	**pBlue LCR 5**	5,28	1,056	0,041	7,29	0,24	5,28	189,39	46,02	0,97
	SDM 1 clone 4	2,45	0,49	0,012	2,13	0,07	2,45	408,16	29,02	0,61
	SDM 3 Clone 1	3,15	0,63	0,038	6,76	0,23	3,15	317,46	71,49	1,51
	Untransfected	1,5	0,3	0,012	2,13	0,07	1,5	666,67	47,41	1
3	**pBlue LCR 5**	13,1	2,62	0,014	2,49	0,08	13,1	76,34	6,33	0,80
	SDM 1 clone 4	8,04	1,608	0,024	4,27	0,14	8,04	124,38	17,69	2,24
	SDM 3 Clone 1	10,4	2,08	0,031	5,51	0,18	10,4	96,15	17,67	2,2
	Untransfected	10,5	2,1	0,014	2,49	0,08	10,5	95,24	7,90	1

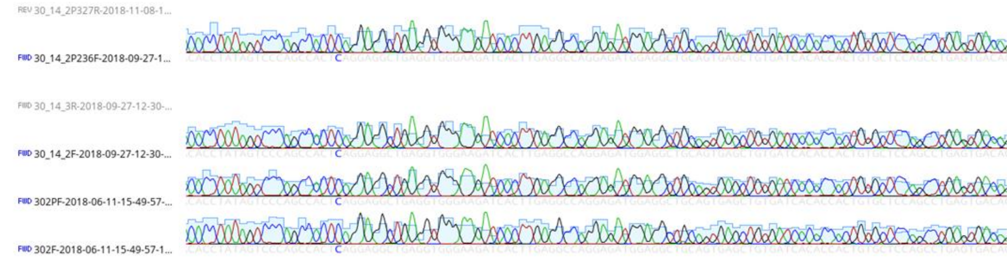
	pBlue_VBD13_14	pBlue_SDM1	pBlue_SDM3	Untransfected
Mean	0,88	1,53	1,67	1
Std Dev	0,09	0,83	0,5	0

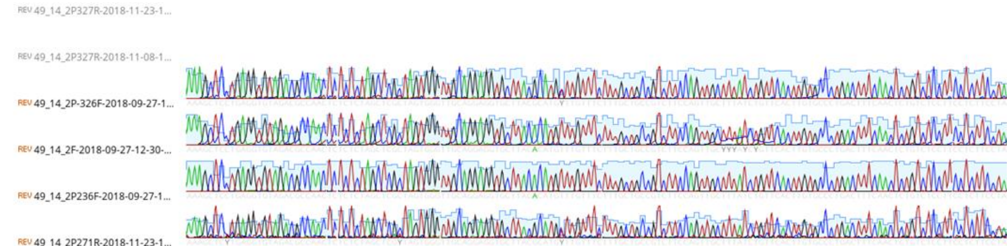
Appendix F

Electropherograms mapped to *p53* sequence to call variants.

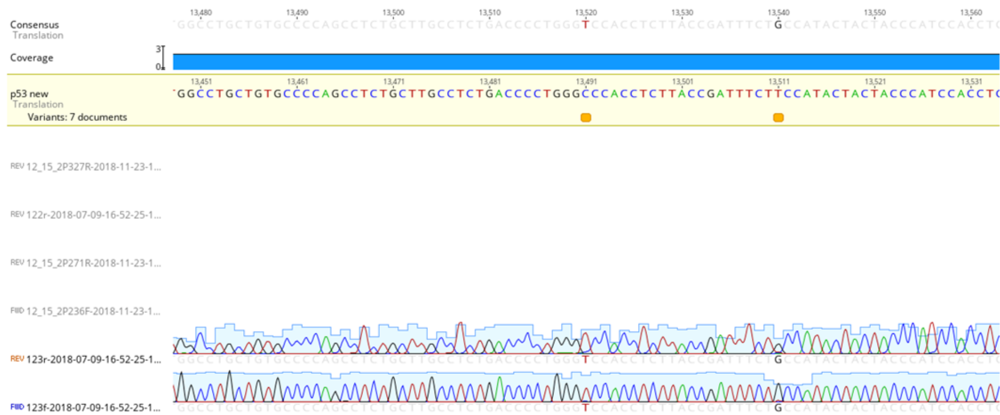




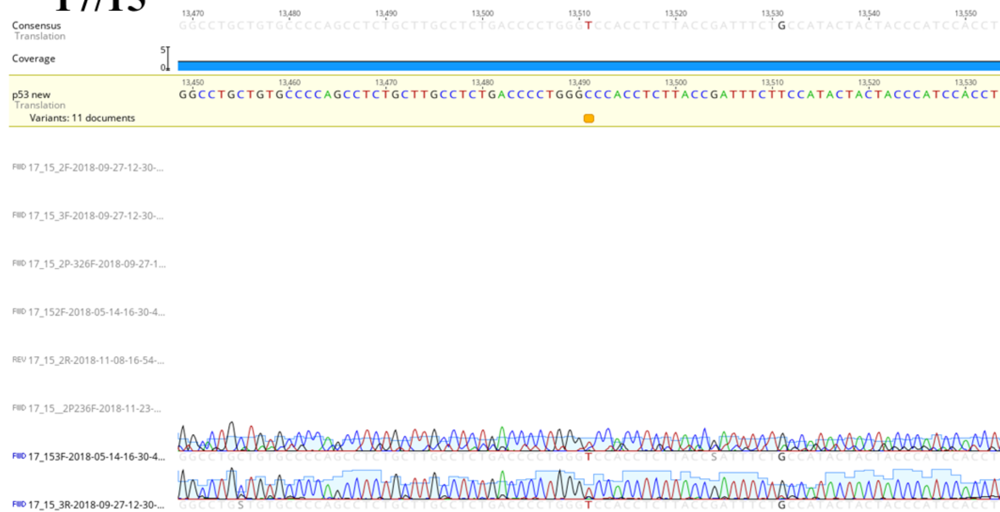




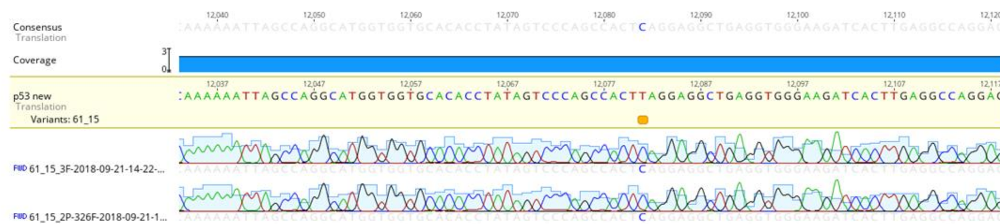
12/15



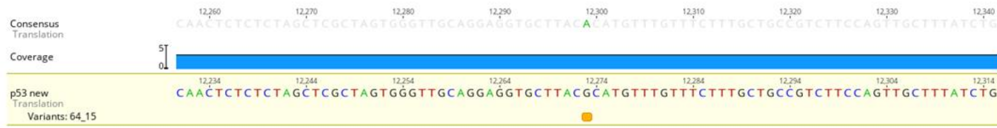
17/15



61/15

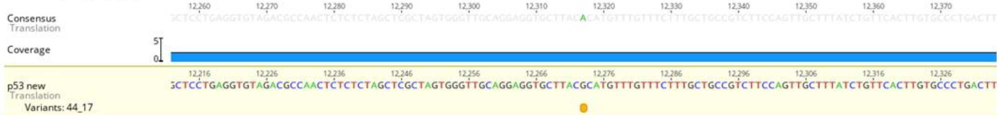


64/15



09/17

44/17



08/18